Statistical Aspects of Surgical Audit

A Dissertation submitted to
The University of Glasgow
for the degree of PhD

by

Catriona E. Hayes

September 1995

SUMMARY

The emphasis of this thesis is on "Comparative Audit", an area which has become increasingly popular in recent years. Firstly, the history of Surgical Audit is explored, from the first data collection to the publication of League Tables. It is evident that any comparisons of outcome measures must involve adjustments for patient case mix. This must be done using a statistical model. The features required of a model for this use are described, and a review of available predictive systems in medicine shows that none of the widely known predictive models satisfy all of the criteria which are appropriate for use in audit.

An interest in Comparative Audit has led to the acquisition of several large data sets. The first of these is from the Royal College of Surgeons of England (RCS) Comparative Audit Service. The methods of presentation they used originally did not utilise the full potential of the data. The change to presentation of death rates as relative mortality confidence intervals rather than ranked bar charts allows consultants to see at a glance whether their rates are significantly different from the mean value. Adjusting these for some aspects of case mix makes a substantial difference to the rank order and, at the very least, highlights the folly of publishing League Tables of raw mortality rates. The adjustments as they stand are crude for three main reasons. The data are of poor quality, and they were never intended for this type of analysis. Also, data are collected as totals for consultants rather than for individual patients, which limits the potential for modelling. Through a case study, the problems of extracting data in the format required by the RCS are investigated, and based on this, together with consideration of more technical issues related to statistical modelling, suggestions are made as to how the RCS data collection exercise could be improved. The issue of modelling based on aggregate data is also explored, using both case studies and simulated data. These investigations show that the analyses are still of worth, as the adjustments have the desired effect if there are a reasonable number of consultants and patients, and if there is no, or only weak, correlation between the explanatory variables.

To progress from the methods introduced for the RCS, it has been suggested that the POSSUM system be used for adjustment. This is currently the most widely used system, but it fails to meet several criteria for models for use in audit. The second data set explored is from a large audit study in Portsmouth, in which the POSSUM variables were collected. It emerges that a greatly reduced model of only four variables performs as well as the POSSUM model. Using this would not only greatly facilitate data collection, but would substantially reduce the amount of missing data, and thus diminish the difficulties of bias which arise from this.

It may be that general surgery is too broad and heterogeneous to model, and that specific areas should be tackled separately. Analysis of colorectal cancer audit data shows that good models can be acquired, and used to compare surgeons. The relationship between patient volume and outcome, a related topical question, is also investigated with these data.

CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

ACKNOWLEDGEMENTS

I would like to thank Dr Gordon Murray for his invaluable help and advice, and the SERC for their financial support. I would also like express my gratitude to Sarah Fowler at The Royal College of Surgeons of England Comparative Audit Service for the four annual data sets and The University Department of Surgery at Glasgow Royal Infirmary for their data, and for allowing me to attend their audit meetings. I am also indebted to David Prytherch and Paul Weaver of Portsmouth Hospitals for the POSSUM Study data, Colin McArdle from the University Department of Surgery at Glasgow Royal Infirmary and David Carter from the University Department of Surgery in Edinburgh for access to the CRAG colorectal cancer database, and Ann Gould of the Scottish Cancer Therapy Network for technical assistance with the Lothian data from the CRAG study.

I would also like to thank the Department of statistics at Glasgow University for their help and support, and the many friends I have made during my stay. In particular, I couldn't have done it without Lucy Stevenson, Graeme Archer, Rodney Wolff and Eileen Wright.

DECLARATION

Part of chapter 4 has previously been published (Murray *et al.*, 1995), and chapter 2 has been submitted to the Journal of Evaluation in Clinical Practice, under the title of "Case Mix Adjustment in Comparative Audit".

1 INTRODUCTION

1.1 What is Surgical Audit?

The Government's white paper *Working for Patients* (1989) defined medical audit as

> The systematic, critical analysis of the quality of medical care, including the procedures used for diagnosis and treatment, the use of resources and the resulting outcome for the patient.

The word "audit" has its roots in accountancy and thus has connotations of financial stringency. In fact, its primary concern is the effectiveness of care, rather than efficiency or economy, although audit should help make the optimum use of the resources available. Perhaps a preferable term would be quality control or health care evaluation. The process of audit has been idealised as a loop (Dudley 1974, Shaw 1980, Crombie & Davies 1993). This audit "cycle" consists of three stages. First, standards must be set, then care should be evaluated and compared with the standards. Last, practice must be reviewed in the light of this and necessary changes made. The standards must then be reviewed and the process begun again, completing the loop. Audit involves collecting data on clinical practice, but differs from clinical research in that it is a review of practice that may uncover problems which must then be investigated if their cause is not obvious. Thus audit can lead to research but is not the same thing.

Surgical audit, then, is the evaluation of the treatment given by surgeons with a view to making changes where necessary in the light of information gained.

1.2 Why audit?

The above mentioned white paper stipulated that audit be an integral part of medical practice. This goes hand in hand with the greater accountability being demanded of doctors. They are now expected to justify resource allocation, and also make decisions about future treatment priorities. Thus, information on the effects of interventions is necessary. In today's society, hard facts are crucial for many reasons. Surgeons will need summaries showing surgical performance that is in line with their contemporaries in defence against litigation if malpractice suits become more common. If the Government introduces contracts for surgeons then the information will be necessary in the application for renewal.

Quite apart from the legal requirement, the process of audit is beneficial in itself. Firstly, it is educational. It promotes discussion, both of raw patient data and results, and thus communication of surgical knowledge, which should lead to better treatment. It also raises questions which can stimulate research projects. Secondly, the very act of recording data may encourage more diligence from surgeons.

Many surgeons are very interested in comparing their results with those of their peers. This could either reassure them that they are performing satisfactorily, or they may have to investigate the reasons for any discrepancies. Audit information can show up any weaknesses and thus should lead to investigation and so policy changes. In several cases audit data have been used as evidence by departments making cases to health boards for more resources. For example the team in Cambridge under DC Dunn was able to demonstrate a continued need for their intensive care unit when the authority wanted to transfer the funding to geriatrics (Dunn 1988). The Lothian Surgical Audit has successfully effected many policy changes. The data on breast disease provided concrete proof of a need for more staff in the unit, and a new vascular unit was established after data showed far lower death rates following ruptured abdominal aortic aneurysms if operations were carried out by specialists (Nixon 1992).

1.2.1 Audit as a way of evaluating interventions

The above mentioned data from the Lothian Surgical Audit on breast disease also showed that breast abscess was better treated by needle aspiration and antibiotics than by the traditional method of surgical drainage (Nixon 1992). This illustrates the value of audit data in comparing surgical interventions, as well as comparing surgeons or hospitals. In fact, it has been suggested that in certain circumstances (e.g. Pollock 1993) careful audit could be a substitute for randomised controlled trials (RCT's). There are many ethical problems in carrying out such trials, and these are particularly acute in the field of surgery (Pollock). For example, it is unlikely that a surgeon believes equally in, or is similarly expert in each procedure, and, unlike in a drugs trial, the implementation of blinding is impossible at the operative stage so bias from patient or surgeons' attitudes can never be eliminated. When comparing two treatments, for example laparoscopic cholecystectomy *versus* the traditional method, patients in a trial have to be randomly assigned to one of the methods. However, as Neugebauer *et al.* (1990) found in their prospective trial of gall bladder removal the laparoscopic method was so much preferred by the time the surgeons had become competent at the technique that it was considered unethical to randomise people to the conventional treatment. For random allocation of a patient to a treatment, both doctor and patient should have no preference for either one. It has been argued that the probability that these two rare events occur simultaneously is so minute that RCT's are ethically impossible. On the other hand, many people see it as unethical to introduce any new treatment without subjecting it to a controlled trial, and would go so far as to say that patients have a moral duty to participate for the common good (Baum 1993). The other reasons for randomisation to treatments, apart from avoiding assignment bias are, firstly to balance the treatment groups with respect to both known and unknown prognostic factors, and so avoid any other possible biases, and second, to provide a basis for performing statistical significance tests. (e.g. Byar *et al.* 1976, Schwartz *et al.* 1980, Gore 1981). This point of view is not held by Bayesians, who disagree with the underlying logic of the statistical inference. They would advocate that a difference between treatment groups in a trial does not logically imply a treatment effect as there could be unknown factors correlated with the treatment (Urbach 1993). Thus, since all factors which could be associated with outcome should be considered, they do not see randomisation as necessary and would recommend historical trials using comprehensive databases. This is very difficult in practice however, as at the time of collection of data possible important contributory factors may not have been recognised. This problem is not so important for audit as it is for a clinical trial situation. As has been illustrated, the information contained in audit databases can be very useful in showing up treatment differences where they may previously have gone unnoticed, or in conclusively demonstrating very large effects, e.g. the introduction of penicillin. A difference between treatments may be shown, but there is no way of knowing whether this is caused by the motivation behind giving a certain treatment to a certain patient or if there was an actual effect (Byar 1980). For example, only patients with good prognoses may have been assigned the new treatment. The information recorded on a database is unlikely to include changes in diagnostic technology or nursing staff, or other variables which could influence results. Then comparisons over time (say the treatments were given in series) will involve large biases that cannot be corrected for. In an RCT, a specific treatment protocol is defined, whereas usually there are variations in surgical technique, the details of which are not recorded. Similarly, there is no way of accounting for observer variation when looking at data retrospectively. For example in a cancer trial, one pathologist will usually assess all the tumours, whereas a database could have input from a number of different experts and the amount of variation will remain unknown. As previously mentioned, the random allocation of treatments should minimise any differences caused by patient characteristics such as age and other contributory factors, both known and unknown. This means we can avoid the subjective judgements about important factors that would have to be made in order to analyse the database data. Often, a small difference can be clinically significant but this is unlikely to be spotted from audit data due to these large amounts of random variation in the data or inherent bias in the allocation of treatments to certain patients. RCT's are still necessary to show these small but important differences. For audit, we are not usually interested in making a decision about treatment so most of the errors are not so important. If changes in technology for diagnoses, say, lead to improvements in outcomes these will be evident but it is not necessary to estimate the actual effect. The results of audit are more likely to lead to a randomised controlled trial than be a substitute for one, and the data from a trial could be part of the audit database. The two types of study must co-exist.

1.3 What should we audit?

The definition given in the white paper suggests that we can evaluate three different aspects of health care. These were first categorised by Donabedian in 1966, who used the following terms (Russell, 1987).

1. Structure.

This is best thought of as the resources available for treatment. It is probably the easiest to evaluate. An audit of structure would be, for example, an enquiry into the most senior surgeon available at each operation or an assessment of the availability of particular equipment. Structure was not included in the remit of the white paper.

2. Process.

This is concerned with the actual administration of treatment. It is audited by assessing adherence of clinicians to standards, which are consensually agreed protocols defining "good" treatment. Process was considered most in the past because it is easier to collect data while patients are present rather than have to trace them for follow-up data.

3. Outcome.

The final effect of treatment on the patient, and arguably the most important and relevant thing to evaluate. It is also the most difficult to define and measure, and it takes much more time and money to acquire the data. Examples of outcome measures include mortality rates, numbers of complications or patient satisfaction.

Several studies have shown structure and process to correlate poorly with outcome (Brook & Appel 1973). Obviously, better facilities and treatment are likely to lead to better outcomes, but the small details of medical care that are important in achieving good results in individual cases are not easy to measure or summarise. It is the inability to gauge accurately the process of care that causes the apparent poor relationship between it and outcome. This relationship between process and outcome can be compared to the one between explanatory and pragmatic clinical trials. The definitions of these are given elsewhere (Schwartz *et al.* 1980), but basically an explanatory trial is designed to answer a specific biological question while any other factors are kept as constant as possible as in a laboratory, whereas a pragmatic trial is to compare treatments in "real life" and requires many more subjects to achieve statistical power. Explanatory trials are required to evaluate individual mechanisms of treatment, for example a drug to lower blood pressure. This type of trial could show that the drug was effective, but would not show up any actual patient benefits, such as a reduced risk of stroke a few years later. Any such real life long term gains would have to be investigated using a pragmatic trial. So, just as explanatory and pragmatic trials must be undertaken in conjunction, audit of process and structure are complementary, depending on the desired purpose of the audit, but it cannot be assumed that one can represent the other.

This thesis is concerned with the overall benefit of treatment to patients, rather than the mechanics of treatment itself, and so is concerned only with outcome audit. The outcome of a treatment is difficult to define, and a relevant measure for an operation will depend on the type of procedure being performed. A commonly used outcome measure is mortality rate. This is particularly relevant for high risk procedures, but is not a good measure for those operations where death is very rare. Mortality is fairly easy to define, although there are still problems of whether to measure in hospital mortality or deaths within 30 days, say, and care must be taken to compare like with like. Other measures are even less clear-cut. There may not be agreement over what constitutes a complication, patients may be more easily lost to follow-up in a study of recurrence rates and patient satisfaction is very difficult to assess consistently. Raw mortality rates, however, are not a useful indicator of surgical calibre although they have often been quoted. For example, in 1986 the mortality rates of Medicare patients in 6000 US hospitals were published. Certain hospitals could boast exceptionally low rates whereas others seemed very high as no account was taken of any possible confounding factors such as diagnostic case mix or demographically different patient populations. It is well known that high mortality rates can be indicative of a highly skilled consultant who is willing to take more risks and, likewise, low rates of a technically incompetent one who rarely operates. (Pollock & Evans 1989)

1.4 History of Surgical Audit

1.4.1 Early history

As long ago as 1858, Florence Nightingale realised the importance of collecting information on hospital patients. She gathered statistics on the soldiers she was treating in the Crimean War, and compared them with statistics for Britain, to show the poor conditions in the British Army hospitals. In 1863 she proposed that a record be kept of all operations, including outcome, with a view to improving efficiency and effectiveness (Devlin, 1990). The first systematic recordings of outcomes after surgery were made in Europe in the 1890's. Two separate studies of inguinal hernia repair operations are classic examples of the necessity of this type of work. In 1890

Dr Haidenthaller published very poor results for Professor Billroth's Vienna clinic; there were 11 deaths out of 195 patients and a 30.8% recurrence rate. As a contrast, in 1894, Professor Bassini of Padua published results of 262 operations with no deaths and only 7 recurrences in a four and a half year follow-up. These showed that large differences can exist, and prompted a rethink in treatment strategy for inguinal hernias.

1.4.2 Development in the USA

Most of the initial progress in surgical audit was made in the USA. Although the arrangement of medical care is very different from the UK, some important principles have been established in the American quest for quality improvement this century. It should be remembered that the standard of data collection in US medicine is much higher than in the UK, and that large amounts are spent on management. The original data collection was concerned with funding, and case notes are detailed and accurate and thus useful for audit. However, with the introduction of NHS Trusts and the demand for greater accountability in the UK, more similarities will exist with the American system.

The publication of the Flexner report on medical education in 1910 inspired the beginning of reforms in the United States that changed medical practice fundamentally. There were many untrained, incompetent surgeons operating unnecessarily for financial gain in poorly equipped hospitals, with no record being made of treatment outcomes. The main pioneer of surgical audit was Ernest Codman, who described the *End Result Idea,* i.e. that hospital care could be standardised and related to outcomes. He proposed that an independent regulatory body be established, and from this the American College of Surgeons (ACS) was formed in 1913. Following Codman, the ACS monitored outcomes but the initial survey reported that so few hospitals were up to standard that the data were destroyed immediately, and a new approach taken for evaluation. A "minimum standard" was defined and a list of establishments fulfilling this was published. The ACS controlled this Hospital Standardisation Program for over 30 years, and succeeded in greatly improving the care given by hospitals. In 1951, due to the increasing number of medical specialities, the Joint Commission on Accreditation of Hospitals (JCAH) was formed. This body continued the structure and process audit, approving hospitals attaining minimum standards, until 1970 when the Federal Government assumed this "policing" role following the introduction of the Medicare scheme. During the late 1940's there were also many independent audits of the process of care, assessing the skill with which procedures were performed. A major problem in the USA was still unnecessary surgery, and previous studies could not be used as weapons against this as they were based on subjective judgements. In 1956, Lembcke took a very significant step forward in surgical audit when he introduced his scientific principles of auditing. These said that the audit should be carried out internally, with only occasional external checks. Objective, verifiable, uniform, specific, relevant and acceptable criteria were to be defined for particular procedures, and standards involving percentage compliance with these explicit criteria introduced. These standards were based on observed values. These principles of Lembcke were very effective in decreasing numbers of unnecessary operations, and if surgeons were repeatedly shown to be involved in these, their practice was restricted. Thus the ideas became associated with punitive action.

In 1972, surgical audit was made a legal requirement in the USA. The JCAH standards expanded to cover more quality assessment and several review bodies were formed, the most important being the Professional Standards Review Organisations (PSRO). In the 1970's the emphasis was shifting from setting standards at the minimum acceptable level to the optimum achievable level. Since these standards relied on process of care, and there was no evidence of the effect of this on outcome, the evaluation of the end result again came into favour. There was a period of controversy over whether to use process or outcome for evaluation, but as the complexity of health care measurement became more evident it was realised that different objectives require different measures. This illustrates again the parallel with clinical trials. A treatment in its infancy must be tested on a small number of people, looking at its mechanics without too much consideration of long term outcome. Similarly, the process of care had to be satisfactory before it was relevant to assess outcome measures. The two types of audit have to co-exist. The PSROs were not very successful in improving quality, nor in fulfilling their main purpose which was cutting costs. Their main problem was getting doctors to agree with their standards.

Poor outcome data were dismissed as being due to external factors and not themselves. Because the audit cycle was not being completed, i.e. there was no feedback, the audit by PSROs only succeeded in collecting large quantities of "orphan" data. These organisations were disbanded in 1984 and Peer Review Organisations introduced in their place. These "PRO's" are groups of professionals who investigate the notes of all Medicare funded patients in a hospital and check them against specified criteria of quality of care and appropriateness of admissions. If they judge any procedure to be unacceptable they can withhold funding to the hospital. They have succeeded in curtailing spending in hospitals, but it is doubtful whether their punitive methods have done much for the improvement of quality. Coinciding with the introduction of the PRO's was the payment of hospitals for their Medicare patients by Diagnosis Related Groups (DRG's). These are 467 groupings which are defined by the type of patient, both clinically and in relation to their demand on resources. Each case can be allocated to only one of the groups on the basis of many factors. DRG's were criticised for being too heterogeneous for severity of illness, and it was observed that there was financial incentive to admit patients at the less severe end of the groups, as the same amount of money would be allocated as for a more severely ill patient in the same group. For this reason, disease staging was proposed (Gonnella *et al.*, 1984). This was a complex way of assigning a severity stage to most diseases, depending on comorbidities present. The coding process was incorporated in a software package which could "stage" diagnoses in a hospital discharge database. While funding for Medicare patients has demanded development of audit techniques, private hospitals are also often evaluated by the state, and the results published. It has even been known for hospitals to advertise based on their superior success rates, even if these have not properly accounted for case mix.

1.4.2.1 COMPARISONS BETWEEN HOSPITALS

The first attempt at modelling hospital mortality data was in 1968, when surgical death rates were adjusted for age, sex, operation and physical status, and a three-fold variation among hospitals was found (Moses 1968). There has been much interest since then in comparing hospitals, although not specifically in surgery. In 1986, it was made a statutory requirement for mortality rates of Medicare hospitals to be published. These were released by the Health Care Financing Association (HCFA) on an annual basis, with adjustments for case mix gradually growing in sophistication. They used a logistic regression model containing the patient's age, gender and comorbidities, number of hospitalisations in the previous year and whether they had been transferred from another hospital. It did not have a measure of severity of illness. Dubois *et al.* produced a model in 1987 that accounted for 64% of variation in death rates between the 93 hospitals in their study. They claimed that this model was superior to the one used by the HCFA. They calculated ratios of observed to expected death rates and identified 11 of the hospitals where the predicted mortality rate was significantly less than the observed one, and 9 where it was significantly more. The outliers were then investigated further, with mortality rates being broken down by diagnosis. In a subsequent study of case notes, more deaths in the high outlier hospitals were judged to be 'preventable', a finding which had not been evident from standard analyses of process of care. Several other models for adjusting hospital mortality have been suggested, all authors claiming more predictive power than the HCFA method. For example, The Medicare Mortality Predictor System (MMPS) (Daley *et al.* 1988). This, however, was specifically for patients over 64 years old admitted with stroke, pneumonia, myocardial infarction or congestive heart failure, and excluded most surgical patients. In 1990, Green *et al.* produced a model which was based on the HCFA model, but added a severity of illness measure based on staging. They found an R^2 value of 2.5% for the HCFA model and 21.5% for their model, and concluded that it was far superior. It is recognised that higher death rates do not necessarily imply poorer quality of care, and that more severely ill patients or random chance all have a part to play. These methods are seen as ways to target hospitals for further investigation. For surgery, we can learn from the many attempts to compare hospitals, but we are interested in results of interventions rather than purely the outcome of hospital stay.

1.4.3 Development in the UK

The first attempt to introduce a national audit of outcome was made in 1908 by Ernest Groves. At that time the idea of any review of surgical treatment or enquiry into postoperative deaths was extremely unpopular with the establishment. Groves suggested that all hospitals should collect statistics on operations with their immediate and remote numbers of deaths. He realised that the information would be invaluable if collated nationally on an annual basis, for knowledge on prognoses of procedures, incidences of diseases and operations and to show improvements over time. In his own survey he found the numbers of institutions collecting statistics to be very poor. After publishing his proposals there was no response to them at all and the topic lay closed for almost 50 years. Traditionally, there has been some regulation of the medical profession through its own institutions, such as the Royal College of Surgeons, but formal audit received very little attention until the mid 1970's.

The earliest national audit in any field was in maternal mortality. Data on obstetric deaths have been collected since 1932 and reviewed by experts, greatly improving the standard of care in this area.

Anaesthetists have had considerable influence on the development of surgical audit. In 1956 Edwards *et al.* published a report on anaesthetic deaths. The national enquiry carried out by the Association of Anaesthetists of Great Britain and Ireland in 1980 concluded that very few deaths could be attributed to anaesthesia alone, and that surgeons and anaesthetists should co-operate. A major study in which this was achieved was the Confidential Enquiry into Perioperative Deaths (CEPOD) (Buck *et al.* 1987). This was started in 1986 with a pilot study, which pointed out the need for an improvement in audit methods. CEPOD found that data from the Hospital Activity Analysis (HAA) were rather inaccurate, as has been shown by many authors (Rees 1982, Whates *et al.* 1982, Baron 1987). The Scottish equivalent of the HAA was the Scottish Hospital Inpatients Statistics (SHIPS). A major problem with these national audits is that the information is not available for 2-3 years and so becomes dated and of little use for practice (Ruckley 1984).

One of the strongest advocates of surgical audit in recent times was Dudley, who published a paper in the BMJ in 1974 entitled *Necessity for Surgical Audit.* The editorial in that edition was also devoted to the topic, and throughout the following years interest grew rapidly, with many publications in the literature. There was some opposition to the idea from various authors, and by 1980 there was still little evidence to show the effectiveness of audit. It was important that audit be seen as educational, rather than as a method to apportion blame, and the abrasive methods used in the US were looked at as a warning of how not to proceed in this country. Many early audits were concerned with workload, but results were compared between two hospitals by Gilmore *et al.* in 1980. Gough *et al.* (1980) noted that computers could make audit easier than performing it manually. Probably the best example of a successful audit is that of the long running Lothian Surgical audit which was started in 1979. At first the data collection was done by hand, then it was transferred to computer. This is the first example of a regional audit, and also before this was set up most audits had been of process rather than outcome. Originally, only mortality was considered as an outcome measure, and then re-operations for complication or surgical failure. The Lothian audit team found the OPCS operation codes unsatisfactory, and developed a new coding system. The discussion of the mechanisms of audit by C.V. Ruckley in 1984 was based on the Lothian experience. He pointed out that too much data should not be collected, and the importance of the audit cycle. He also stressed the importance of good data presentation: "Information must be intelligible, easily digestible and attractively presented." The report on five years of the Lothian audit (Gruer *et al.* 1986) described the trends observed and the benefits achieved by the audit. There were significant falls in mortality over the observed period. The data were of great use for more than the immediate audit requirements.

Another important audit was started by D. C. Dunn in 1982 (Dunn 1988). He started collecting audit data on a microcomputer, along with a diary until it was decided exactly what variables were important. Dunnfile was developed gradually in the early 1980's, eventually incorporating a word processor for the automatic generation of discharge documents (Dunn & Dale 1986). It has since been renamed as the Cambridge Audit System and installed in other hospitals. The development of a different system for surgical audit has been described by Ellis *et al.* (1987). This has evolved into the *Micromed* package. Many computer systems have been developed and marketed since these original ones, and these will be discussed in more detail later.

During the 1980's there emerged a general consensus as to the importance of audit and methods for data collection were much improved. However, the problems of simply using complication or mortality rates for comparisons were rarely tackled until the latter part of the decade. In 1987, Deans *et al.* proposed that deaths be classified as “avoidable” or “unavoidable”, as consensually agreed by a group of surgeons considering all the contributory factors. This was fairer than looking at raw mortality rates, but was time consuming and relied on subjective judgement rather than scientific methods. Similarly, Avoidable Mortality Rate’ (AMR) was studied in Zambia (Heywood *et al.*, 1989). It was concluded that regular audit meetings would help to reduce the rate and that the index AMR would show improvements in management over time. The definition of ‘avoidable’ mortality was where there was evidence of mismanagement that could account for a death. This type of audit is perhaps more appropriate for a developing country, as death rates are higher, and the types of mismanagement considered are far less subtle than those usually occurring in a country such as the UK. For example, almost 1/3 of the patients in this study who died were affected by some sort of administrative factor such as lack of blood. It is useful to label deaths as ‘unexpected’ or ‘avoidable’ for audit, as total numbers of deaths do not give any information as to how many patients were admitted for palliative care. As well as requiring personal judgement, categorising deaths as avoidable in this country could be highly politically sensitive, and it is unlikely that consultants would release this information themselves. The National Confidential Enquiry into Perioperative Deaths (CEPOD), however, released information on numbers of these for large samples of surgeons nationally, and response rates were very high. Details on every death were submitted to the Enquiry, and these were then judged by an independent team of experts as to whether they could have been avoidable. The initial study found that 8% of deaths analysed were due to deficiencies in surgical care, and between 8% and 25% of all deaths were avoidable. (Buck, Devlin & Lunn, 1987). The success of this study lies in the confidentiality. The data were all destroyed before publication, and no information on hospital, surgeon, or patient was attached to the pro-formas. The study also received “Crown Privilege”, whereby none of the data could be used as evidence in court. Very useful lessons have been learned from this study, and individuals were never referred to. There is, however, increasing demand for information on individual surgeons.

The British Medical Journal published an article at the start of 1989 (Mortensen) which recognised the need for audit and a standard definition for mortality, which should take into account the condition of the patient using a scoring method. The audit comparing four health regions using Dunnfile showed large differences in complication rates, but pointed out the variation in case mix and patient types that have to be taken into consideration (Dunn *et al.* 1992). This study was the fore-runner of the Comparative Audit carried out by the Royal College of Surgeons. POSSUM is a scoring system which was specifically devised for use in audit (Copeland *et al.* 1991). It involves calculating scores for operative severity and physiological derangement, and including them in a logistic regression model. It will be discussed in more detail later. An alternative approach to standardising outcome measurements for use in comparisons was proposed around the same time as POSSUM by Cale *et al.* (1991). This involved constructing a "morbidity profile" rather than mortality rates.

Interest in comparative audit has grown in recent years, with the Royal College of Surgeons introducing their Comparative Audit service (Dunn & Fowler 1992). The originators of POSSUM have also been interested in this area (Copeland 1993), and have carried out comparative studies in the specific areas of vascular surgery (Copeland *et al.* 1993) and colorectal surgery (Sagar *et al.*, 1994) as well as in General Surgery (Copeland *et al.*, 1995). There is, however, still a risk of comparative audit data being misinterpreted in the media (Brindle 1994, Toynbee, 1991, 1993) and of them being used for the wrong purposes.

In September 1993, the first NHS hospital 'league tables' were published. These were of waiting times in the West Midlands. They caused uproar among doctors, who claimed that they were unfair and meaningless (Beecham 1993, Jones 1993). Long waiting times were blamed on lack of resources, rather than efficiency. There was anger at raw figures being published without explanation. The first official national league tables were published in June 1994. These awarded between one and five stars to hospitals in England on 23 performance measures of waiting time, speed of attention in casualty and out-patient departments, cancelled operations and use of day-case surgery. They were criticised as being flawed, irrelevant and only measuring efficiency rather than effectiveness of care. Purely administrative data have, however, inspired far less emotive journalism and public interest than the publication of death rates. Tables of mortality rates were released in Scotland in December 1994. The mortality rates from cardiac arrest, stroke and broken femur were given separately, and wide variations were observed. It was stressed that these were likely to be due to differences in patients, especially social class. The use of outcome figures rather than information on throughput was a step in the right direction, but there is still a long way to go before outcome measures will be sophisticated enough to give a reliable indication of quality. A recent study examined methodological issues related to mortality league tables, and attempted to adjust for age, emergency admission and disease stage (McKee & Hunter 1995). It highlighted several problems with analysing the data. The first is that sample sizes are often too small to allow conclusions to be drawn, if a particular disease is studied for a year. A longer time period renders the information irrelevant. Adjustment for severity is also a problem. These authors chose to use the American system of disease staging rather than develop their own model. The poor coding of data is also highlighted. A major issue in the discussion of audit recently has been the poor quality of data, and the difficulty of extracting information from databases (Cleary *et al.* 1994). Much improvement is still required in the collection and standardisation of data.

A related area which has received much attention in the literature is the relationship of case load with outcome. Many authors have advocated that specialisation of surgical units leads to improved treatment, by looking at audit data on number of particular procedures carried out. There are some who do not agree that this relationship exists. The arguments have been reviewed recently (Houghton 1994). Although there are no data unequivocally supporting the need for increased specialisation, there are many practical reasons for it. For example patients can be near other people with similar illnesses, facilities can be more up to date, and specialists can work together rather than being alone in individual units.

1.5 How is audit effected?

The process of audit will obviously depend upon the level at which it is carried out. This could be local, as in a single department of surgery, regional as with the Lothian Audit, national, like CEPOD or even on an international scale. The level, or breadth of participation has been described as inversely proportional to the accuracy and usefulness of the audit (Ruckley 1984). This is because it is possible to collect more detailed information in a more thorough way from fewer surgeons, the results can be seen after a much shorter time and so more effective feedback can be instituted. However, the use of large data sets collected nationally can show up differences in, for example, mortality rates which could not be spotted on a smaller scale. In any case, a decision must be made about which information to collect. All experienced practitioners of audit would recommend keeping it as simple as possible to minimise workload and maximise accuracy (Pettigrew *et al.* 1991). The usual method of collecting surgical audit data is for surgeons to fill in a form for each patient, the information from which is fed into a computer. To illustrate this, we consider the following examples, where a national audit, a local audit of General Surgery and an audit of a specific disease are described in more detail. We shall return to these audits in proceeding chapters, where we shall investigate data from them and discuss the methods of collection and analysis.

1.5.1 The Royal College of Surgeons Comparative Audit

The Royal College of Surgeons of England started its Comparative Audit service in 1990 (Emberton *et al.* 1991). Forms are sent out to consultant surgeons all over the country asking for data on their practice for the year. These involve sections on resources such as staff and beds, workload, i.e. numbers of operations and admissions, and clinical data which includes patient ages and diagnoses, operations and outcomes. There are also specific clinical enquiries into two different procedures each year. These have low response rates, but give very useful insights into current treatment methods. The information is requested in the form of total numbers in a particular category over the year so there is no individual patient data. This means we can not, for example, tabulate age by complications. There are many missing values due to the differing stages of development of computer data collection between hospitals. Not all hospitals have the available equipment to record data so participation in the scheme is voluntary. This ensures that the data are more accurate as the involved consultants will be interested in the results, and there is little incentive to falsify information.

The idea is for surgeons to gauge how well they are performing compared with their peers, and to see how local practice differs from that in other areas. Each consultant is assigned a confidential number and the data are mostly displayed as histograms in rank order. The consultant can then see where he lies in relation to everyone else. If for example he sees that his rates of a certain complication are particularly high he can go back and investigate the reason for this by looking at his rankings on other factors, which will either explain the adverse result or stimulate an investigation into practice to find the cause, which can then hopefully be rectified. In chapter 4, we will introduce new ways of displaying the data which remove the need for such subjectivity.

The comparative audit is designed to be informative and does not seek to incriminate people. The fact that the numbers are not traceable to individual consultants means there is no incentive to misquote any figures. Since only raw data are used, individual mortality rates, say, must remain confidential, as they could be open to misinterpretation by the public. If the rates could be standardised they would be more useful and would give a truer picture of performance (Dunn & Fowler 1992).

1.5.2 Audit at Glasgow Royal Infirmary

The University Department of Surgery consists of two firms, each with two consultants, working in General Surgery. Information on diagnoses, procedures and outcomes as well as patient details is entered onto a form by a junior surgeon and then the medical secretary transfers it to the computer audit system. The data are used to produce GP letters and discharge documents, as well as producing the required weekly reports. These list all the discharges and deaths for the week for each firm, giving patient name, age and length of stay and their diagnoses, procedures and complications. The lists are then discussed in the weekly audit meeting to check for accuracy of the information and to pick out any cases with unusual outcomes, or those where there is controversy as to the appropriateness of the procedure, for further discussion at the monthly meeting. The secretary notes which patients should be highlighted and also any data omissions or errors. The monthly meeting then consists of a discussion of the previously selected cases. The process is educational and should lead to a consensus agreement as to whether a particular outcome could have been avoidable, and perhaps stimulate research projects. The third type of audit meeting is the annual one, which should consist of a presentation of summary statistics for various operations. This could include information on numbers and ages of patients, lengths of stay and complication and death rates broken down into avoidable and unavoidable. Information can then be compared with previous years or other units to spot trends or problem areas. These can then be investigated and appropriate adjustments made to clinical practice.

We shall explore data collected from GRI later, in order to gain more insight into the problems of collecting and utilising audit data.

1.5.3 The Colorectal Cancer Study

This study is funded by the Clinical Resource and Audit Group (CRAG), which is a branch of the Chief Scientist's Office. We will thus refer to it as the CRAG Study. It involves several consultants in Lothian and the West of Scotland, so is a large regional audit of a specific disease. Each of the two areas collate their data separately, although a standard audit form has been developed to use with every patient. In auditing a particular disease, more specific details can be collected than with General Surgery. Thus with this study, we have information on the histology of the tumour and other variables which are known to influence survival from colorectal cancer. A disadvantage of studying a specific area is that it takes longer to accumulate substantial amounts of data than with a general study. The purpose of this audit is to monitor the outcome from resection of colorectal cancer, and to compare consultants and hospitals. This will allow procedures used by those having the most successful results to be recommended as standard. The relationship of workload to outcome is also of interest. Some results of this study have been published (Consultant Surgeons and Pathologists of the Lothian and Borders Health Boards, 1995)

In chapter 6, we shall describe the data, and make comparisons between the consultants. We would expect to achieve better models from these data than from the RCS data as they are at patient level rather than consultant, and because they are from a specific disease rather than from general surgery.

1.6 Computer Systems for Surgical Audit

In the last decade, there have been many software packages developed specifically for clinical audit. There are three basic approaches to the set up of computers for audit. The first is to have the audit computers completely separate from any administrative or other information systems. The benefits of this are that it is cheap, and its introduction will not interrupt the day to day running of the department. The second approach is to have an audit system which also carries out managerial and administrative tasks. These tasks include scheduling and producing patient discharge letters and other documents. This type of package is the most common, with most commercially available packages falling into this category. These packages include *Micromed* (Medical Systems Ltd), *Metabase* (Metasa), *Proton* (Clinical Computing Ltd.) and *Clinics* (ICS). These are more likely to gather a comprehensive data set than the stand-alone type of system, as they are part of the administration of the department. The third, and most efficient way of running an audit is to integrate it with the overall Hospital Information System. This means there is no replication of data entry, and patient details are consistent throughout the hospital. However, it is expensive and complicated to get such a system in working order. The ideal situation would be when all information was entered directly onto computer by clinicians at the time of treatment. This will remove the need to fill in forms, and thus obviate the data errors that ensue.

1.7 Some problems and limitations in surgical audit

1.7.1 Use of Resources

The audit procedure requires considerable resources, both human and financial, especially in the initial stages of setting up a system. Investment must be made in computer hardware and software, and perhaps in extra staff. Much thought must also be given to the organisation and format of the audit. As well as these, if the audit is to be successful, there must be sufficient enthusiasm for it.

There have been many data accumulated over recent years on surgical patients, on various microcomputer systems. In many cases these data have not been used to their full potential, if at all. With all this information available, it seems a great waste of effort and money not to exploit it. This work involves looking at some such data to see how much information can be gained retrospectively from databases, with a view to interpreting any methods for use on an ongoing basis.

1.7.2 Quality of data

The data are often incomplete and inaccurate but could still give some indication of trends in practice if explored. It is to be hoped that, with the introduction of more efficient computers and the growth in interest in audit, future data will be far more comprehensive. Not all surgeons are entirely enthusiastic about audit, and see it as an imposition on top of their workload. It is necessary that surgeons treat the process as important, otherwise the data collected is likely to be inaccurate (Ellis 1989). It should also be stressed that audit is an educational process and not a "witch hunting" one where poor results incur penalties. (Wilkin & McColl 1987). If an audit is approached confrontationally, doctors will be tempted to massage their figures for fear of recriminations. Entering the data into a computer from cards uses a considerable amount of secretarial time and so these cards should be filled in carefully. The coding of diagnoses and procedures has been a problem in surgical audit. Various systems exist such as the OPCS classifications and Read Clinical Coding. Some groups, for example the Lothian Audit found no codes detailed enough and invented their own (Gruer *et al.* 1986). Accuracy of the codes is also a problem. People may assign a general code rather than looking up the precise one, and severity of illness within a particular disease classification can differ. Data are collected in the form of "Consultant Episodes". This makes it difficult to obtain data on individual patients, and where patients are transferred within an admission, the number of episodes is increased and so the mortality rate appears lower. Record linkage is required so as outcome measures such as 5 year mortality can be ascertained more accurately.

1.7.3 Presentation of information

There is not much use made at present of graphics in displaying information on a regular basis. It would be favourable to use these more often in order that summaries could be digested at a glance and particular features of the data could be more easily spotted. Pictures would also be far more interesting to look at than the lists of figures which are presented currently at audit meetings.

1.7.4 Analysis

As previously mentioned, it has not been uncommon for comparisons to be made between surgeons or hospitals using raw mortality rates or complication rates. These are obviously unfair as bias will be intrinsic due to the non random way in which patients are allocated to a particular surgeon. There could be systematic differences in the severity and type of operations, age, social class and illness of the patients and so on. An object of this research is to take relevant factors into account, by using statistical modelling either to adjust the rates for them or to compare observed with predicted values. There have been many attempts over the years to predict surgical outcome. We shall review some of them from the point of view of use for audit in chapter 2. There has been relatively little work done on using risk indices as a tool in audit to compare actual numbers of a specific outcome with expected numbers, although a need has been recognised since the mid 1980's (Deans *et al.* 1987, Mortensen 1989). Most of the existing models are unsuitable for use in audit for various reasons which will be detailed later. If a good model can be found for a procedure that may give some representation of surgical skill, it could be incorporated into an existing audit package to give automatic adjusted values.

Surgery, it would seem, is an inherently unpredictable activity. If we consider an individual patient there are any number of unforeseen things that could go wrong. A statistical model is unlikely to be sensitive enough to predict individual outcomes, especially as there are always extreme cases. However, if a series of patients over a length of time is considered these things should balance out - the old person who makes a remarkable recovery and the teenager who has chronic illness - and the adjusted values can be expected to give a good idea of performance. It is important to realise that the whole health care process is so complex and that such a large number of immeasurable factors contribute towards outcome, that it is impossible to achieve total accuracy or even to say that results are completely accreditable to surgeons. Nonetheless, an attempt to adjust outcomes for contributory factors must surely be an improvement on the situation where people are allowed to judge by crude figures.

2 PREDICTIVE MODELS FOR AUDIT

2. 1 The need for models in audit

As stated in chapter 1, the aim of this thesis is to investigate ways to make comparisons between surgeons by studying outcomes. This has been done in the past by looking at raw mortality rates. In America, hospitals have been compared in this way, and resources allocated in the light of results. This is obviously an inappropriate method of assessment due to the large differences between the patient intakes of hospitals. This fact has been pointed out by many commentators. Poorer outcomes (for example, more patient deaths) may reflect that a more competent surgeon receives more complex patients, or it may reflect that a surgeon is not actually performing as well as he should. Many studies have shown up large differences between surgeons. An example of this is in Fielding's 1980 study of large bowel cancer, where the rate of anastomotic leakage after resection varied among surgeons from 0. 5% to 30%. In the colorectal cancer study by McArdle & Hole (1991), Cox proportional hazards regression analyses were carried out on their data, and they found that the hazard ratios of the surgeons in the study were significantly different even after adjusting for significant prognostic factors. For curative resections, for example, these adjusted rates varied from 0.56% to 2.03%. In a study of post-operative wound infection by Mishriki *et al.* in 1990, 'surgeon' was found to be a highly significant factor. Evidence of differences in surgical performance is abundant, but we require methods of quantifying these differences more objectively, by looking at data, and adjusting outcomes for case mix. The scoring of patients can help in evaluating quality of care if 'unexpected' deaths and survivals are considered (Schein 1988). Thus, it is necessary to use some form of predictive system to adjust mortality rates by comparing observed with expected. The properties of a suitable model for audit will now be explored, followed by a review of some of the systems already published, with a view to assessing whether they are appropriate for our purposes.

2.2 Properties required of a statistical model for audit

2.2.1 Well defined and relevant outcome measures

The definition of the outcome measure is a major problem in health care evaluation (Delamothe 1994). It could range from a rating of patient satisfaction or quality of life through various morbidity measurements to mortality. The most appropriate measurement depends strongly on the type of disease and surgical procedure being audited. For example, there is no point in using mortality rate as the outcome measure for operations on ingrown toenails or varicose veins, but it is suitable for an audit of colorectal cancer. In order to achieve reasonable predictions of any outcome, it must be well defined and suitable for the population being considered. Although death is not subject to the ambiguities of other outcome measures, it is not always clear-cut which measure of mortality is appropriate. Postoperative mortality could be examined, which gives an idea of immediate technical success. This could be defined as death in hospital or within 30 days of operation, but it should always be clearly stated. For comparative purposes, 30 day mortality is a fairer measure, as it has been shown that in-hospital mortality depends largely on local discharge policies (Jencks *et al.*, 1988). It may be more appropriate to study long term survival, say 5 years, as this will show the more important success of an operation, i.e. whether the patient had a substantial long term benefit. It is unclear how best to balance short term risk against the chance of a long term cure, and the shape of the entire survival curve is relevant to the evaluation of the outcome.

Definitions become even more difficult when they involve "softer" events such as complications. For example, different practitioners vary in what they would define as a wound infection. A scrupulous surgeon might record very minor signs of infection, and they would then appear to have a very high morbidity rate due to their conscientiousness. Complication rates can also reflect local discharge policies, as minor problems may go unrecorded if they develop after discharge from hospital.

To date, most well developed models have related to serious medical conditions where mortality is a relevant outcome measure. For example, severe head injury (Murray *et al.* 1986), intensive care (Lemeshow & Legall 1994) and surgical oncology (Deans *et al.* 1994). More research is required to establish standardised, objective measures of outcome in other less severe medical conditions if comparative audit is to be refined in these areas.

2.2.2 Easily obtainable and objective prognostic variables.

In order to make the predictions of outcome, a decision as to which variables should be included in the statistical model must be made. They should be easily measured and not subject to observer bias. For use in audit, it is essential that they should reflect the health status of the patient and not depend on their treatment. An extreme case of this can be seen in the various reports that the best predictor of outcome is the surgeon's post-operative evaluation of the technical success of the operation. For example, Baker *et al.* (1982), Pettigrew & Hill (1986), Pettigrew, Burns and Carter (1987), Hirshberg & Adar (1990), Oguz *et al.* (1990) and Hartley & Sagar (1994). This is not only highly subjective, but could also mask incompetence. Following a technical disaster, the surgeon predicts a poor outcome, and not surprisingly the outcome is as predicted! Brenner *et al.* (1989), however, found that only the most experienced surgeon in their study could predict more accurately than the simple scoring system that they used. Whether or not this is true, for audit we need objective estimates of surgical risk in order that fair comparisons can be made, avoiding bias, whether conscious or unconscious. For example, any consultant who wished to make himself look superior could simply record poor prognoses for all his patients. A more subtle manifestation of this can be seen in McArdle and Hole's paper of 1991 on colorectal cancer. A key factor in determining outcome was whether the resection was palliative or curative. A curative resection indicates that the tumour was not too far spread and that the patient's health was not critical, and as such is a strong prognostic factor. On the other hand, the type of operation is at the discretion of the operating surgeon, and could mask differences in skill. One could envisage a situation where a more adventurous surgeon would undertake a curative resection resulting in the expected good outcome, where a more "timid" surgeon would choose a palliative one with less immediate risk, but poor long term outcome. Both surgeons could perform "as expected", but clearly this would not be a sensible comparison.

Some variables measured during the operation would be suitable for inclusion in a model, such as size or stage of a tumour. Others, for example blood loss or length of time under anaesthetic are more debatable. They could reflect the severity of the surgery and so be relevant for the prediction of post operative morbidity. They could also be related to the competence of the surgeon and again would give the situation where observed performance is very close to expected. For this reason these types of observations should not be considered for inclusion in a prognostic index for audit.

A final example of the potential for bias relates to scores based on physiological measures. Often such scores are based on the most deranged value observed for each parameter over a certain period. With a good standard of care, a patient might be stable, whereas poor care may result in more variability. With such variability, the most deranged value observed will be more extreme than for a stable patient, and again the system inappropriately compensates for poor management. (Boyd & Grounds, 1993).

2.2.3 Calibration and Discriminatory Power

A statistical model for predicting outcome will depend to some extent on the intended type of audit. The purpose we are mostly concerned with is to use the model with a series of patients to compare surgeons or units with each other. In this case, the calibration of the model is of great importance. That is, that calculated probabilities of death, say, correspond to the actual probability of a patient dying. Thus when the predicted probabilities are averaged over a series of patients, this corresponds to the actual mortality rate, so that observed and expected rates can be compared.

The other application is to highlight individual patients for discussion, whose outcome is not as would be expected. A poorly calibrated model can still give a good idea of the relative probabilities of death and enable the patients to be ranked in order of survival prospects. This can then be employed in selecting individual cases meriting further discussion. An arbitrary cut off level can be chosen for risk, below which any unfavourable outcomes can be pinpointed, or above which any unexpected favourable outcomes occurred. A model such as this must have good discriminatory power, i.e. high sensitivity and specificity. We require a method of predicting what should happen so that performance can be assessed by comparing actual and predicted outcomes, and thus instigate investigation into the causes of any discrepancies. As stated above, accurate probabilities are not required for the purpose of selecting unexpected outcomes. It is necessary, however, to quantify the risk in some way so as to avoid subjective judgements as described above.

2.3 *Statistical approaches to modelling general surgery*

There are many statistical techniques that have been used to develop prognostic systems. The simplest of these involve using scores as predictors. Several factors are given scores for severity and summed to give a total risk score. The individual scores are often derived subjectively, using clinical knowledge of the important factors and their relative contributions to the risk. For example, the Fitness Score (Playforth *et al.*, 1987) consisted of a sum of scores for four different factors, each with possible values from 0 to 4 depending on severity. This has been used in Scarborough, where patients who die with a score of less than 6 are picked out for discussion. The Fitness Score is not calibrated so it is not possible to use it to make comparisons in a series of patients. Other scores have been derived from the weights in regression equations, for example the physiological and operative severity scores which make up the POSSUM scoring system (Copeland *et al.*, 1991). These consist of several factors which are given scores of 1, 2, 4 or 8 according to strict guidelines. Other types of systems take the actual values of variables such as age and physiological measurements, and combine them directly in a regression equation or in discriminant analysis. A classic example of this type of model was the Prognostic Nutritional Index of Buzby *et al.* (1980). This model gave risk of complication as a function of four nutrition related variables, for the specific area of gastrointestinal surgery. The variables to be included in the model are usually selected by a stepwise method.

The majority of work in modelling outcomes has been done using multivariate discriminant analysis or logistic regression. Many of the techniques that can be used were reviewed by Titterington *et al.* (1981). Since then more modern, computer intensive methods such as neural nets have been developed (Ripley 1994), which appear promising for prediction of outcome. Connectionist models have been compared favourably to traditional statistical techniques in Intensive Care (Buchman *et al.* 1994) and have successfully been employed predicting outcomes after liver transplantation (Doyle *et al.* 1994). A disadvantage of using neural networks is that they are "black box" systems, and the exact relationships between outcome and prognostic variables cannot be determined. This means that it is not known which pieces of information are important, and could lead to wasted effort in collecting large amounts of data. An improvement in discrimination is traded for loss of insight (Hart & Wyatt, 1990). Most reports on predictive systems evaluate them using sensitivity analysis, showing Receiver Operator Characteristic curves. However, very few published reports evaluate the calibration of predictive systems, which is important if one wishes to compare the expected frequency of a particular outcome over a series of patients.

2. 4 Review of predictive models in medicine

Many attempts to model the outcome of surgery have been made over the years, and we shall not endeavour to describe them all here. This review is a summary of the major areas of modelling, including some of the most important and interesting publications, assessing their suitability for our purposes.

2.4.1 Models in intensive care

Much work in clinical prediction has been done in the area of critical care. This area is easier to model than general surgery for two reasons. Firstly, the outcome measure of death is more relevant as it is more common. Also, critically ill patients have more extreme symptoms which make powerful prognostic factors. Aside from these, the data are likely to be more complete as the patients are constantly monitored.

Extremely accurate predictions were made in studies of severe head injuries (Murray 1986). These facilitated comparisons between international centres which used different methods and had very different mixes of patients. The centres also had significantly different rates of survival. Using a predictive model calculated from one centre's data on another centre's patient mix, the exact number of deaths observed in that centre was predicted. The system works so well because the prognostic variables for recovery from coma are known to be very powerful, and the study population is well defined. Also, the short term mortality rate is approximately 50%, making this a relevant, sensitive and objective outcome measure. The severe head injury model is one of the few areas where good calibration has been demonstrated (Murray *et al.*, 1986), and where the model has been evaluated formally as a decision aid in the clinical context. (Murray *et al.*, 1993). There are other models for prediction of outcome in intensive care where the condition is very specific. For example, the Abdominal Trauma Index is a score which predicts intra-abdominal sepsis based on which organs in the abdomen have been injured. (Moore *et al.*, 1981).

The most widely used system in intensive care is APACHE II, which stratifies severity of disease (Knaus *et al.*, 1985). It is calculated using scores for abnormal measures of twelve physiological variables, together with age and a score involving the patient's history of chronic health problems. The risk of death is then calculated using a logistic regression equation. This has weights for diagnostic category, as death rates were found to differ between different diseases, and also a term for whether the treatment was emergency. For series of scores, the system is well calibrated and could be of use for comparing different units by looking at expected and observed death rates. This was done by Knaus in 1988, when he compared ratios of observed and expected mortality rates at 13 different hospitals and showed up large differences which were not evident from crude rates. The original version of APACHE was based on 34 physiological variables, and was a subjectively assigned scoring mechanism. (Knaus *et al.*, 1981). This was too complicated, as even in an intensive care unit this amount of data cannot reasonably be routinely recorded, and was never validated in different medical centres. The more recently developed APACHE III has more statistical accuracy, as weights were calculated by regression techniques rather than being subjectively assigned, and can be used to calculate risk estimates for individual patients (Knaus *et al.*, 1991). In APACHE III the probability of death given a particular score is dependent on the diagnosis, and these are tabulated in detail.

A problem of modelling intensive care is that one is working with a dynamic problem with no "Time Zero" to act as reference. This problem does not arise with the head injuries studies, as time of injury gives a definite starting point to which subsequent assessments can be related. Timing of entry to the unit depends on local resource provision for ICU's, so patients are not measured at the same point in their treatment. APACHE scores can be measured on entry to the unit, but usually the most severe score within 24 hours is taken. A small, under funded unit will tend to receive patients with more advanced illness, and thus more deranged physiology than a larger, well resourced one, where patients may be admitted sooner. If one assumes that early admission is crucial in terms of the patients' prospects of survival, then the differences between the units will be masked. The large unit will receive fitter patients who will do well, as expected. The small unit will receive more critical patients, who will do badly, as expected. Alternatively, if admission to the ICU actually confers little benefit in terms of survival prospects, an artificial difference will emerge. The large unit, which admits patients with favourable APACHE scores, will achieve the same outcomes as the small one which admits patients with poorer values. In either situation, organisational differences cause a bias which makes the "adjusted" comparisons highly misleading.

It has been pointed out (Boyd & Grounds, 1993) that, since the most extreme values within 24 hours of admission are usually used in calculation of an APACHE type score, quality of treatment could have an effect on the score. Thus a patient receiving good, appropriate treatment could have a lower score than a similar one receiving poor inappropriate treatment. These two patients could then have the same standardised mortalities despite the difference in quality of treatment. Thus, the authors advocate that these types of models which use physiological measurements are not suitable for audit.

Accurate predictions of outcome for individual patients were gained from the Continuous APACHE Score which basically smoothes out the daily time series of APACHE values in an ICU. (Moser *et al.*, 1989). This idea is good for intensive care, but could not practically be applied to general surgery as it involves too many measurements. Also, admission to ICU is during hospital treatment and not at the start, so modelling is quite different. There is an interest in the progress of the patient, rather than merely the effect of an intervention as in surgery. With general surgery we are dealing with pre-treatment measurements and so a continuous model would not be suitable.

2.4.2 General Surgery

We have seen that good predictive models are achievable in critical care medicine, although there are still problems with bias. Their overall success is promising, although it should be kept in mind that general surgery is very different from critical care. There is less difficulty with timing as we are interested in the result of a well defined intervention, although even here local organisational policies may be relevant. For example, a conservatively managed patient might deteriorate and require emergency surgery, when a more timely surgical intervention might have been appropriate. Preoperative predictions under either management strategy are likely to match outcome, and so, again, the comparison of adjusted outcomes could mask a difference in quality of care. Some of the predictive systems in use in Intensive Care are fairly complicated. This is acceptable in that field, as constant measurements are made, and it is a highly staffed area. For general surgery, simpler models are required. As stated previously, there are problems with general surgery in that outcomes are not so well defined, preoperative measurements less extreme and data more scarce. Even so, there have been many attempts to make predictions in general surgery for various reasons, but most often with treatment interventions in mind.

One of the earliest systems was to classify patients into five groups according to physical status. The American Society of Anaesthesiologists (ASA) system was devised in 1941 (Saklad 1941) and was subsequently updated (ASA 1963). The classes range from I (healthy) to V (moribund). They depend entirely on subjective observations by physicians and as such are fairly arbitrary. The relationship between these classes and postoperative mortality was shown in a very large study by Vacanti in 1970. The classifications are still often used, as an indicator of preoperative health status.

In 1977, Goldman *et al.* produced their index to determine risk of a cardiac complication or death from any non cardiac surgery. A discriminant analysis was carried out to find the significant variables, and a score calculated for use in treatment decisions. For audit, it may be of interest to only consider a specific type of morbidity, such as cardiac, but this should be related to a specific procedure. In this example, looking at the specific complication would not give a good measurement of the overall quality of care, but one could envisage circumstances where it would be a more sensitive measure.

In the early 1980's there was much discussion in the literature about the prognostic value of variables associated with nutrition. These can be biochemical, anthropometric or clinical measurements. The correlation between nutritional status and outcome of surgery has been investigated many times (Dempsey *et al.*, 1988). It is generally accepted that the relationship exists, although there have been disagreements (Ryan & Taft, 1980). In the cited paper by Ryan and Taft, only univariate comparisons were made between patients suffering postoperative complications and those who were not. The variables used were thought to represent nutritional status, but no actual nutritional assessment was made. It is not known whether the relationship is causal, or if the symptoms of malnutrition merely reflect the advanced state of the disease. This is not important for audit purposes. For example, if serum albumin is found to be a powerful predictive factor, it does not matter why its level is low, and if anergy is useful for prediction of outcome it is of no consequence to the results whether the disease suppressed the immune system or the patient was immunodeficient and thus more susceptible to disease.

The main work in prediction involved with nutrition was the calculation of the Prognostic Nutritional Index or PNI (Buzby *et al.*, 1980). This model gave risk of complication as a function of four nutrition related variables, for the specific area of gastrointestinal surgery. It gave good predictions in the prospective trial carried out in the same hospital unit, with a sensitivity of 93% for prediction of mortality. The outcome measures were well defined and objective and the model may be suitable for the categorisation of patients into high or low risk for the purpose of highlighting unexpected outcomes. It does not appear to be very well calibrated, so would not be useful in predicting actual numbers of deaths. It is also based on a specific type of patient, that is those who were considered 'nutritionally deficient', although no rigid definition of this is given. In a trial in Germany (Kohler *et al.*, 1988), the PNI did not seem useful for calculation of risk, although they used different outcome definitions for complications from the original study, and their patient population had a higher proportion of cancer patients. This suggests using presence of cancer as a predictive variable. It has also been suggested that age be included as a prognostic factor in an index such as the PNI (Warnold & Lundholm, 1984).

Another similar index, with coefficients calculated by discriminant analysis was developed by Harvey *et al.* in 1981. This included a variable for presence or absence of cancer. As with the PNI, it was developed on a specific, critically ill group of patients and considers risk of complications and mortality.

Much of the controversy in this field has been around the suitability of particular types of variable for prediction of outcome. Some authors have claimed that morbidity is strongly related to anthropometric measurements (Klidjian *et al.*, 1982) and others have concluded that they are not at all related (Ramsay *et al.*, 1986 and Pettigrew & Hill, 1986). Even more contentious was skin testing with antigens (Meakins *et al.*, 1980; Christou *et al.*, 1981; Ausobsky *et al.*, 1982; Ottow *et al.*, 1984; Schackert *et al.*, 1986). The inclusion of anergy in a predictive index is not very practical from a clinical point of view, as it is inconvenient to carry out and thus not easily adopted into routine practice.

An easily calculable index involving only two variables was developed in Glasgow (Ramsay *et al.*, 1986). The risk is of complication or death where the patients have undergone laparotomy, and is a function of age and lymphocyte count. The outcome measures are not rigidly defined and so accurate calibration is not likely. An index of this sort which had more predictive value would be ideal for audit as it is simple to calculate.

Another prognostic index for survival considers only surgical patients aged 80 or over. (Krenzien *et al.*, 1989). The variables used are all dichotomous, with the exception of age which is grouped into five categories. When a prospective trial was carried out to test the model, two of the original variables were no longer significant. The predictive accuracy of the index is lower when it takes a high value.

The indices discussed so far were all calculated using discriminant techniques and validated using sensitivity and specificity analysis, as well as in prospective trials. Many other studies have used regression analysis in order to find the factors with the most significant effect on outcome, and thus produced linear predictors. Pederson *et al.* (1990) were mainly interested in the effect of anaesthetic on mortality, but naturally this cannot be separated from the surgery effect. Their multifactorial risk index can be used to estimate whether an individual's risk in general surgery is high or low. With an overall mortality rate of only 1.2% in the study population, this outcome measure is too rare to model accurately.

The analysis of survival of patients after resection for large bowel cancer (Chapuis *et al.*, 1985) predicts outcome in a very specific area. Most of the significant variables were measured during operation and were related to size and spread of the tumour. This type of survival analysis unfortunately demands more follow-up information than is routinely collected. However, the use of an area where mortality is more common means that it would be easier to model were the data available.

A study where logistic regression analysis was used on a specific population was on total gastrectomies for stomach tumours (Miholic *et al.*, 1988). The outcome being considered was again mortality. Since the surgical field was so specific, the number of patients in the study was rather small (98). The fit was tested by counting the number of patients correctly classified as dead or alive. The numbers were quite high, but a probability of 0.1 was used as the cut off point, suggesting lack of calibration. The area is too small to achieve large enough numbers, and the model would not necessarily transfer to other hospitals.

The effect of psychological variables on postoperative length of stay was investigated by Boeke *et al.* in 1991. These include assessments of feelings of anxiety and inadequacy. The index involves variables measured after operation so is not suitable for our purposes. Using length of stay as an outcome measure can be an inaccurate measure of surgical success as it can depend heavily on social factors.

Regression analyses were also carried out by Mishriki *et al.* in 1990, to find factors associated with their chosen outcome of wound infection. This outcome was defined very precisely. They found that 'operating surgeon' was a significant factor, even when included with other explanatory variables. This type of study could be useful for an audit of wound infections, but for overall quality of surgical treatment one would wish to cover all types of morbidity.

The POSSUM (Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity) scoring system (Copeland *et al.*, 1991) was specifically developed for use in audit, and has been adopted in several hospitals throughout the UK. It is fairly complex in that there are two scores - physiological and operative severity. Separate risks for mortality and morbidity are calculated from these scores, by logistic regression. The physiological score involves twelve factors, some of which are not routinely measured or recorded for all patients undergoing general surgery. It is also possible that these factors could be affected by treatment. The operative severity score consists of 6 items. Two of these are volume of blood lost in millilitres and peritoneal soiling, which may be a function of surgical ability as well as the technical difficulty of the operation. The models are fairly well calibrated, especially for mortality, when tested in the hospital of origin, although for low risk procedures the predictions are not good. In fact, the minimum possible value for the risk of death obtainable from the published POSSUM equation is 1.08%. This is very high for minor procedures, and thus leads to overestimation of the death rate (Whiteley *et al.*, 1995). This means that any general surgeon, working on a large number of low risk cases will have too many deaths predicted by the POSSUM system, leading to an unrealistically favourable picture of their performance. This is another potential bias which could mask poor performance.

2.5 Discussion

There have been many attempts over the years to predict outcome from surgery - some more successful than others. The main ones discussed in this chapter are summarised in table 1. Each system is rated according to its properties as required for use in audit, with three stars being good and one star poor.

For an optimistic view on surgical prediction see Knaus (1988) who pointed out that a few hundred years ago temperature was not seen as a quantifiable entity. Consider the following three factors that determine outcome:

- The patient's ability to withstand the operation
- The severity of the disease and procedure
- The surgeon's skill (Playforth *et al.*, 1987)

If we predict outcome using data about the patient, while either adjusting for the second factor by including severity in the model, or by keeping it constant by considering one disease or operation at a time, then comparing observed and predicted outcomes should leave only an estimate of surgical skill.

It is important to avoid making comparisons using crude league tables, and rather to adjust the outcome rates in some way. As we have seen, this is not an easy task, and there is great danger of adjusted values actually masking real differences in quality of care.

Possession of required property ★ poor ★★fair ★★★ good

System	area	type/ method	PREDICTORS		OUTCOMES		calibration	accuracy	Overall suitability for audit
			measurement	objectivity	relevant	definitions			
APACHE Knaus 1981	Intensive Care	Complicated, subjective score.	★	★	N/A	N/A	N/A	N/A	★
APACHE II Knaus 1985	Intensive Care	Subjective scores assigned. Risk from LR equation.	★	★★	★★★	★★★	★★	★★	★
APACHE III Knaus 1991	Intensive Care	Score, weights calculated statistically and used in LR.	★	★★★	★★★	★★★	★★★	★★	★
HEAD INJURIES Murray 1986	Critical Care	Scores used in model.	★★	★★★	★★★	★★★	★★★	★★★	★
CARDIAC RISK Goldman 1977	(Non-cardiac) surgery	Discriminant function	★	★★	★	★★	★	★★	★ Complicated. Outcome too specific.
ASA 1963	general	Classification into 5 groups	★	★	★	★	N/A	N/A	★
PNI Buzby 1980	Gastrointestinal surgery	Discriminant function	★	★★	★	★	★	★★	★ Wide range of outcomes on small patient subset.
HOSPITAL INDEX Harvey 1981	general	Discriminant function	★	★★	★	★	★	★	★ Wide range of outcomes on small patient subset.
RISK INDEX Ramsay 1988	Laparotomy	Discriminant function	★★★	★★★	★★	★	★	★	★ Wide range of outcomes. Lacks accuracy.
OCTOGENARIANS Krenzien 1989	general surgery	Discriminant function gives score.	★	★	★★	★★	★	★★	★ Very specific population (over 80, trauma patients)
POINT SYSTEM Brenner 1989	surgery (cancer)	Score statistically calculated	★	★★	★	★	★	★	★
PAFS Playforth 1987	abdominal surgery	Arbitrarily assigned scores	★	★★	★	★	N/A	★★	★ OK for studying individuals
POSSUM Copeland 1991	general surgery	Physiological (statistical) Op Severity (subjective)	★	★	★	★★	★★	★★	★★ Predictive power due to wide range of procedures Poor predictive ability for low mortality

Table 1: Summary of major works on surgical prediction and their suitability for audit use.

3 THE RCS COMPARATIVE AUDIT DATA

3.1 Introduction

The original Comparative Audit study was carried out by Medical Systems Limited, the originators of the MICROMED program. In mid 1990 they sent questionnaires to all users of the package asking for certain reports on surgical activity in 1989. The analysis included 45 general surgeons, covering 49 005 admissions. Each item was ranked and displayed in bar chart form. The work is described by Emberton *et al.* in their 1991 paper.

The objectives of the study were to gain an insight into current UK practice, to collect large amounts of data on an annual basis, and to allow local results to be compared with others. The study should also stimulate surgeons to collect data, and lead to some uniformity in data collection. Although it was never intended that this study should complete the audit cycle, it should result in quality improvement if consultants are motivated by the data to investigate why they are performing differently from their peers in any area.

In the above-mentioned paper it is stated that "Since no sampling takes place because every patient event is recorded there is little requirement for statistical analysis."

The data collection process involves only summary data from each consultant and individual patient information is lost. Thus we can not relate factors to each other, nor do we have an idea of within surgeon variation. This does limit the scope for statistical analysis, but it is still necessary if we want to make fair comparisons. The practice of ranking rates and proportions gives some indication of performances, but the statistical significance of any differences will obviously depend on the numbers of patients involved.

This initial study was published in the anticipation that Comparative Audit would become more widespread in the future, and that methods of making the comparisons could be refined to include some form of standardisation.

Following the initial work by Medical Systems Limited, the Royal College of Surgeons of England adopted the cause of the Comparative Audit. They circulated forms among their Fellows in 1991, requesting audit results for the year 1990, and received 160 completed ones out of the initial 1025 which were sent out. These data covered 147 882 admissions, although there were several values missing, and so different information was available for each surgeon.

The details of these data have been published (Dunn & Fowler 1992). The paper stresses that the size of the database compensates for any irregularities in data collection, and that the raw data can be considered reasonably to represent current practice. It seems likely, however, that the consultants who return data, who have been making an effort to gather information, will be among the most conscientious and that their responses will therefore not be representative of the country at large.

One cannot attribute low response rates solely to a lack of interest from surgeons, as a more major limitation is the unavailability of the required data in most units. This is due to the small number of computer systems in operation, and the inability to extract the specified information from a system even if an audit is in place. Those units using MICROMED were at an advantage, as the data collection forms were originally based on this system and so all the data can easily be obtained using specific built-in reports.

The data were presented to a meeting of interested consultants in June of 1991, in the form of ranked bar charts for each variable. The participating consultants had each been assigned a confidential number, by which they could recognise their data. At the meeting they each received a summary printout, telling them where they lay in each category so as they could compare their positions with others. They could then, say, compensate for the fact that they had one of the highest mortality rates with the knowledge that their proportion of emergency admissions was also one of the highest. This was very subjective, however, as they could not know of the relationships between variables for any other consultants. This procedure of collecting the previous year's audit data and presenting it at a meeting in June has since become an annual one.

After attending the Comparative Audit meeting in 1992, where the 1991 data were presented, it became apparent to us that there was much room for statistical input. The ranked charts gave an idea of the spread of totals, but no real conclusions could be drawn by anyone who had not personally contributed data. The organisers of the Comparative Audit Service had realised the requirement for some more analysis when we approached them with some suggestions of more effective presentation. We requested, and were given, the 1991 data, in order to work on the development of the methods and to explore their feasibility. Before the 1993 meeting, we acquired the 1992 data, and analysed it using the new procedures. The data were officially presented in ranked charts, as had become customary, but the new ideas were put to the consultants present, who approved of using them in the future.

There have been several drawbacks of the data, in that there are many missing values, there are only aggregated figures for each consultant, and some of the data are of poor quality as information has not been recorded for every patient admitted. We should hope that some improvements in data collection can be made in the future so that we can have more confidence in the conclusions drawn from any analyses.

The aim of this work, then, is to explore the particular form of data generated by this audit and to find more rigorous ways of displaying some of the important factors, for use at a meeting of consultants. Individual printouts for each consultant, giving a personal summary of data must also be designed. An important step in the presentation of these data is to take random variation into account and show confidence intervals for mortality, rather than just the rates as has been previous practice. We have explored the relationship between the "case mix" variables and mortality, in the hope of finding a model with which to adjust these outcome values, in order to make fairer comparisons than simply looking at raw mortality. It will be shown that this task is rather complicated and that no one model emerges to describe the data perfectly. The adjustments do, however, make a substantial difference to consultants' rankings and as such, at the very least, show the absurdity of making comparisons without any attempt at accounting for patient variation.

The data from 1991 and 1992 will now be described in more detail, and then, in the next chapter, some methods of analysis and presentation discussed.

3.2 The data

3.2.1 Collection

The data collection forms for the 1991 and 1992 data consisted of the following sections.

A Resources

1. Manpower: Percentage of time worked by each grade, e.g. if a registrar spends half his time working with a consultant he makes a contribution of 50% of a registrar to the unit.
2. Beds available
3. Theatre sessions

B Workload

1. Admissions: Total and numbers of emergency, elective and day case
2. Operations Severity defined by BUPA classifications (British United Provident Association. 1989)
3. Stay nights

C Clinical Data

1. Ages Number of patients aged in each decade for the 1991 exercise, collapsed to numbers aged 0 to 10, 11 to 60 and over 60 in 1992.
2. Diagnostic groups Number in each of 10 groups
3. Operative groups Number in each of 12 groups. In 1992 an "other procedures" section was also included.
4. (a) Complications and mortality
   (b) Specific complications

In the above Clinical Data section, it should be noted that the data on age and diagnoses were requested for all admissions, whereas the operative groups included only those patients who received surgery. This makes the modelling of postoperative outcomes difficult as there is no information on case mix of the operated patients.

There were also more specific clinical enquiries into Cholecystectomies and Abdominal Aortic Aneurysms in 1991, and into Colorectal Resections and Open and Laparoscopic Appendicectomies in 1992, as part of an ongoing scheme to study specific procedures in greater detail. Two procedures of interest are chosen each year and analysed in depth by a specialist in that field. It is planned to repeat these investigations intermittently so that comparisons can be made over the years. The data returned on these were very sparse and will not be considered here.

It should be mentioned that there was a lack of standardisation in the data collection, with some definitions open to subjective interpretation. Also there may have been wide variations in the rigour with which information such as the number of complications were documented. Even the recording of deaths was not standardised at those within 30 days, but was in-hospital mortality which could vary greatly due to discharge policies and the type of hospital. This must be kept in mind when making any assertions about the data.

3.2.2 The 1991 data

The original data file, as received in ASCII and PARADOX database form from the Royal College of Surgeons, consisted of 255 fields, but this was reduced to 138 when those regarding the specific enquiries were discarded. This was then trimmed down to a file with 83 variables by removing those which could be calculated from others. There were 215 records in the file, although the report produced for the meeting in June 1992 was on 209 surgeons. The discrepancy was caused by the addition of late arrivals to the database, and so an attempt to reproduce the means and other summary statistics given in the original report did not produce identical results.

It was observed that several consultants had exactly the same values for every variable, and some had fractions of numbers of patients. These were where the data had been submitted from a unit as a whole and then divided by the number of consultants. Since this gives no more information, these were re-merged to make use of the larger numbers of patients. Following this adjustment there were 199 records.

There were many values missing from these 199 consultants' data, some among potentially important variables. For our initial analyses, the outcome measure considered was overall mortality, so any consultant with no value for postoperative or non operative numbers of deaths had to be excluded. It was decided, also, to include only those consultants with complete data for certain variables which, as will become clearer later, are useful for modelling mortality. Table 2 shows the numbers of consultants left if we omit those with no data for these particular variables.

Another irregularity of the data was that many of the variables which gave numbers in separate groups did not correspond with the total number of admissions given for that consultant when summed over all the groups. This was especially noticeable for the numbers in the 10 diagnostic groups, none of which added to the total number of admissions. This is because the diagnostic groups used are not by any means comprehensive, and no consultants only treated patients who fitted into one of these categories. In fact, if we take the sum of the numbers of patients given in each diagnostic group as a percentage of the stated number of admissions for each consultant we find the values fairly evenly distributed around 65%, with a value as low as 3% and one as high as 127%. This shows that the fact that the groupings are not exhaustive does not account for all the discrepancies. It should be noted that the groupings were never intended to be comprehensive, but were chosen because they were of interest. The correspondences between the totals of the 10 age groups and numbers of admissions are better, with most values being fairly close to 100% and a mean of 93%. However there are even more extreme values than with the diagnoses, of 0% and 154%. Admission status has been most accurately recorded, with few totals being different from the stated number of admissions. The discrepancies may have been due to the incomplete recording of data, or they could have been due to mistakes in the initial hospital data collection, or misreading of figures in the data entry at the Royal College.

For analysis of proportions in each group, it had to be assumed that the distribution of those patients whose data were recorded was identical to that of those who were not, i.e.

proportion in group i of variable j= number in group i/Σ all groups of variable j

This assumption is probably fairly reasonable for admission status and age as the differences are mostly very low and the proportion will be very close to the true value. For the diagnostic groups, it is rather doubtful as most have totals much lower than the number of admissions, due to other diagnoses. We are assuming firstly that the spread of risk in unrecorded groups is the same as the spread in the ones which have been considered, and less importantly, that the chance of a patient's diagnosis not being recorded, and thus causing the total to be too low could be expected to be similar in each group so that the calculated proportions are representative. This is unlikely to be true, and this must be remembered when making inferences about diagnostic risk. These analyses are preliminary, and would obviously be refined in time. We would hope that data collection could be changed to incorporate all diagnoses, preferably based on severity rather than anatomy.

	Numbers left in study		
Variable	**1991**	**1992**	**both years**
total mortality	142	157	188
age	151	189	234
admission status	182	200	256
diagnostic groups	146	195	241
all the above	96	132	163

Table 2: Numbers of consultants remaining when those with missing values for particular variables are omitted from the study

The number of admissions per consultant in 1991 ranged from 126 to 7297, with a median value of 1160. Seven consultants had no data on number of admissions. The median total number of deaths was 24, ranging from 0 to 213, on average evenly spread between non- and post- operative fatalities. These gave an overall mortality rate of 2.2% with the lowest being 0% and the highest 5.6%, suggesting quite a difference between surgeons. Figures 1(a) and 1(b) show the totals over all consultants in each age group and diagnostic group respectively, to illustrate the overall spread. The number of patients aged 0 to 10, for example, ranged from 0 to 377 (median 55.5), whereas those in the middle age group numbered 64 to 4404 (median 578.5) and those over 60 from 56 to 2858 (median 440). It seems that 10 is perhaps too low a cut off point for the lower age group.

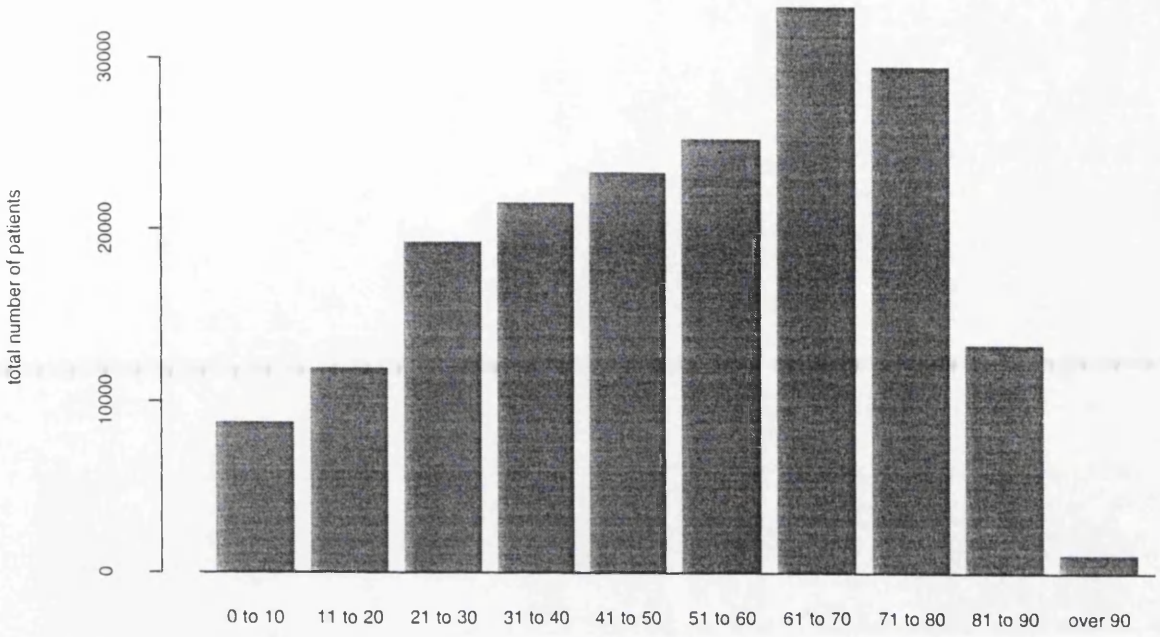


Figure 1(a): Spread of Age Groups (1991 RCS Data)

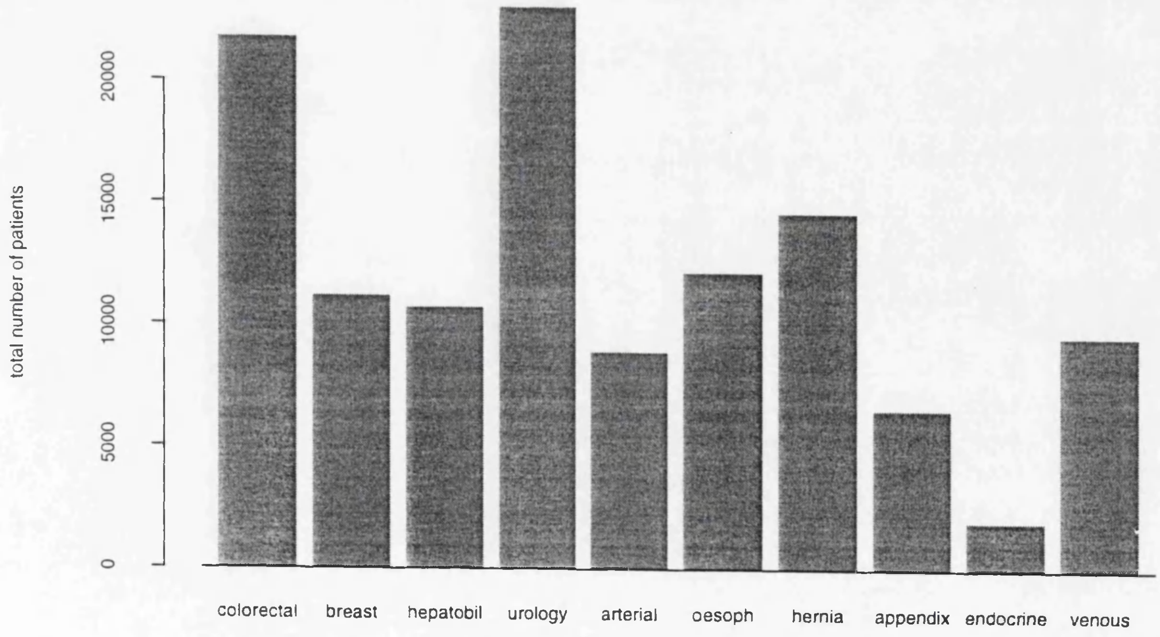


Figure 1(b): Spread of Diagnostic Groups (1991 RCS Data)

3.2.3 The 1992 data

Similar data forms as were sent out for the 1991 data were circulated in 1993 to collect information on practice in 1992. 1004 of these were sent out, and replies received from only 208 consultants. The main difference between the pro forma for 1991 and 1992 was that data were only requested for three age groups instead of ten. Only the numbers of patients aged under 10, 10 to 60 and over 60 were requested in 1992, since they were the only details presented in bar charts the previous year. Also more information on outpatients was gathered.

After excluding those consultants with missing values for the required variables (mortality, age, diagnoses and admission status) there were 132 remaining so the quality of data appeared to have improved slightly from the previous year (see table 2). One consultant was omitted because there was no value for number of patients with endocrine diagnoses, although the rest of the information was complete. The other options would have been to include the consultant and assume there were none in that category, or to assume the number was the difference between the total number of admissions and the total of the other diagnostic groups. As before, many of the totals for the age groups and diagnostic groups did not add up to the total numbers of admissions so the same assumption was required. The proportion of admissions accounted for by the sum of the age groups had a median value of 100%, whereas the median for diagnoses was only 63%.

The data were slightly different from the 1991 ones in that less consultant firms contributed their total results, but submitted them for individual consultants. This meant that the range of number of admissions was smaller (137 to 2776). The median number was 1260, and there was no information for 5 out of the 208 consultants. The mean mortality rate for the 157 with available data was 2.0%, ranging from 0.0% to 6.0%. The diagnostic spread was similar to before, and again there were very few patients in the youngest age group, with most in the middle group.

3.2.4 The 1991 and 1992 data together

The data files for 1991 and 1992 were reduced to compatible files with the columns in the same order, then stacked and sorted in order of consultant number. Since several consultants had data for both years these had to be added together. This was done using a simple FORTRAN program. After those with missing values for any of the variables considered to be important were deleted there were 163 consultants remaining. The breakdown of this can again be seen in table 2.

3.3 Case Study: Glasgow Royal Infirmary

We have seen that there are many problems with the RCS data. In order to gain an understanding of why this is so, we will now study the audit data collected for 3 years from August 1990 in the University Department of Surgery at Glasgow Royal Infirmary (GRI). These data give an invaluable insight into the problems of collecting audit data in a hospital, and highlight some reasons for the poor quality of the data collected by the Royal College of Surgeons. There is a long way to go in the standardisation of data collection in hospitals before a national comparative audit will be reliable. However, the process is very worthwhile, even if it only serves to bring the large differences in data integrity to the attention of those concerned.

The GRI data were retrieved from a now obsolete audit system, which was based on the SQL database Oracle 5. There were four copies of the system, one for each consultant's secretary. These had to be individually restored onto a computer, and the relevant data base files located and extracted. The files were copied into Paradox using the dbms/copy package. There were several technical problems with this as many of the files were too large, and so had to be manipulated in Oracle. A large proportion (approaching half) of the overall patient data was duplicated several times. The database after all four secretaries' data were combined initially had almost 14000 records. After the repeat entries and those with no information had been excluded, there were data on 7435 admissions. Selected variables were then transferred to an Excel worksheet and then analysed using Minitab.

A problem with these data is that a substantial number of patients were admitted several times, usually for the same complaint. This means that there is a lack of independence in the data. It would perhaps be preferable to consider patient episodes, where a succession of admissions for the same problem would count as one treatment episode. In this way, only the final outcome would be considered, and a more accurate reflection of the treatment's success would be achieved. These episodes would be difficult to determine from the current data. In some cases the same patient was assigned more than one identification number. Where this has obviously led to the same admission being recorded more than once, the extra ones were deleted. It is, however, not always possible to tell, so there could have been duplicates remaining in the data set.

There are 4 consultants in the team at GRI, and their total numbers of admissions can be seen in table 3. These numbers of admissions are the numbers which are available on the database, and may not reflect the actual numbers of patients treated during the time period, due to the phasing in of the audit system.

consultant	# admissions
1	1547
2	2250
3	1610
4	2020

Table 3: Number of patient admissions by consultant - GRI data

The data were originally collected on a card which was designed for use with the Micromed system. One of them can be seen in Appendix 1. These cards were filled in on admission and the data entered into the computer. The records were then updated on discharge. It seems that this updating did not occur on every occasion, as much information is missing, and the diagnoses are often vague, for example 'abdominal pain' or 'old age'. There are many missing values, due to the data not being entered into the computer, or perhaps never being recorded on the form in the first place. Table 4 gives the total numbers of missing values for some important variables.

Variable	Number missing (% of admissions)
Sex	14 (0.2)
Age	120 (1.6)
Admission Priority	1778 (24)
Admission From	5 (0.1)
Discharge Code	163 (2.2)
Diagnostic Code (Read)	886 (12)

Table 4: Numbers of missing values in GRI Data

3.3.1 Production of data required by RCS Comparative Audit

In an effort to explore the manageability of the data, and to give a feeling for the problems faced by hospitals supplying data to the RCS, it was attempted to rearrange the data so as to achieve the information requested by the RCS which are required for the presentation methods introduced in this thesis. This involved making several subjective judgements on categorisation of patients, which must also have been done by consultants around the country.

For example, the variable 'admission status' for the RCS data required numbers of day case, elective and emergency admissions. The GRI database had two variables which could have been related to this. Firstly, 'admission priority' which was categorised into routine, soon or urgent and secondly 'admission source' which consisted of 8 categories. These were day case, GP, inpatient, outpatient, re-booked, planned, Accident & Emergency and other. Each of these contains useful information, but admission priority contained many missing values. The method decided upon for classifying patients into the RCS categories is summarised in table 5. This gave only 3 missing values for status overall. These definitions gave in total 623 day cases, 4382 elective admissions and 2427 emergencies. These figures are not incongruous with those received by the RCS for consultants in England and Wales.

		Admission From			
	status	Day Case	others	A&E	missing
Admission Priority	Routine	DC	elective	elective	elective
	Soon	DC	elective	emergency	elective
	Urgent	DC	emergency	emergency	emergency
	missing	DC	elective	emergency	emergency

Table 5: Definitions of Admission Status from GRI data

Another problem arose with diagnostic categories. Many of the common diagnoses do not belong to a group as defined by the RCS. The groups originally chosen by them were those thought to be "of some interest", and are by no means comprehensive. Thus many patients are overlooked when diagnostic risk is included, for example, those with pulmonary or lymphatic disorders. The risk groups considered in the RCS study are likely to be highly inaccurate, as the types of diagnoses not included are important and could seriously affect the balance of risk groups. The diagnostic groups selected by the RCS are based strictly anatomically, and do not easily correspond with the Read Codes, which have many more sections and separate out cancer. The RCS diagnostic groups include too wide a range of risks. For example colorectal diagnoses can include anything from haemorrhoids to terminal cancer. With the GRI data, as seen in table 4, 886 patients had no diagnostic code given. A further 936 of the codes had no diagnosis in the diagnostic classification table. Often the diagnosis could not be categorised, and was unrelated to the operation, for example 'old age', 'asthma' and 'confusion'. Some diagnoses were actually postoperative complications such as wound dehiscence or renal failure, and often occurred where no postoperative complication was recorded for that admission. This would make the analysis of morbidity rather difficult, as well as not providing diagnostic groups for several patients. In an attempt to produce the RCS diagnostic categories, new codes were assigned to each of the patients by looking at their data, including diagnosis and operation. These codes are as follows.

1 colorectal
2 breast
3 hepatobiliary/pancreatic
4 urological
5 arterial
6 oesophago-gastric
7 hernia
8 appendix
9 endocrine
10 venous
11 minor others
12 pulmonary
13 lymphatic

The first 10 are as used by the RCS, and the last three were created in order to include more of the patients. Category 11 includes nail and skin problems, and if no diagnosis was recorded, where the procedure was coded as minor. Category 12 contains lung and respiratory disorders, and 13 incorporates diseases of the lymph glands, including cancers. The codes assigned were based mainly on the diagnoses, but also on the operations if a specific part of the anatomy was not apparent from this. In the 936 cases where no medical term was available for a code, the nearest code with a diagnosis was assumed. New codes were assigned to 6943 of the patients. Some could not be categorised, including septic shock and alcohol withdrawal syndrome. Leukaemia also did not fall into any of the above categories. A major problem was in categorising the diagnosis of diabetes mellitus. In some cases this seemed to be irrelevant to the procedure carried out, for example a gastrointestinal one. If this was the case, it was assumed that the operation gave a clue as to the main reason for admission. If no operation was recorded, the diagnosis was coded as an endocrine one. These types of coding problems must occur in every hospital, and lead to patchy and unreliable data.

This exercise has served to highlight the fact that data should be collected with a specific purpose in mind if they are to be of use. Although these data were never intended to be included in such a study, they have given an insight as to why the data received by the Royal College of Surgeons are so poor. The majority of surgical units do not collect data specifically for the RCS study, and so cannot produce the information they require. Likewise, the RCS request a large amount of data, mostly without a specific reason. If such a national comparative study is to be extended, individual hospitals should perhaps add the required categories to their data set so that they can easily obtain the values at the end of the year, and the data collected should also be changed. We will discuss this further in chapter 5.

In this case data were collected for the weekly audit. This is an audit of in-patients, and so, while the data would have been correct at that time, it is likely that the final diagnosis for many patients did not reach the computer database. Several problems of using these data for a more long term analysis have been shown, for example mistakes in diagnosis coding, and missing values, as well as patients with different hospital numbers and duplicate entries.

3.4 Original RCS data presentation methods

The 1991 and 1992 data were presented at meetings of surgeons in June 1992 and June 1993 respectively. All items, except diagnostic and operative case mix variables which were shown as means and ranges, were shown as bar charts with values in rank order. Figures 2 and 3 show the types of bar chart display used. A consultant could identify himself on the charts, as he had a personalised confidential printout of his position for each one. This meant that an individual consultant could assess his relative performance by looking at his ranks for different variables. For example, if he found his postoperative complication rate was exceptionally high, he could compare other factors like proportion of emergencies or patients over 60 to see if his ranking in these could explain his adverse outcomes. The presentation has the advantage that no one can identify the results of an individual consultant, and that no conclusions can be drawn by any external observers. Statistics were also calculated for the "Average General Surgeon", but no mention of the amount of variation was given.

Although these presentation methods maintain absolute confidentiality, it seems that much more information could be gleaned from the data, for example analyses of relationships between variables and statistical assessments of outcome measures, taking into account the variation present. The next chapter deals with the data investigation and some new ways of showing the data more clearly and of actually making comparisons between consultants.

Total Mortality Rate 1992

Total Deaths Reported: 3,597 Mean: 1.93% (95% CI 1.83 to 2.17)

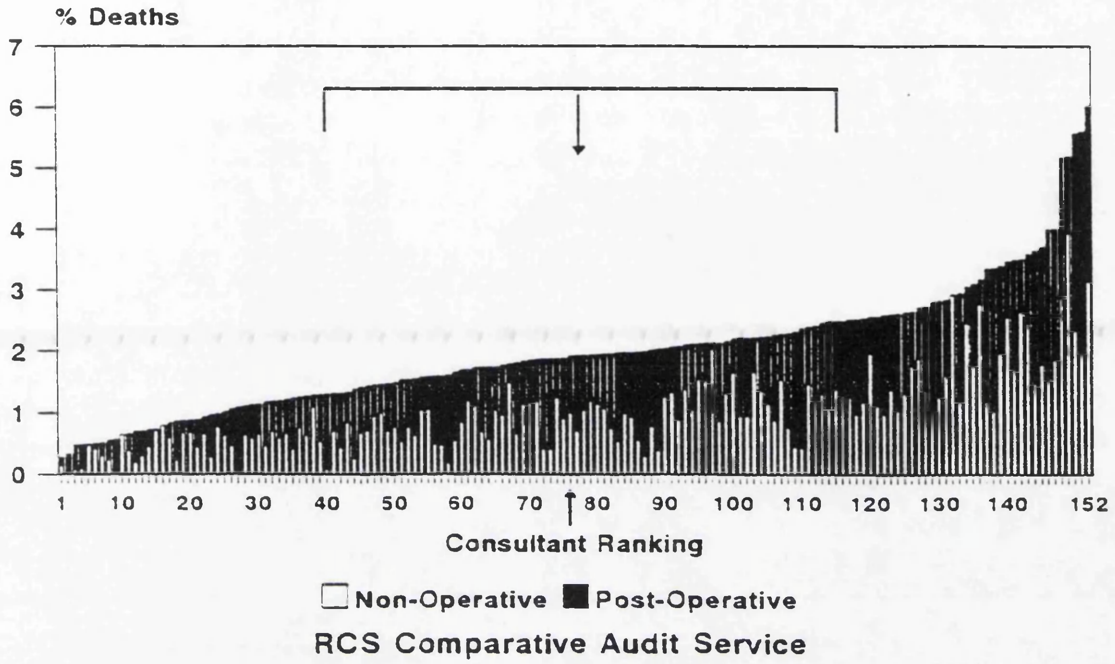


Figure 2: Example of Royal College of Surgeons' Data Presentation Methods

Diagnostic Case Mix 1992

Database Means plus Ranges

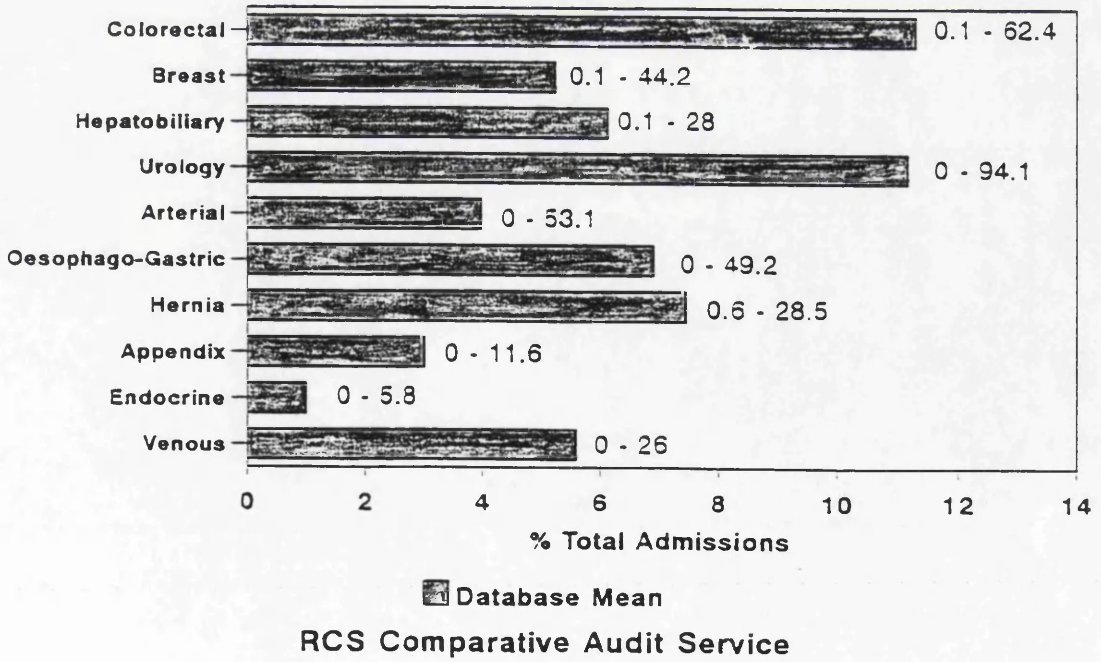


Figure 3: Royal College of Surgeons Presentation Method for Diagnoses

4 NEW PRESENTATION METHODS FOR RCS AUDIT DATA

4.1 Introduction

In the preceding chapter, we described the data obtained from the Royal College of Surgeons of England, and highlighted some of its drawbacks. We also described the methods which were originally used to present these data to consultants. These were unsatisfactory as they meant little to outside observers, and did not consider numbers of patients when making comparisons. We now describe the methods of analysis which we have introduced in order to improve the presentation of these data. These methods show confidence intervals for relative mortality, triangle diagrams of case mix, and finally adjustments of the confidence intervals for case mix. Due to the drawbacks of the data, which were described in the last chapter, this work should be considered more as a presentation of ideas about how to deal with these data than as a definitive analysis of differences between surgeons.

4.2 Relative Risk Confidence Intervals

Instead of presenting raw mortality rates, relative mortalities were calculated for each consultant, by dividing their mortality rate by the mean rate of all the others. Here, we are concentrating on total mortality, both postoperative and non-operative.

We think of the data as being an outcome frequency in N groups of patients, i.e. the number of deaths for each consultant, and it can be expressed in a table as follows

			Consultant		
	1	. . . . .	i	. . . . .	N
Dead	d_1	. . . . .	d_i	. . . . .	d_N
Alive	a_1	. . . . .	a_i	. . . . .	a_N
Tot Adm	n_1	. . . . .	n_i	. . . . .	n_N

or equivalently in N separate tables such as:

	Consultant	
	i	not i
Dead	d_i	Σd_j $(j{\neq}i)$
Alive	a_i	Σa_j $(j{\neq}i)$
Tot Adm	n_i	Σn_j $(j{\neq}i)$

The probability of death for a patient treated by surgeon i is estimated, then, as d_i/n_i, and an estimate of R, the relative risk of death for consultant i compared with all the other consultants is

$$\hat{R}_i = \frac{d_i/n_i}{\sum_{j\neq i} d_j \Big/ \sum_{j\neq i} n_j}$$

An approximation to the variance of logR can be found using the first two terms of the Taylor expansion of the function. This gives the estimated standard error as

$$se\left(\log\hat{R}_i\right) = \sqrt{\frac{1}{d_i} - \frac{1}{n_i} + \frac{1}{\sum_{j\neq i} d_j} - \frac{1}{\sum_{j\neq i} n_j}}$$

and thus an approximate 95% confidence interval for R can easily be obtained as

$$\exp\left\{(\log\hat{R}) \pm 1.96 \times se(\log\hat{R})\right\}.$$

This method of calculating confidence intervals was recommended by Katz *et al.* in 1978 and has since been endorsed several times in the medical literature. (e.g. Gardner & Altman 1989, chapter 6 p.51). The method has also been heavily criticised in the statistical literature. The estimate of the variance is unlikely to be precise, as it is a linear approximation to a non-linear function, and the confidence intervals have been accused of being unstable and of having inaccurate coverage probabilities.

Several other methods for calculating these intervals have been proposed, all of them computationally more complex than the logarithm method, and some involving numerical procedures for solution. In the next section, we shall look at two of the methods which have claimed superiority over the one of Katz *et al.*, and, using the data from the Royal College of Surgeons, compare the three techniques.

We could, alternatively, use the odds ratio

$$O.R_i = \frac{d_i/a_i}{\sum_{j\neq i} d_j \Big/ \sum_{j\neq i} a_j}$$

as our measure to compare mortality. With data such as these, where n is large, and d is very small, this is practically the same as the relative mortality rate. This value is equivalent to the exponential of the coefficient obtained when a binary variable which takes the value 1 for consultant i and 0 for the others is fitted in a logistic regression.

Table 6 compares the intervals obtained for three consultants using the two methods. There is little difference between the relative mortality and odds ratio estimates. The relative mortality is preferable as it is intuitively easier to understand.

Method	Consultant A	Consultant B	Consultant C
Ratio of mortality rates	(1.13,1.86)	(1.08,1.99)	(0.20,0.89)
Odds Ratio	(1.13,1.88)	(1.08,2.02)	(0.20,0.89)

Table 6: Confidence intervals calculated by two different methods for three consultants (1992 data)

The values for each surgeon were easily calculated in MINITAB. They were arranged in rank order and the 95% confidence intervals plotted on a log scale against rank using the *Splus* "matplot" function. The plots are fairly accessible to surgeons as they are similar to the way meta-analysis data are presented. The calculations were performed for various subsets of surgeons, according to other variables which had to be present. Figure 4 shows the relative risk confidence interval plots for the 96 surgeons in 1991 who had complete data on diagnostic groups, age and admission status, except for one who had zero deaths and therefore for whom a value could obviously not be calculated using the standard procedures. Looking at the plot, it can be seen that the majority of consultants have confidence intervals which contain 1, and so there is no evidence of them deviating from the norm. It should be borne in mind that these confidence intervals are not joint, but separate, and so we would expect, on average, 5% of them not to contain 1 purely by chance. However, approximately one quarter of the consultants have confidence intervals either entirely above or below 1, giving evidence that there are actual differences between the consultants with causes other than random variation. This may be explainable by differences in case mix as will be explored later. This step from displaying only simple mortality rates to showing the relative risk confidence intervals is a large one, despite being straightforward statistically. By plotting only ranks, individual confidentiality is maintained, and a printout of surgeon number and rank can be obtained in order to inform consultants of their position.

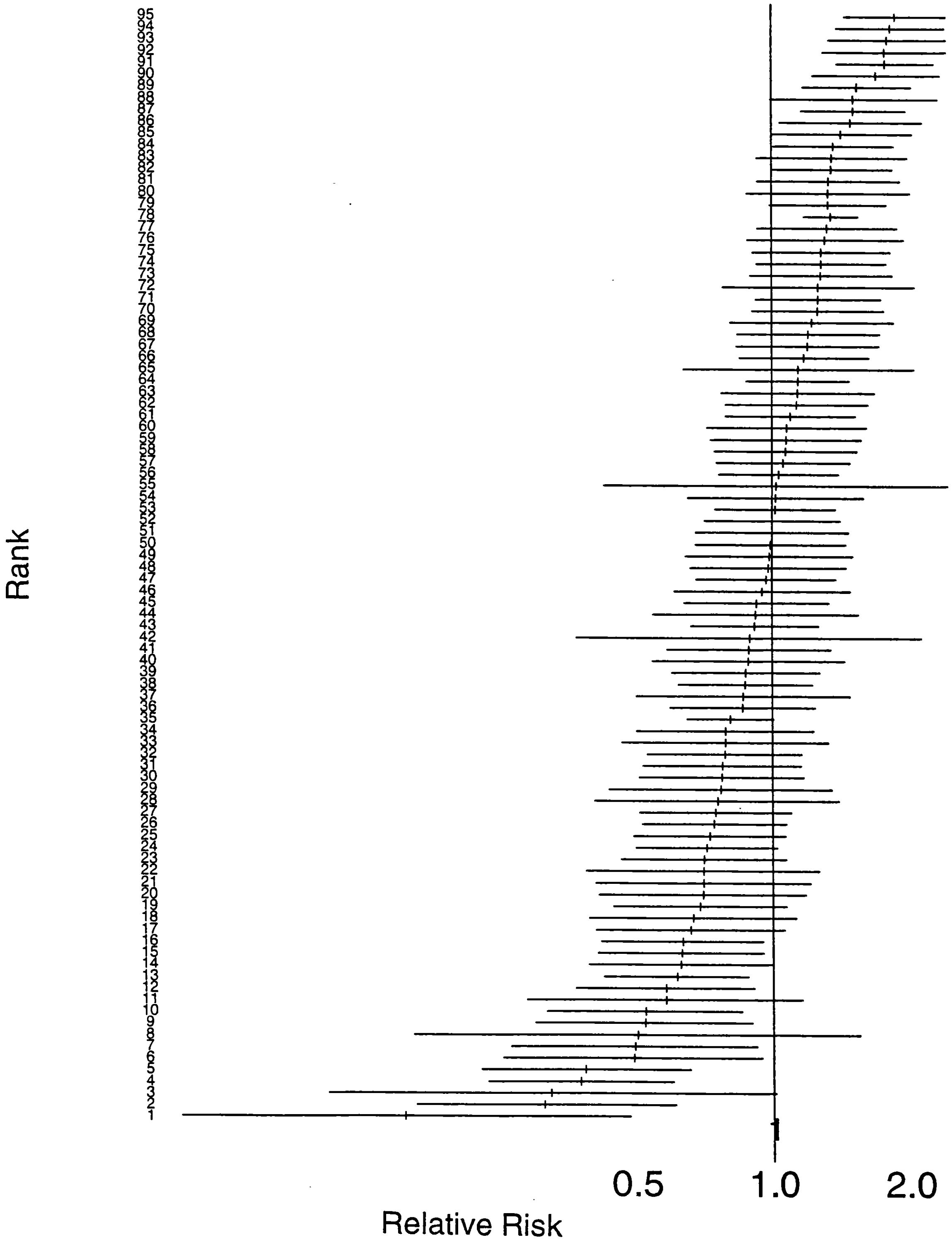


Figure 4: Unadjusted Relative Risk Confidence Intervals for ranked Consultants (1991 RCS Data)

4.2 Comparison of confidence interval estimation methods

As previously mentioned, there has been much criticism of the logarithm method of calculating relative risk confidence intervals. There have been several publications in the 20 years since the recommendation of this method by Katz *et al.*, advocating new methods which do not rely on such asymptotic theory. Every method suggested since the 1978 paper has made comparisons with the logarithm method, as it has become a standard. The other general approach taken is to use a likelihood scoring method, as summarised in 1988 by Gart and Nam, who concluded that this approach was preferable to taking logarithms of binomial proportions for small and moderate sample sizes. Koopman's publication in 1984 was the first to use the Chi-square method, and the methods subsequently proposed by Gart, and Miettinen and Nurminen (both 1985) were very similar. Criticism of Koopman's method came from Bailey in 1987, who suggested a direct formula for the calculation, rather than a method requiring iteration. The methods have all been compared in the literature, but never with the magnitudes of the sample sizes prevalent in the Royal College of Surgeons' data. It is of concern that, although we have, in most cases, large sample sizes, the rates in question are very low. For this reason it is necessary to investigate the accuracy of estimation of different methods using this particular data set. The two alternative methods chosen for comparison were those which appeared to be the strongest contenders with the logarithm method so far, and are representative of the work to date. These are

1. "Chi-square" Method (Koopman 1984)
2. "Power" Method (Bailey 1987)

These methods, which will be described below, were applied to the data from 1991 and 1992 as well as the logarithm method, using a FORTRAN program.

4.2.1 Notation

In this section, the notation used will be slightly different. The 2×2 table for consultant i would be

	con i	not con i
dead	x	y
alive	m-x	n-y
admits	m	n

Then x and y are binomially distributed on sample sizes m and n, with parameters p_1 and p_2 say. Then the relative risk R is p_1/p_2.

4.2.2 Chi-square Method

This method is based on a test statistic for the null hypothesis that R is equal to some value R_O, versus the alternative that R equals some other value. This chi square test statistic is given by

$$U_{R_o}(x,y)=\frac{(x-m\hat{p}_1)^2}{m\hat{p}_1(1-\hat{p}_1)}+\frac{(y-n\hat{p}_2)^2}{n\hat{p}_2(1-\hat{p}_2)}$$

where $\hat{p}_1$ and $\hat{p}_2$ are maximum likelihood estimators for p_1 and p_2 under $R=R_O$, as given in Koopman's paper.

To calculate a 95% confidence interval (R_L,R_U) for R we must solve

$$U_{R_L}(x,y)=U_{R_U}(x,y)=\chi^2_{(1,0.95)}=3.841$$

U is a convex function of R so the graph of U *versus* R only crosses the line U(x,y)=3.841 twice; at the end points of the confidence interval. To find these points, we have to employ a numerical method as there is no explicit solution to the above. This has been done using repeated bisection separately for the lower and upper limits. The starting intervals used the estimates from the logarithm method and an easily found point known to have a value of U on the other side of 3.841. Sufficient iterations to achieve convergence to 4 decimal places were used and the values of the lower and upper limits output to a data file. These confidence intervals were then plotted in order of the relative risk and compared with those from the other methods.

Asymptotically the methods are equally efficient, but we are interested in the efficiency for the data we have.

In the 1984 publication, this method was compared directly with the logarithm method by computing actual coverage frequencies for different choices of the parameters. Probabilities that the lower limit overestimated and that the upper limit underestimated the true value of R were also calculated. These should, of course, equal 0.025 for a 95% interval. The Chi-square method appeared to be far more accurate for most chosen values of the parameters. However for $p_1(=x/m)$ very small and n greater than m, the logarithm method was more accurate for the lower limit but not for the upper. These are the most similar conditions to our data, but Koopman only calculated values for a total sample size of 200 compared with our totals of 288,488 for the two years of data.

4.2.3 Power Method

This method is based on a power of the observed ratio, and is claimed to be more stable than the logarithm method. The publication in which the method was introduced recommended a power of 1/3, and gave a formula to calculate the 95% confidence interval as

$$(R_L, R_U) = \left(\frac{p_1}{p_2}\right)\left(\frac{1 \pm 1.96\sqrt{(q_1/x + q_2/y - 1.96^2 q_1 q_2/(9xy))}/3}{1 - 1.96^2 q_2/(9y)}\right)^3 .$$

These were easily calculated from the data in a FORTRAN program.

This method was also compared to the logarithm one in the original publication, again by looking at coverage probabilities for various parameter values. It was shown to be superior to the logarithm method, particularly for small sample sizes. The closest sample sizes to the RCS study in this publication were m=50 and n=200. This combination gave the smallest range of coverage probabilities for the log method than any of the other combinations considered.

4.2.4 Comparisons

Looking at the plots of the confidence intervals for the three methods (fig 5(a)-5(c)) we can see a general agreement between them, especially at the top end. In fact, whichever method was used we would come to the same conclusions for almost every consultant. As we would expect, the few exceptions lie at the low risk end, where the upper limit reaches 1 in a few cases for one method but not the others.

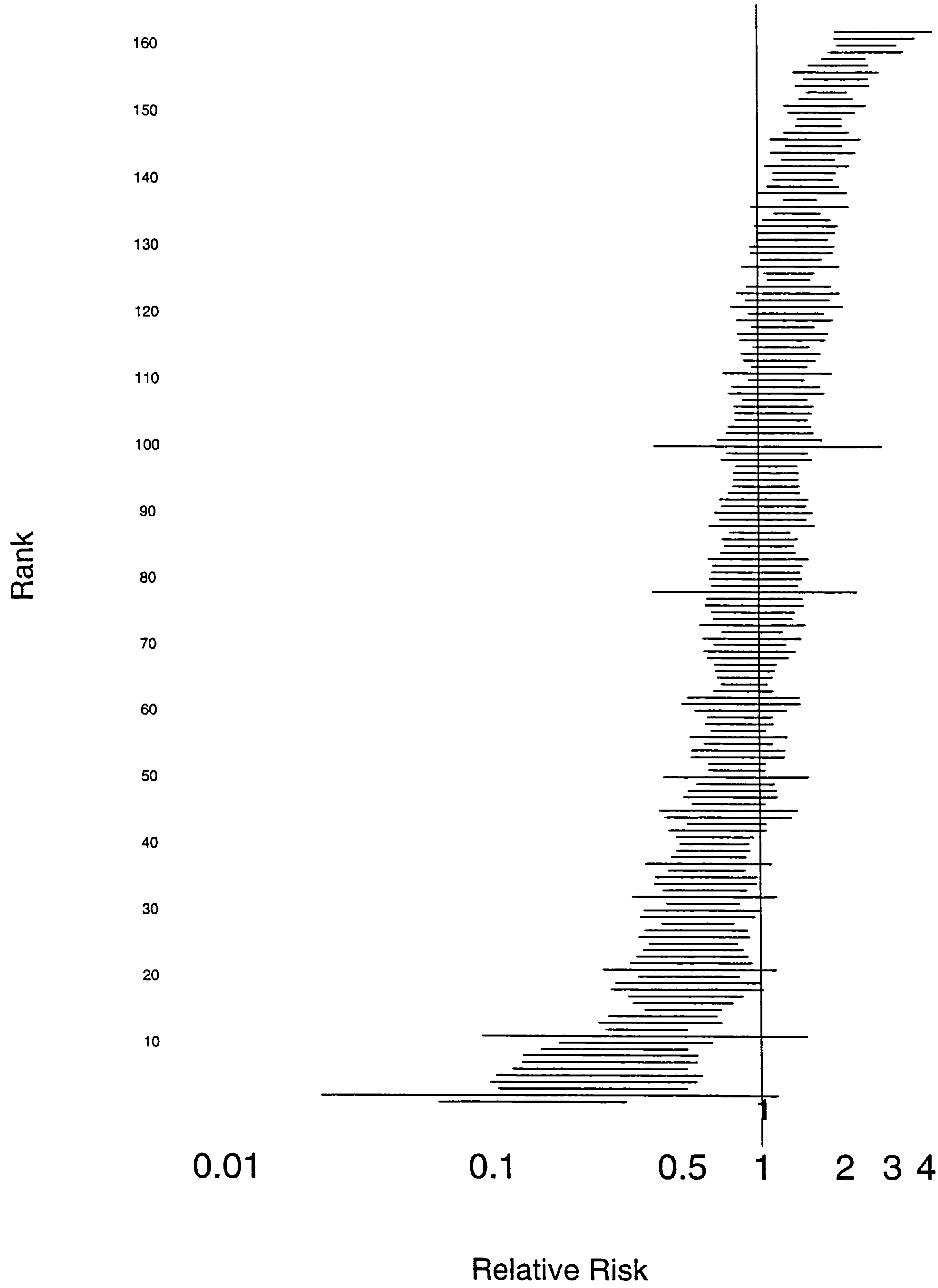


Figure 5(a): Unadjusted Relative Risk Confidence Intervals for Ranked Consultants - Logarithm Method

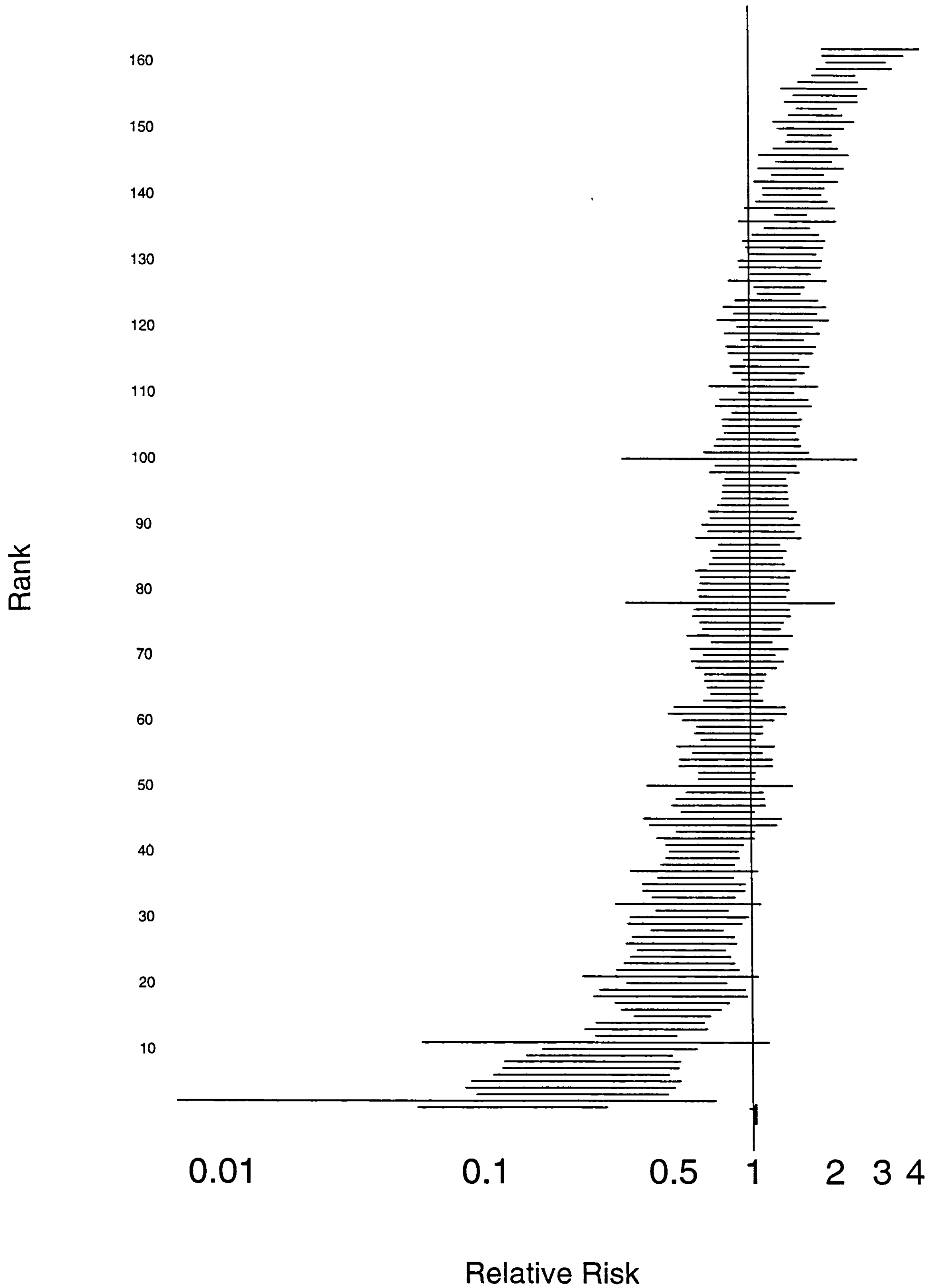


Figure 5(b): Unadjusted Relative Risk Confidence Intervals for Ranked Consultants - Power Method

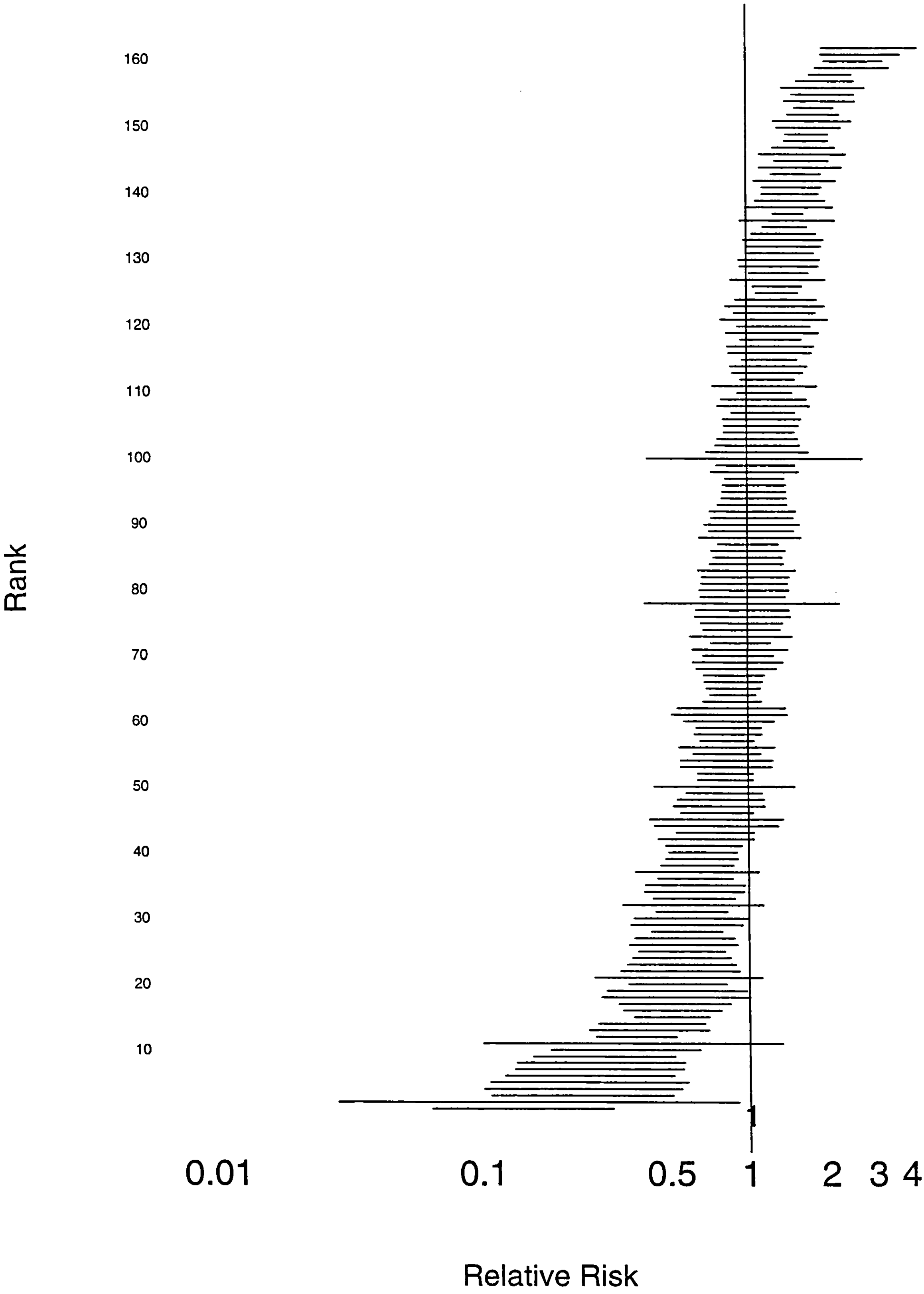


Figure 5(c): Unadjusted Relative Risk Confidence Intervals for Ranked Consultants - Chi-square Method

To see the differences more easily, the corresponding limits have been plotted against each other to compare the three methods in a pairwise way. It can be seen (fig 6(a)) that, for the lower limit, the logarithm and Chi-square values correspond almost exactly, apart from two slight outliers. These are the consultants ranked 2nd and 11th, who had 1 death out of 302 admissions and 2 out of 265 respectively, so these two methods disagree for a combination of low death rate and low number of admissions. The power method, however, underestimates many of the values compared to the logarithm method (fig 6(b)), and does not conform to the Chi-square method nearly as closely (fig 6(c)). It remains unknown whether the estimates from the Power method with either small numbers of deaths or of admissions are too low or are in fact superior to those of the Logarithm and Chi-square methods. For the upper limit, there is more correspondence between the methods. Again, however, those with low death rates or numbers of admissions have lower estimates using the power method than the log method (fig 7(b)). The three outliers, who all have their upper limits estimated as much lower by the logarithm method, can be seen in fig 7(a). They are the two consultants (2 and 11) as described above for the lower limits and the consultant ranked 100, who had 4 deaths out of only 181 admissions. Again, the number of admissions appears to be the main factor in causing discrepancies between logarithm and Chi-square methods. These cases have the widest confidence intervals so the differences are not as important as one may initially think.

For these reasons, we conclude that the use of the logarithm method is justified for use in calculating relative risks for these data. Not only does it compare more favourably with the Chi-square method than the more recently proposed power method but it is far easier to calculate than any other option. In the cases where it does make the interval too narrow, there are very few deaths or admissions which is very rare. Since these cases do not tend to be in the high risk area, they are not so crucial, and we can be safe in the assumption that those with higher numbers of deaths, in whom more interest will be taken, will not have their confidence intervals wrongly estimated by using the logarithm method. For the majority of consultants, with “typical” workloads, there is no difference between the methods for the upper limit.

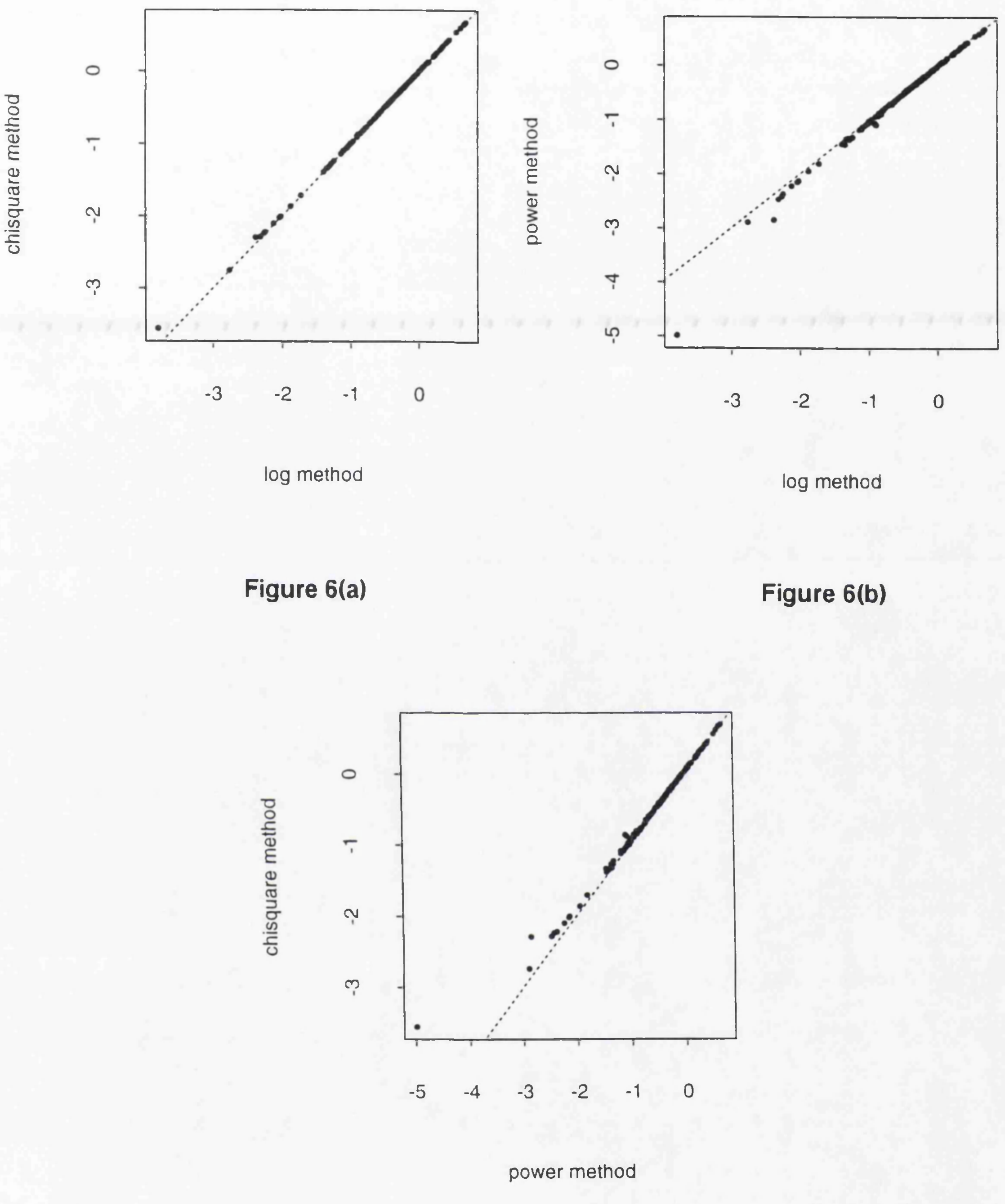


Figure 6(a)

Figure 6(b)

Figure 6(c)

Figure 6: Plots comparing Lower Confidence Limits for log relative risk calculated using different methods

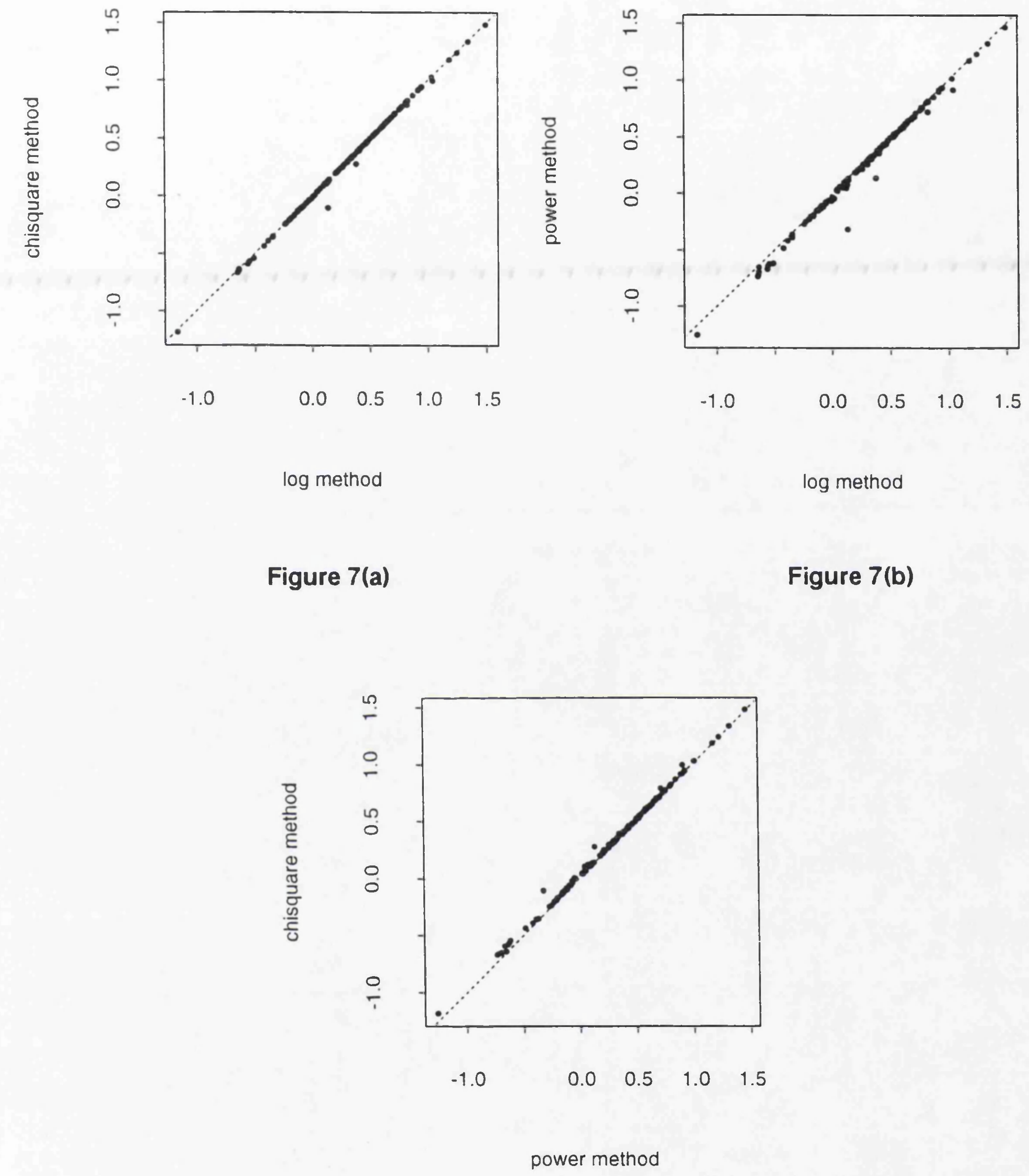


Figure 7(a)

Figure 7(b)

Figure 7(c)

Figure 7: Plots comparing Upper Confidence Limits for log relative risk calculated using different methods

4.3 Triangle Plots

Triangle Plots are a way of representing case mix graphically to show the variation among consultants. For example, a patient's admission status could either be day case, elective or emergency, but the only information available is the total number of patients in each category for each consultant. So, using proportions in each group, we have a data set which has the property that the three values sum to 1, i.e. compositional data, and methods of analysis must be found which cope with this constraint. Triangle plots are a standard way of displaying this type of data. The plots use the property of an equilateral triangle that, taking perpendiculars from any point to each side of the triangle, the sum of the lengths of these lines from their respective sides to the point where they all meet is equal to the height of the triangle. So, with a triangle of height 1 unit, we make each corner represent a group and the length of the perpendicular from the opposite side represents the proportion of patients in that group for a particular consultant. Thus the nearer a point is to a corner of the triangle the higher the proportion of that surgeon's patients are in that category.

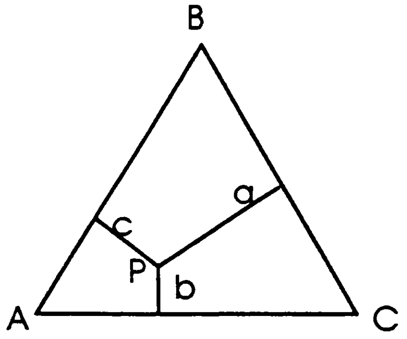


Figure 8: Drawing a Triangle Plot

In order to plot the triangles, we need to find the point P where the three lines intersect. If we take A as (0,0) then the y co-ordinate of P is simply b (the proportion in category B) and the x co-ordinate can be calculated using simple trigonometry to be $(b+2c)/\sqrt{3}$. They have been plotted in *Splus*, using the relevant line segments instead of axes.

Although the admission status data are naturally suitable for this type of display, it has also been done for age distribution and diagnostic distribution. With the 1991 data age proportions were first of all grouped into under 40, 40 to 60 and over 60, as these were thought to be appropriate low, medium and high risk categories. As can be seen in the triangle (Fig 9(a)), there is not a wide spread of age distributions. The pattern looks much the same, except the cluster is moved up more towards the middle group, if 20 and 70 are used as the cutpoints instead, and similarly if ages 10 and 60 are used as the cutpoints (fig 9(b)). These latter groups will be used since the data for 1992 only contains information on them. The age triangle for 1992 shows a similar pattern to the 1991 one. The diagnostic groups were collapsed into three risk categories, as consensually agreed by a group of surgeons. The groups are as follows

Low risk: breast and hernia

Med risk: urological, appendix, endocrine, nervous and hepatobiliary

High risk: colorectal, arterial and oesophago-gastric

Originally, it was unclear whether hepatobiliary diagnoses should be regarded as medium or high risk. It seemed sensible to include it in the middle group, in order that the overall proportion in the high risk category was not too large.

The triangle plots for admission and diagnoses show considerable spread among surgeons. These are shown for the two years' data together in Figures 10 and 11. In the triangles, we would expect those consultants nearest the bottom right hand corner to have the worst outcomes. To illustrate the relationship between case mix and mortality, the points can be plotted using mortality rate quartiles, with 1 representing the consultants in the lowest 25% for mortality rate and 4 those with the highest proportions of deaths. It can be seen from figure 12 that those with the highest proportion of emergencies tend to have the highest mortality rates and similarly, from figure 13 those with most patients having high risk diagnoses.

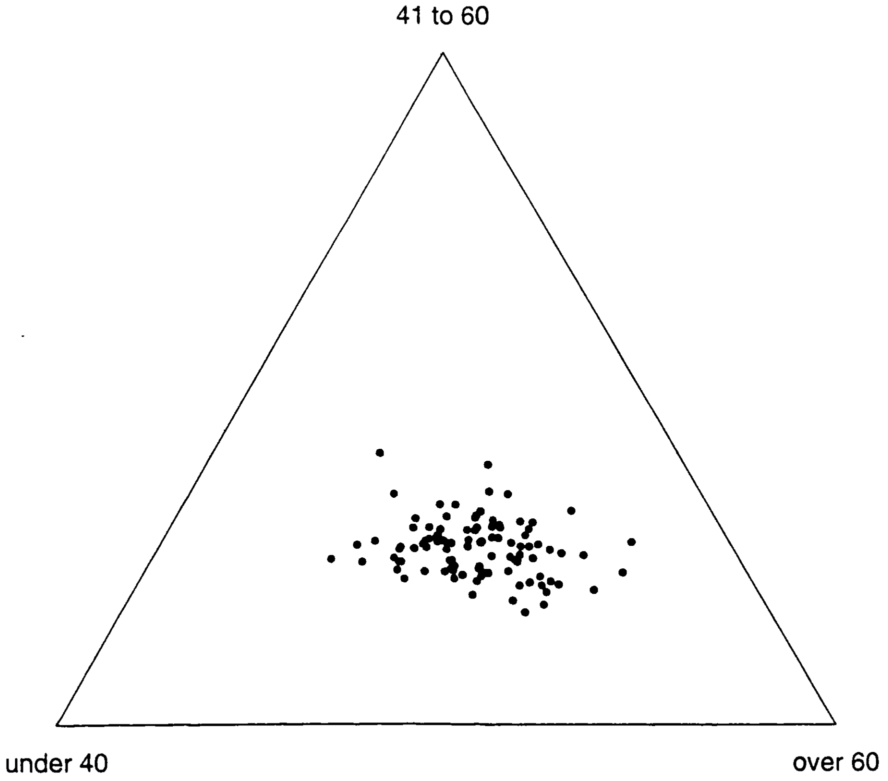


Figure 9(a): Triangle Plot Showing Spread of Age Groups with Lower Cut-off 40 (RCS 1991 Data)

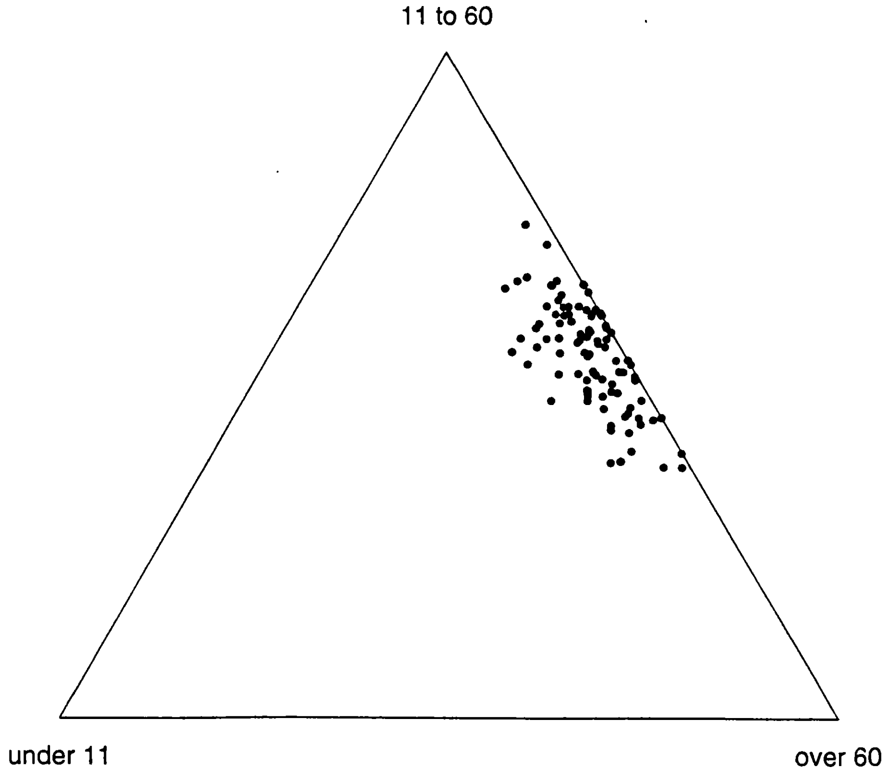


Figure 9(b): Triangle Plot Showing Spread of Age Groups with Lower Cut-off 10 (RCS 1991 Data)

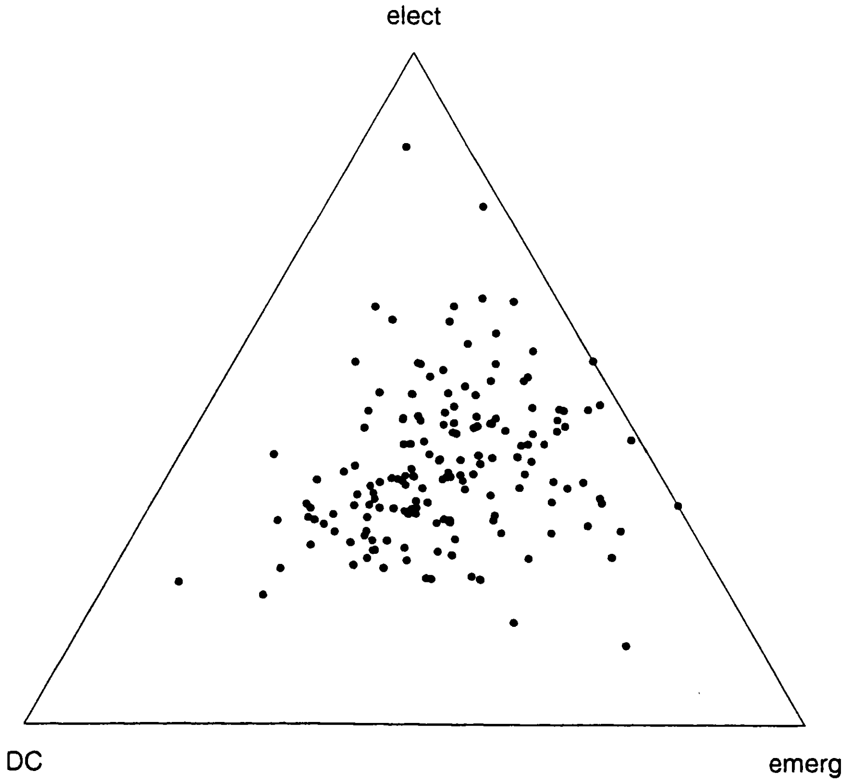


Figure 10: Triangle Plot Showing Distribution of Admission Category (RCS Data from 1991 and 1992)

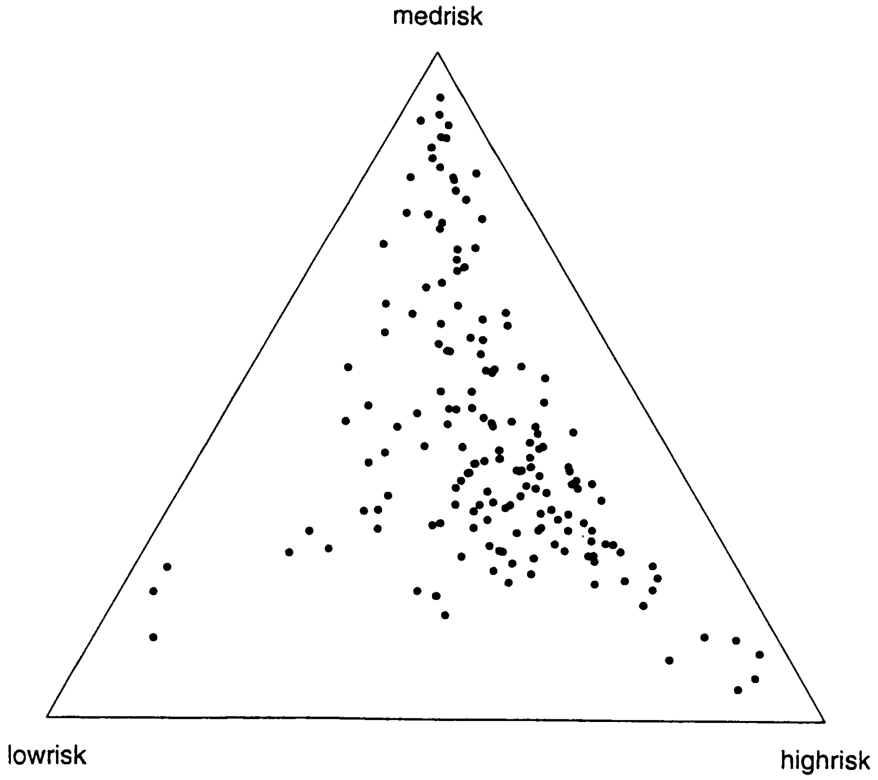


Figure 11: Triangle Plot Showing Distribution of Diagnostic Risk Category (RCS Data from 1991 and 1992)

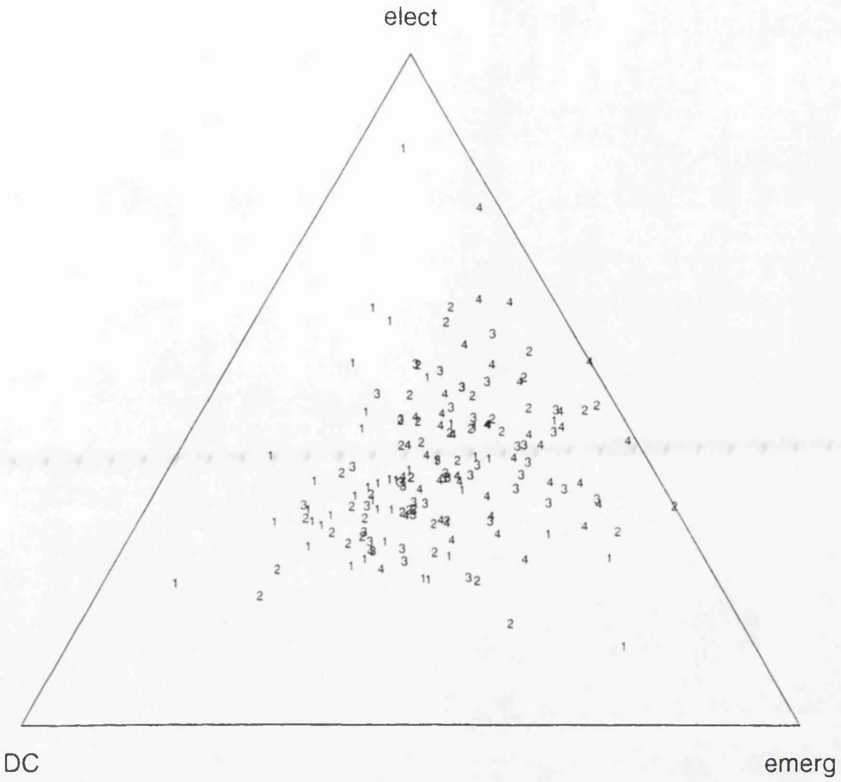


Figure 12: Triangle Plot Showing Distribution of Admission Category, labelled by Mortality Rate Quartile (RCS Data from 1991 and 1992)

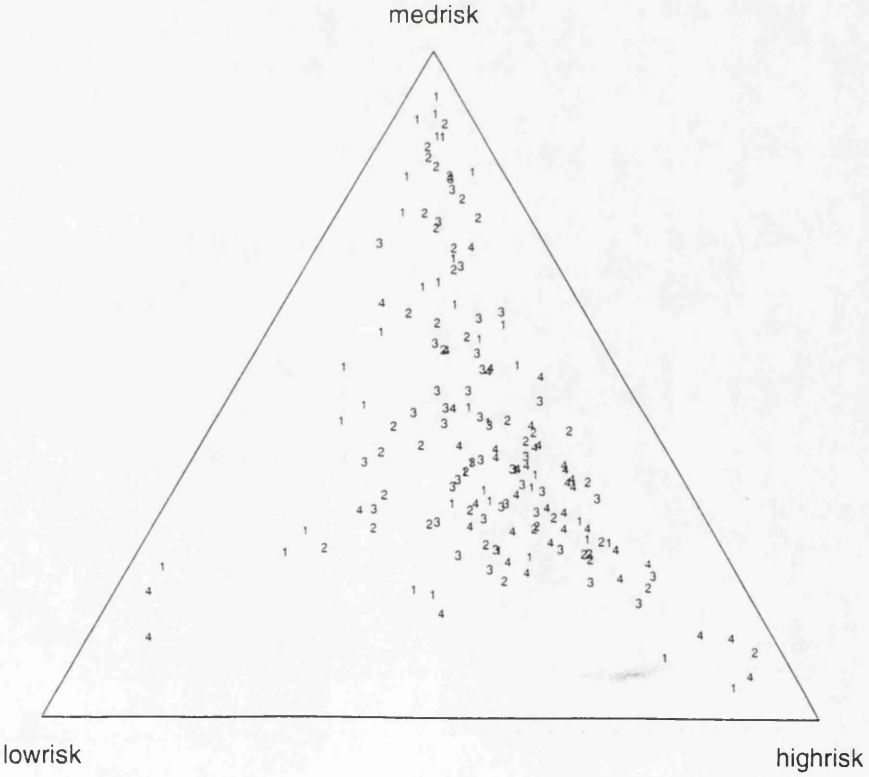


Figure 13: Triangle Plot Showing Distribution of Diagnostic Risk Groups, labelled by Mortality Rate Quartile (RCS Data from 1991 and 1992)

Another way of showing the relationship of case mix with mortality is to divide the triangular diagram into several sub triangles and then calculate the mean mortality rate for each small triangle as in Figures 14(a) and (b). From these pictures we can see that both for admission status and diagnostic risk group, the mean mortality rates in general increase towards the higher risk categories. There is not an exact correspondence as one might hope, but those sub triangles with surprising results are those with the most sparse data, so that the numbers are small, and moreover that particular mixture of patients is unusual and the outcomes may be atypical. Also, there are many other factors which could be related to mortality which cannot be accounted for in this highly simplified way of looking at the data. It is possible that the reason for the inconsistencies with expectation is that surgical skill is an important cause of variation, and differences cannot be explained purely by looking at case mix. The triangular plots do, however, present these case mix factors in a way that is easy to digest, and show more information than the original bar charts could.

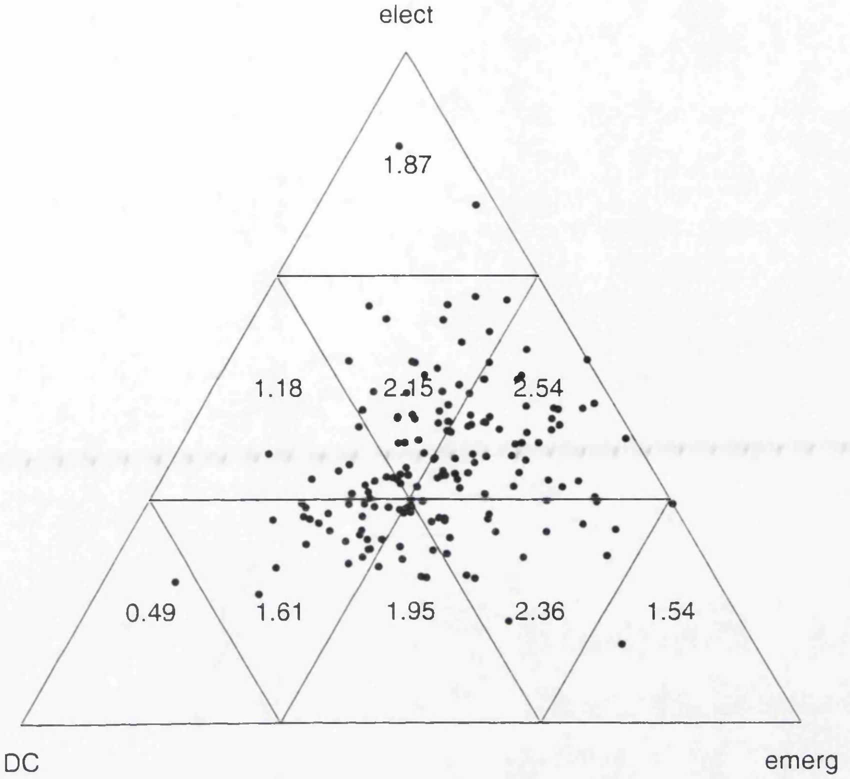


Figure 14(a): Triangle Plot of Admission Category showing Mean Mortality Rates of Consultants in Sub-Triangles

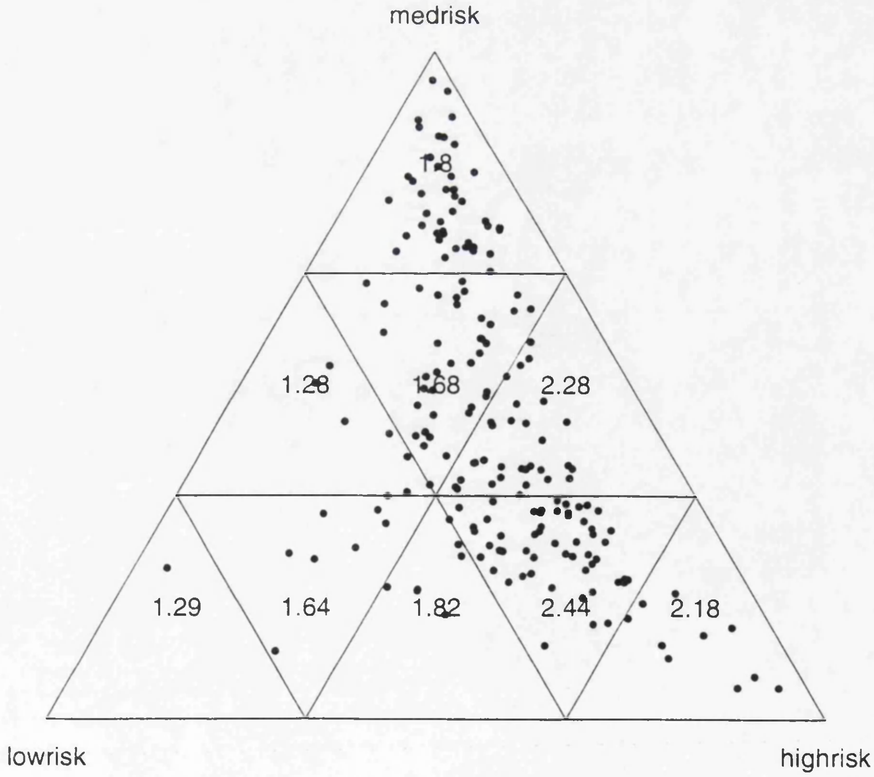


Figure 14(b): Triangle Plot of Diagnostic Risk Category showing Mean Mortality Rates of Consultants in Sub-Triangles

4.4 Adjusted Relative Risks

The relationship between case mix and mortality was explored using logistic regression. This was carried out using the BMDP program LR, with the number of deaths for each consultant representing the “success” count. This is rather an unusual method of performing logistic regression, as the response is the number of deaths rather than dead/alive as would be used with patient level data. This particular form of data means that the explanatory variables cannot be exactly related to the outcome. In chapter 6 we shall investigate how much information is lost by having only this summary data rather than having data on individual patients. This will be done using an actual data set from a study on colorectal cancer, and on simulated data.

A stepwise approach was taken at first, with forward selection of variables relating to age, diagnosis and admission status. Other variables were discarded because too few consultants had supplied information on them or because they applied only to operative patients as opposed to all patients, and so were not relevant for total mortality. Each of the two categories included for all three variables were entered into the equation. Although the pairs of variables are related, they are included in the model. This could be thought of as similar to fitting two indicator variables when there is a categorical variable with three possibilities, except here we have summed these up over the patients for each consultant. If under 40 is used as the low age group rather than under 10 with the 1991 data, then the over 60 term is not entered into the model. It would seem that under 40 is a more suitable low risk category for death than under 10, but the information is not available for 1992 so it is not so useful. The results of the logistic regressions on these variables for 1991, 1992 and the two years combined can be seen in table 7

Term	1991	1992	both years
Proportion:			
aged 0 to 10	2.26	–	–
aged over 60	0.92	1.84	1.12
day case admissions	-0.49	-1.26	-1.08
emergency admissions	0.62	1.24	0.87
low risk diagnoses	0.76	–	–
high risk diagnoses	0.51	0.69	0.49
constant	-4.53	-5.01	-4.48

Table 7: Coefficients of variables in logistic regression equations modelling total numbers of deaths for 1991, 1992 and both years combined

It can be seen that the proportion of low risk diagnoses in 1991 has a positive coefficient, where we would expect it to have a negative contribution to the predicted log odds. This could be explained by the fact that the variables are correlated with each other. So, for example, since the proportion of low risk patients is negatively correlated with the proportion over 60, their coefficients in the model are positively correlated (r=0.495). If the proportion over 60 had been given less weight then the low risk diagnoses could have had a much smaller, or negative coefficient. Another explanation for this could be that the broad categorisation of the diagnostic groups into risks is not accurate. For example, some high risk groups may contain a particular low risk diagnosis, although generally describing more serious conditions. This is, unfortunately, a disadvantage of the data and we will presume, for the moment, that the overall model fits the data and will work in adjusting individual consultants' data.

The coefficients in the equations calculated from the 1992 data and from the 2 years' data are more intuitively appealing. It may be that the data for 1992 were of greater quality, or that the higher numbers resulted in a more reasonable model than in 1991.

Having calculated a model, we then calculate the adjusted relative risks by taking the ratio of the observed and expected log odds, i.e. the estimate of the value for surgeon i is

$$\hat{R}_i\text{adj} = \frac{d_i/a_i}{p(\text{death})/(1-p(\text{death}))},$$

where d_i and a_i are as before and p(death) is the predicted probability of death for a patient of surgeon i from the logistic regression equation. Thus the log of the adjusted value is simply the difference between the observed and predicted log odds. The observed values are easily calculated in MINITAB, and the predicted ones can be extracted from the BMDP output if the PRINT CELLS command is used. Alternatively, they can be calculated in MINITAB using the above equation.

The approximate standard error of the adjusted value is then obtained by

$$\mathrm{s.e}(\log \hat{R}_i^{\mathrm{adj}}) = \sqrt{\frac{1}{d_i} + \frac{1}{a_i} + \mathbf{x}_i^T \mathbf{V} \mathbf{x}_i}$$

where $\mathbf{x}_i$ is a vector of the data values for the significant variables for surgeon i, of the form (DC, em, over60, hirisk, 1) and V is the covariance matrix of the regression coefficients. This is obtained from the BMDP output and, for the 1992 data, is approximately

$$\begin{bmatrix} 0.033 & & & & \\ 0.023 & 0.046 & & & \\ 0.001 & 0.001 & 0.008 & & \\ 0.001 & 0.007 & -0.000 & 0.047 & \\ -0.018 & -0.026 & -0.004 & -0.023 & 0.026 \end{bmatrix}.$$

As can be seen these values are very small, and since the $\mathbf{x}_i$ are all between 0 and 1, the variance accounted for by the denominator (predicted log odds of death) of the adjusted relative risk ($x_i^T V x_i$) is very small and so most of the variability comes from the numerator (observed log odds) and depends on the number of deaths. Exploration of the relative contributions to the total variance of the adjusted values by the numerator and denominator can give us an idea of how well the model is predicting. If the values for the denominator are very small, then little variability comes from the model and most depends on the actual number of deaths (the lower the number of deaths, the less faith we can have in our estimate of the relative risk). In table 8 the contribution from each part of the variance of the three consultants, “A”, “B” and “C” is summarised for each year. As expected, we can see that the number of deaths has the largest effect on the value of the variance of a consultant’s relative risk. For each of the consultants, the value of the variance of the predictive part (denominator) is smaller for 1992 than 1991, even if the number of admissions is less, suggesting that the model may be a better fit.

year	consultant	# deaths	# admits	numerator var (%)	denom var (%)
1991	A	56	2227	0.0183 (92.1)	0.0016 (7.9)
1992	A	61	2173	0.0169 (92.5)	0.0014 (7.5)
both	A	137	4400	0.0088 (91.8)	0.0008 (8.2)
1991	B	5	222	0.2046 (99.4)	0.0013 (0.6)
1992	B	41	1438	0.0251 (98.3)	0.0004 (1.7)
both	B	46	1660	0.0224 (98.8)	0.0003 (1.2)
1991	C	3	866	0.3335 (98.6)	0.0048 (1.4)
1992	C	7	838	0.1441 (98.5)	0.0022 (1.5)
both	C	10	1704	0.1006 (99.0)	0.0010 (1.0)

Table 8: Relative contributions to the variances of adjusted relative risks of the observed and predicted parts of the estimator, for three surgeons

In section 4.1, the calculation of confidence intervals using the coefficients from logistic regression equations was discussed. Fitting a separate equation for each consultant with a binary variable gave the unadjusted log odds for that consultant, which was approximately equal to the relative risk. If we fit a logistic regression model with the significant variables and consultant, this will give us the values corrected for these variables. Table 9 shows the adjusted confidence intervals for the relative risks of the same consultants A, B and C as before, as calculated by the two different logistic methods described. Again, fitting a separate regression equation for each consultant gives almost equivalent intervals to the method described above, of dividing the observed by the predicted log odds. This highly labour intensive method fits a slightly different equation for each consultant, whereas our ratio method uses the same coefficients of the significant variables for every one. This does not make much difference to the results because of the large numbers involved. The ratio method is easiest to carry out, and unlike the other method, computer time is not so dependent on the number of consultants.

Method	Consultant A	Consultant B	Consultant C
Ratio	(0.88,1.50)	(0.93,1.73)	(0.25,1.13)
Separate equations	(0.89,1.51)	(0.93,1.74)	(0.25,1.12)

Table 9: Adjusted confidence intervals calculated by two different logistic methods for three consultants (1992 data)

The adjusted confidence intervals were plotted in the same order as the observed ones, again on a log scale, and as can be seen from figure 15(a), the adjustments have resulted in a moderate reduction in variability and the rank order of the consultants has changed. While, initially, 11 confidence intervals at the top end (high relative risk) and 13 at the bottom end did not contain 1, after adjustment 5 lay wholly above and 10 completely below 1. If case mix accounted for all the differences between the surgeons, then we would expect 95% of the confidence intervals to contain 1. In fact almost 16% of the adjusted intervals exclude 1 and it seems fair to conclude that there is some other reason for the differences. This could be attributed to unknown factors, which may become clear after some investigation, for example, a unit may admit large numbers of patients for palliative care which shows up in a high total death rate. On the other hand the differences could actually be due to variation in surgical skill or even resources available and so it is important that they are discovered.

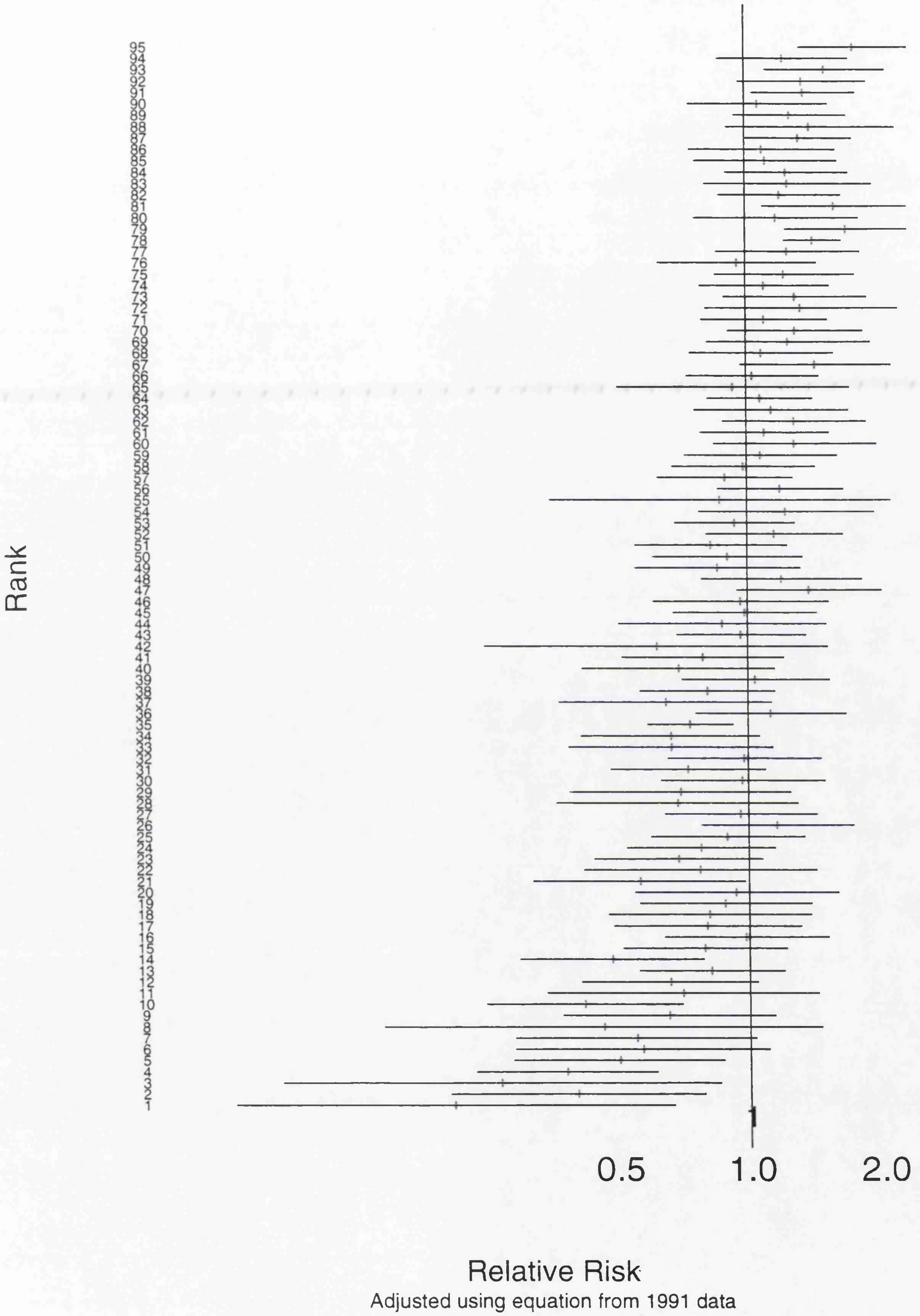


Figure 15(a): Relative Risk Confidence Intervals for Ranked Consultants (1991 Data, adjusted using equation calculated from 1991 data)

Looking in more detail, we can see that the direction of movement of the confidence interval on adjustment can usually be explained by the consultant's position in one or more of the triangles. For example, in 1991, the surgeon ranked 90 originally had a very high relative risk but adjustment caused it to decrease considerably. This is because 60% of his patients were older than 60 years and almost 80% of the diagnoses were in the "high risk" category. As a contrast, the consultant ranked 79 had an unadjusted confidence interval which just contained 1 at the lower end, but which on adjustment was considerably higher. This one had one of the lowest proportions (32%) of patients in the oldest age category. Thus when these factors are accounted for, this consultant appears to be performing worse than average. At the other end, a consultant who looked to be performing particularly well was ranked 26. On adjustment, however, his relative risk became greater than 1 and his rank became 52, although the interval still contained 1. This is due to the high proportion of day case patients and the particularly low (only 6%) proportion of high risk diagnoses seen. A consultant who performed rather better than one would have concluded from simply studying raw mortality was ranked tenth. His adjusted relative risk is fifth smallest, due to the fact that more than half of his admissions were emergency ones.

In contrast to the above examples, some confidence intervals did move in a counterintuitive direction on adjustment. For example, the consultant ranked 94 had his confidence interval decreased by adjustment so that it included 1. If we look at his case mix, we see that most of his patients (51%) were admitted electively, 25% of them were aged under 10 and 80% of them had low risk diagnoses. With this "easy" case mix, one would expect his predicted risk to be low, and thus his adjusted value to be higher. The reason for the strange result is the positive coefficient of low risk diagnoses. This made the predicted risk using the 1991 equation with the positive coefficient for low risk higher, and thus the adjusted value lower. All those consultants with a high proportion in the low risk diagnosis category had counterintuitive adjusted values. Perhaps the 1991 data were not sufficient to build a reliable model for adjustment and we would achieve more reasonable adjusted values using the model calculated from two years' data. These have been calculated, and have been plotted against the values calculated from the 1991 equation in figure 16. It can be seen that most of the values correspond fairly well. The outliers are those consultants with particularly high proportions of low risk diagnoses.

The confidence intervals adjusted using the 2 year equation are plotted in figure 15(b). The adjusted interval for the consultant ranked 94 is now much higher, as we would have expected from his case mix. Considering again, the consultant ranked 90, we can see that his relative risk is reduced even more by the 2 year equation than by the 1991 one due to his extreme case mix. These results seem more reasonable than before, and we would prefer to use the equation from both years to adjust these data.

A problem with using a model which is not calculated from the data set of interest to obtain adjusted relative risks is that the predicted values will be based on a different baseline mortality. The unadjusted risks are relative to the current group of consultants, and so depend on the overall case mix. The adjusted risks are compared with predicted and so the effect of case mix is removed and these values do not depend on the other consultants. Thus it is not strictly appropriate to compare the unadjusted and adjusted values, and counter-intuitive results may occur. The mortality rate of the 1991 data is very close to that of the 2 years' combined data, so this does not make a noticeable difference in this case, and we shall ignore it for the time being. In general, however, an adjustment should be made when using a model from a different data set to calculate adjusted relative risks, if it is desired to compare them with unadjusted values. We shall discuss this in more detail in section 5.6.

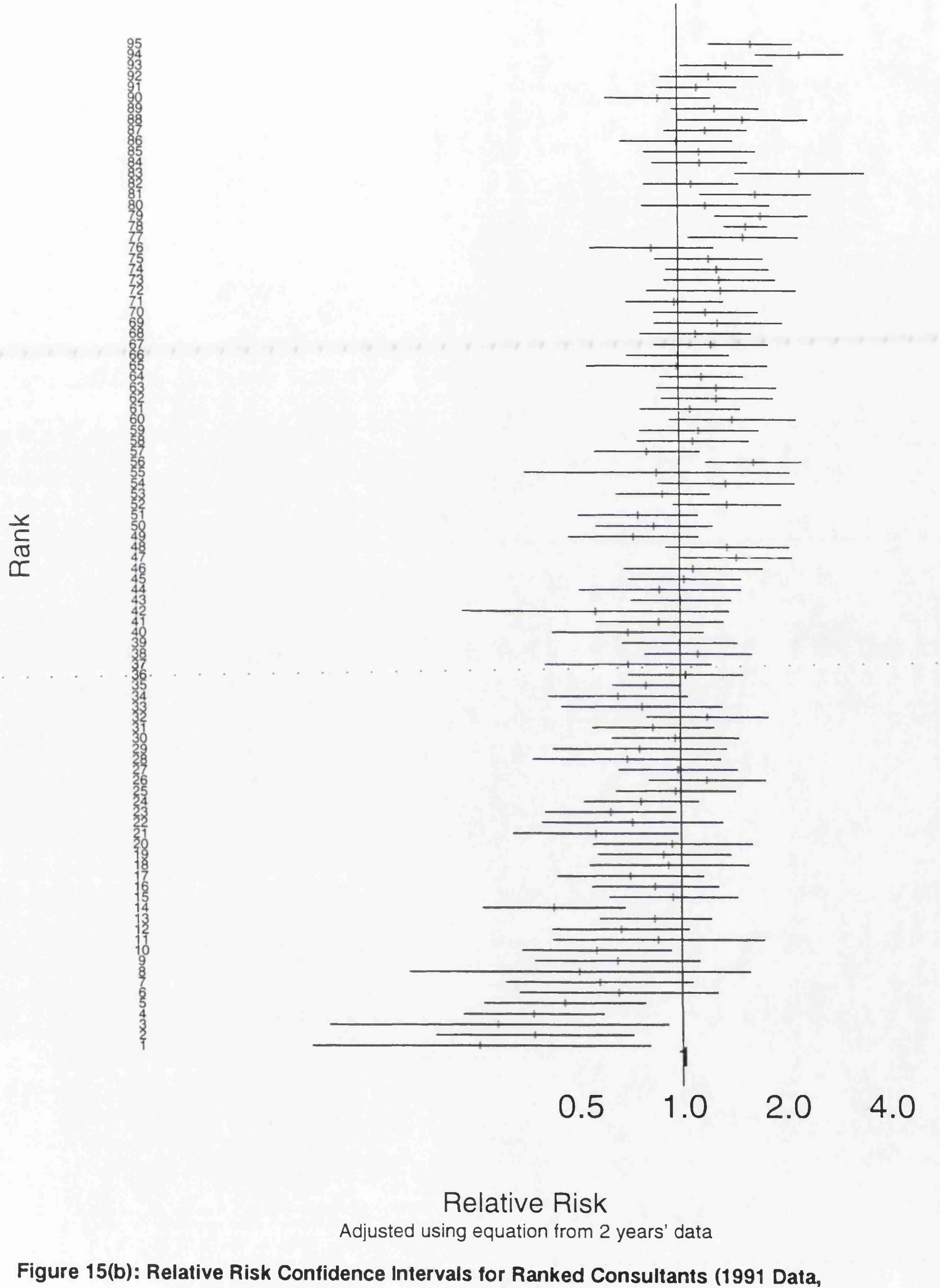


Figure 15(b): Relative Risk Confidence Intervals for Ranked Consultants (1991 Data, adjusted using equation calculated from 2 years' data)

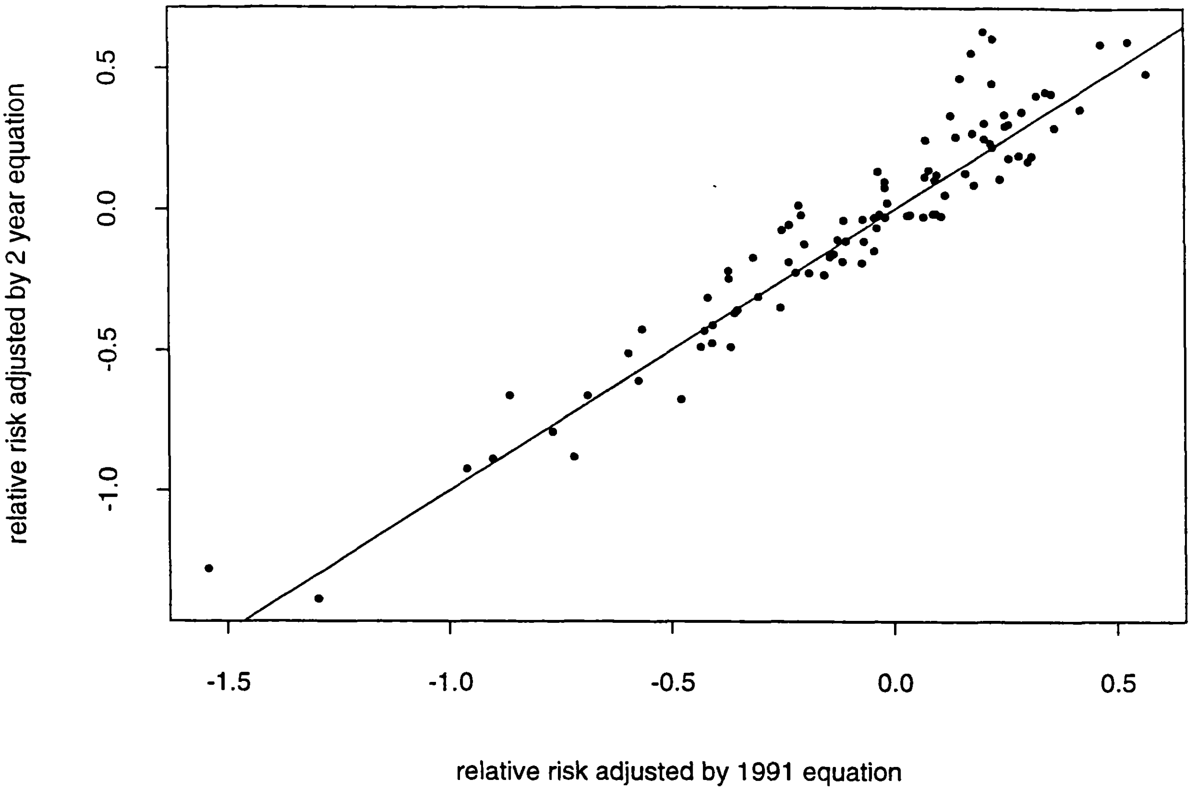


Figure 16: **Comparison of log Relative Risks adjusted using 2 different equations (RCS 1991 data)**

We turn now to the 1992 data. The adjusted intervals are plotted in figure 17, with arrows showing the change from the unadjusted to adjusted values. Of the unadjusted intervals, 48 did not include unity. This is over a third and would lead us to conclude that there are substantial differences between surgeons. Adjustment reduces this figure to 27, so some of the differences have been removed by including case mix. However, this figure is still much more than the expected 5% and suggests that there is still a difference between mortality rates of surgeons. As in the previous year some confidence intervals move to include 1, whereas others increase or decrease away from the average. An example where the adjusted confidence interval does not even overlap the unadjusted one is the consultant ranked 68. His original interval was centred on 1, and after adjustment his relative risk was fourth largest. This could be explained by the fact that more than half of this consultant's patients were day case admissions and only 9% of them had high risk diagnoses. Thus the adjustment showed him to be performing much worse than average, considering his low risk case mix. In contrast, the consultant ranked 77 had the eighteenth lowest relative risk after adjustment. The confidence interval moved from being centred slightly above 1 to being wholly below 1. Looking at his case mix, it can be seen that 62% of his admissions were emergencies and 45% high risk diagnoses, thus his mortality rate for this difficult mix was actually very low.

This seems to be a good model, but the question of what should be done on an annual basis is not answered. Should we calculate a model using all available data or should we just adjust using the equation developed using that particular year's data? It would obviously be a great administrative advantage if a model could be found which always gave reliable adjustments, but it may be fairer if the model with the best fit to that particular data set was used.

If we adjust the 1992 data using an equation calculated from data from two years, we find that the results correspond fairly closely with those calculated from the 1992 equation. In Figure 18 we can see that discrepancies lie fairly evenly in both directions. Looking at some of the outliers we find that the reason for them is the extra weight given to admission status by the 1992 equation. For example, the point marked "3" on the plot has the biggest difference between the adjusted values calculated from the two models, with the 1992 value being larger. This consultant had 69% of his patients admitted on a day case basis and so would have been predicted as being lower risk by the 1992 equation with its coefficient for day case of -1.26 than for the two year model with its day case coefficient of -0.63. Similarly consultant 6 has his adjusted value reduced even more by the two year equation than the one year one. This is because he has 86% of his patients in the "high risk" diagnostic groups, but also 50% of them were day case groupings. Since relatively more weight is applied to diagnostic risk than admission category by the two year equation this consultant's risk is considered higher. It is difficult to tell which model is making the most intuitively appealing adjustments, but due to the fairly arbitrary nature of the risk categorisations of the diagnoses and the more evident relationship between admission status and mortality, it would seem logical to favour the model which puts more emphasis on the emergency and day case proportions. Thus for 1992 it seems that the model calculated from that year's data only provides a superior means of risk adjustment, whereas the 1991 model is inferior to the two year model. There may have been a substantial improvement in the quality of the data from 1991 to 1992.

The fact that the analyses of the 1991 and 1992 data gave different results leads us to question the validity of the models and the data and suggests perhaps too small numbers, both in numbers of patients and numbers of consultants with complete data. In the following section we will investigate the effectiveness of models and compare results from year to year by looking at only those consultants having data for *both* 1991 and 1992.

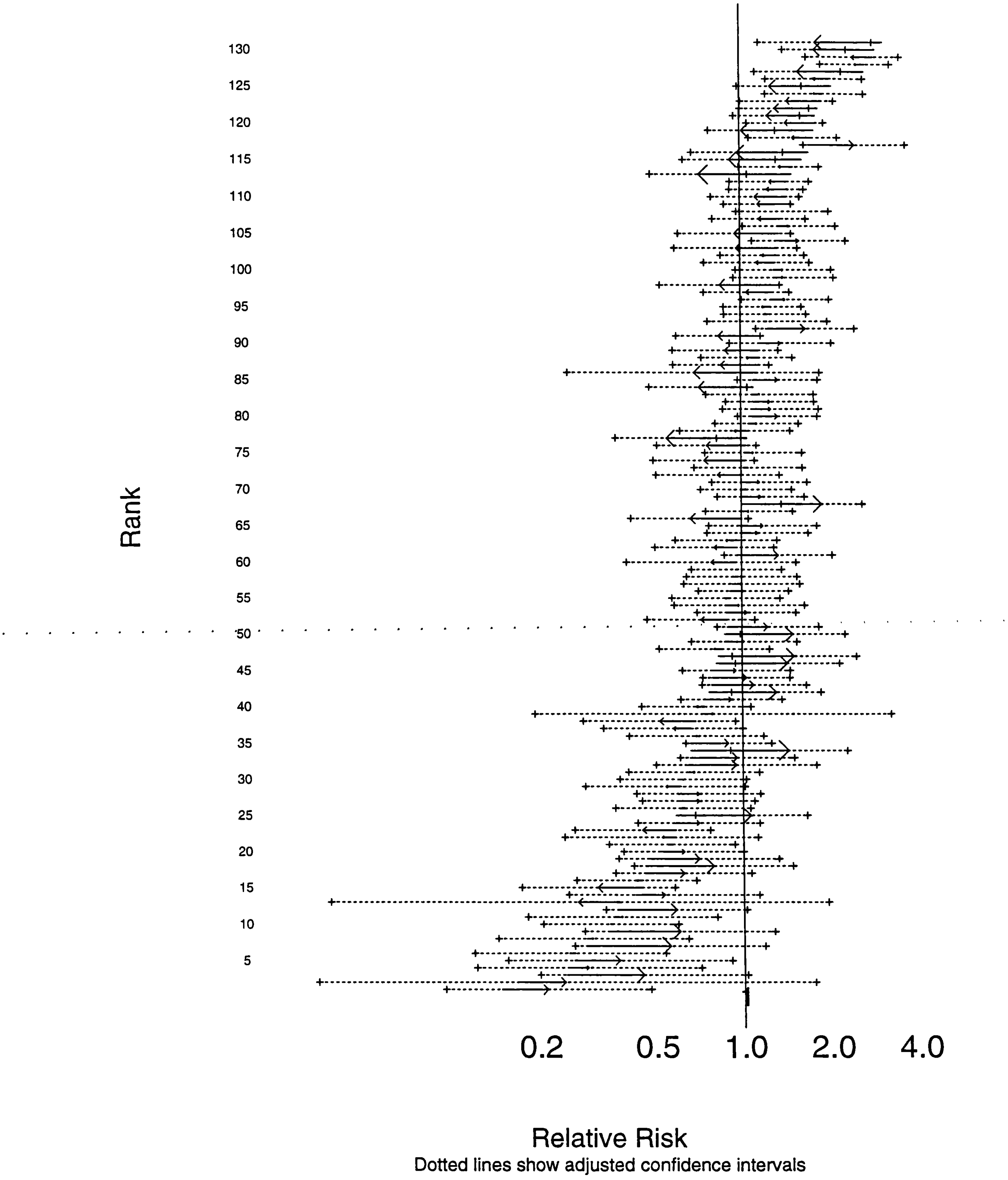


Figure 17: Change between unadjusted and adjusted relative risks for ranked consultants (1992 data)

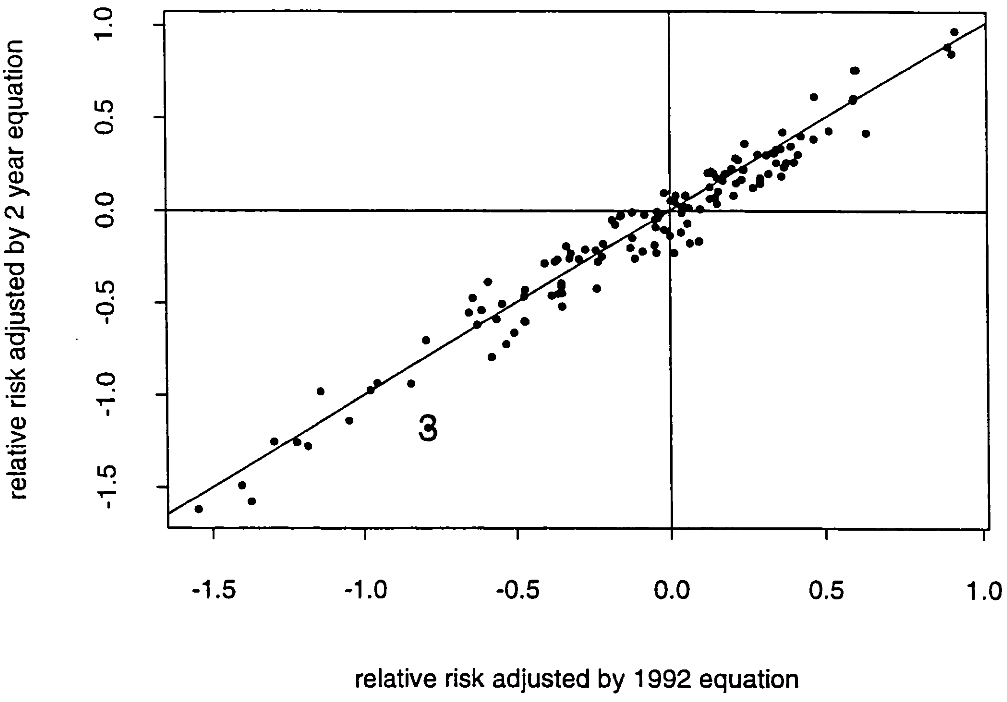


Figure 18: Comparison of Adjusted Relative Risks for 1992 data, using 2 different equations

4.5 Comparisons between 1991 and 1992

In order to evaluate the techniques, and to give an idea of the changes in performance and case mix over the two years studied, those surgeons who submitted data in both years have been investigated separately. There were 64 of them, and again the data were summed for the two years using a FORTRAN program. The same triangle plots as before were drawn separately for each year, and then for the totals over the two years. To show the change in distributions over the two years, plots were drawn with arrows from the 1991 to the 1992 position in the triangle for each consultant. The total mortality rate for the two years was considered and then the plots split up into whether these rates were above or below the median (Figures 19 and 20). Those with higher mortality rates tend to lie further towards the right in each triangle. The variability in change in proportions in each category from year to year is quite high, so we cannot expect all the adjustments to have the same effect in 1991 and 1992, although if they make sense and give some idea of surgical competence we would hope for some degree of compatibility between the adjusted values from year to year.

If we consider the rank order of the consultants between the two years, we find that Spearman's rank correlation coefficient is higher for the unadjusted values than for the adjusted ones (0.69 *versus* 0.56). Pearson's correlation between the actual relative risks for the two years is also higher for the unadjusted values (0.64 *versus* 0.56). This is as one would expect, as mostly the case mix variables are fairly similar for each consultant from year to year, especially for distribution of admission status as can be seen from the prevalence of shorter arrows in Figure 19, and so adjusting for these will remove some of the correlation between mortality rates of the two years. The fact that the adjusted values are still highly correlated may tell us that the consultants have performed fairly consistently, or that the model has not accounted for some important characteristics of their patients. Most of the consultants had a change in adjusted relative risk from 1991 to 1992 of around 25% of their 1991 value.

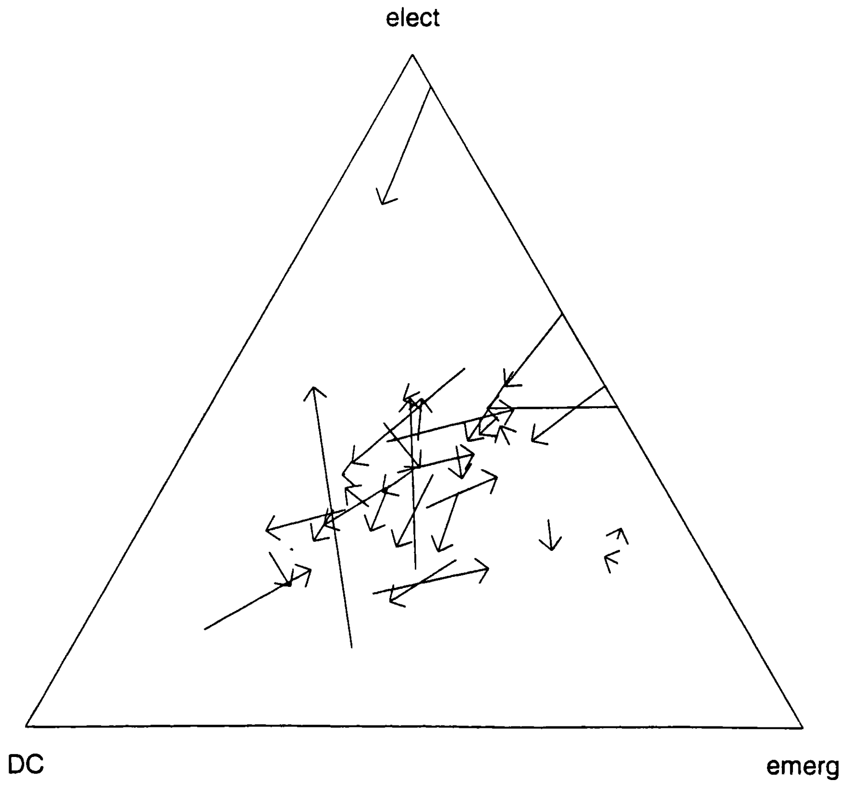


Figure 19(a): Change of Admission Status Distributions from 1991 to 1992 for consultants with lower than median mortality rate over the 2 years

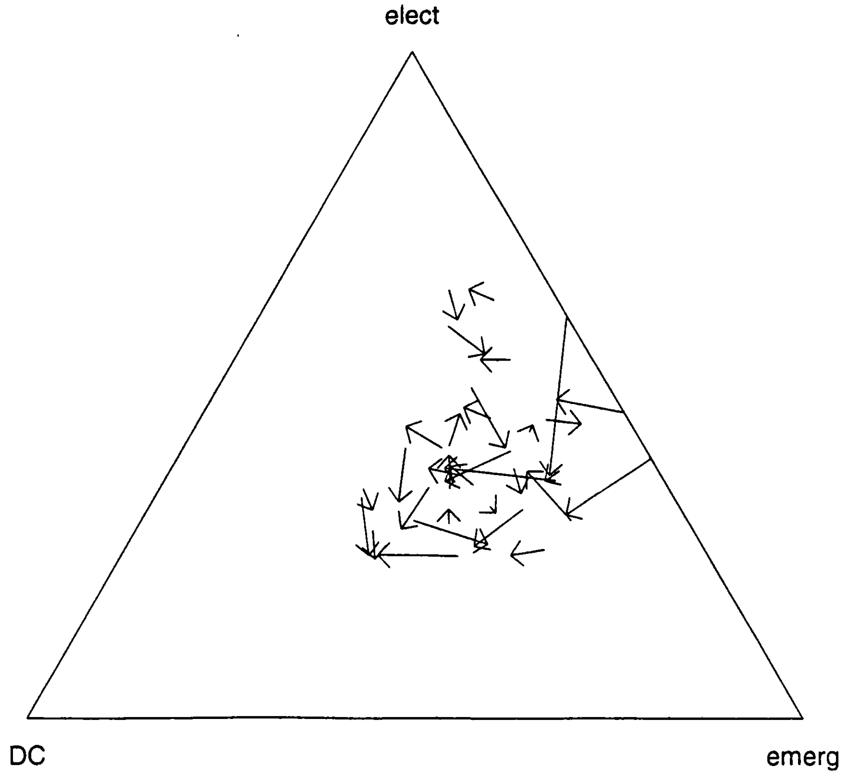


Figure 19(b): Change of Admission Status Distributions from 1991 to 1992 for consultants with higher than median mortality rate over the 2 years

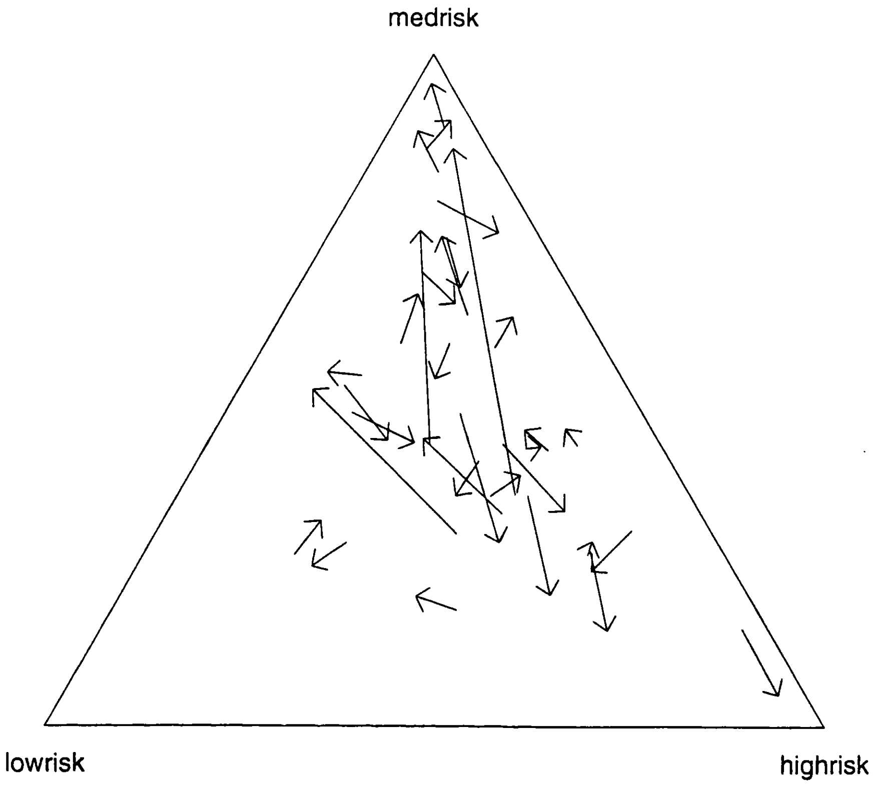


Figure 20(a): Change of Diagnostic Group Distributions from 1991 to 1992 for consultants with lower than median mortality rate over the 2 years

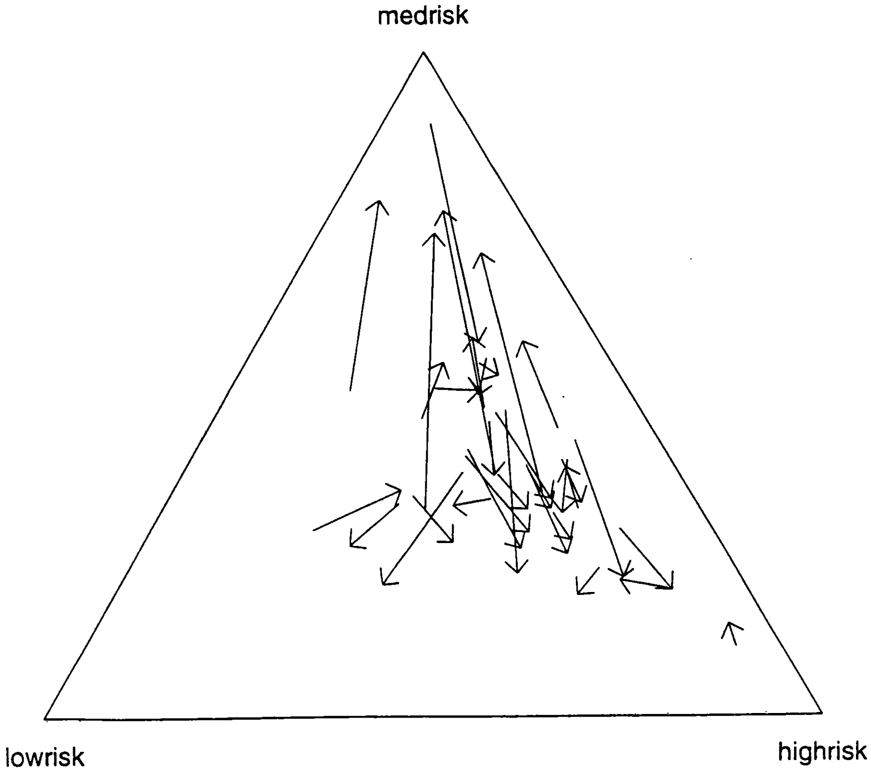


Figure 20(b): Change of Diagnostic Group Distributions from 1991 to 1992 for consultants with higher than median mortality rate over the 2 years

4.6 Discussion

Several new methods of presenting the Royal College of Surgeon's Comparative Audit data have been described. There are many limitations of these data, which still require to be overcome. The first of these is quality of the data, in that values are completely missing for many consultants and that many have not recorded information on all patients. This problem should lessen in future years as computer packages become more user friendly and the input of data is adopted as a standard procedure. Also, having taken part in the audit, consultants would know what data to collect for the following year. Another area in which quality could be improved is in standardisation of the data. For example there is no standard definition of postoperative mortality, which could just have been recorded as death while still in the unit, or 30 day mortality. The second limitation is the quantity of data, where the number of consultants responding has been less than one fifth of those eligible, but a great amount of data has been requested from each one. As interest grows in the scheme, as it must with recent media publicity, and as more units acquire computing facilities, more consultants should participate, thus providing a larger database with which to calculate more reliable models. The third, and main, limitation of the data is, of course, the fact that it is pooled for each surgeon. This makes analysis difficult, and much less accurate than if patient level data were available, as was discussed in section 4.3.

We would recommend that, in order to improve data quality, less information should be requested, and the importance of accuracy stressed. Stricter definitions of the information required should be circulated with the questionnaire. It would be preferable to have some patient level information, or some sort of breakdown of information, for example giving patient diagnostic group by age. At the very least it would be useful to know the case mixes of patients undergoing surgery separately from those who do not.

Obviously, we can never account for all possible contributory factors, but these methods are a step in the right direction and go some way towards correcting for major sources of variation. The information that a surgeon is "performing worse" does not mean that they are negligent or less able, only that some investigation into why this is occurring should take place.

The outcome of this work is to present a meaningful summary of a large amount of data to a group of consultants. This involves giving their mortality rates, both unadjusted and adjusted for different factors, in order that they can receive some idea of their performance compared with their contemporaries. An example of the individual printout of the data, with which each surgeon at the meeting would be supplied can be seen in Appendix 3. The adjusted values for the variables on their own have been included in order to give information to those who did not have complete data for each factor. It will be necessary to continue with many of the old presentation methods, as there are many things, for example work load, which we have not considered, but which are of great interest. This work has been carried out for the 1993 data, and will be discussed in chapter 5, along with the consultants' reactions to the methods described in this chapter.

5 PRESENTATION OF RECENT RCS DATA

5.1 The 1994 Meeting

The data collected in 1993 were presented to a meeting of consultants in June 1994. The original RCS ranked bar chart presentation was given first, which consisted of a booklet of charts, as exemplified in figure 3, covering all the information submitted. We then introduced triangle plots of admission status, age and diagnostic risk groups, and confidence interval plots for unadjusted and adjusted relative mortality rates. A personalised printout was given to those who had submitted the necessary data. The consultants then had an opportunity to discuss the results, and were given a questionnaire asking their opinions on the new presentation methods. The questionnaire can be seen in Appendix 2. A similar form was sent to those consultants who did not attend the meeting, but who did receive a personal data sheet. The response rate to the questionnaire was approximately a quarter, with 39 consultants responding. The questionnaire asked for ratings from 1 (poor) to 5 (excellent) of the clarity and usefulness of the three above mentioned presentation methods, as well as comments on the results. The new presentation methods were described as experimental, as the results are by no means accurate and could easily be misconstrued. Each method will be discussed in turn, followed by a discussion of criticisms of the methods and how possible improvements could be made.

5.2 The Data

The data received from the Royal College of Surgeons contained 221 records, but 2 of these were duplicates and a further 30 contained no information apart from a consultant number. The data collected were similar to those of previous years, except for some more detailed questions on outpatients, and ASA grades were requested for the first time. The ASA grades could be useful in modelling outcome, but only 45 of the consultants were able to submit this information. Of the 189 consultants remaining, 150 had reported the total number of deaths. Where only one type of death (non- or post- operative) was supplied, it has been assumed that there were no deaths in the other category. Of these 150 consultants with mortality data, 148 had complete data on admission status, 136 on age and 140 had diagnosis information. The mean number of admissions per consultant was 1314, ranging from 26 to 3115. It was questionable as to whether the consultant with only 26 admissions should be included, but there were no other irregular factors about his data to justify exclusion. The overall mortality rate was 2.0%, with a minimum of 0.0% and a maximum of 5.5%. Admissions were approximately evenly distributed between day case, elective and emergency, and again there were very few patients aged under 11. The most common diagnosis was colorectal with a mean of 169 admissions per consultant (20% of known diagnoses), closely followed by urological (15%), and the least common overall was endocrine, with a mean of only 23 admissions (under 3%), and a median of 7. In all cases the mean number in the diagnostic groups is much higher than the median, suggesting the distributions of numbers in each group are positively skewed.

As in previous years, there were several discrepancies in the data. The numbers in each of the admission status groups did not add up to the given number of admissions for approximately one fifth of the consultants. For age groups, this figure was over a half, and for diagnostic groups none of the figures added up due to the groupings not being comprehensive. The assumption that the known proportions represented the overall ones was made, as discussed in chapter 3. In several cases, the difference between the total of the three age groups and the given number of admissions was exactly equal to the number of day cases. The reason for this is probably that some systems (notably *Micromed*) do not include day cases as admissions, so this information is not available for these patients.

5.3 Triangle Plots

The triangle plots from the 1993 data can be seen in figures 21(a) to 21(c). The presentation slides were in colour, with red representing those below the median mortality rate and yellow representing those above it. The plot of admission status groups shows that most of those with the highest mortality rates lie towards the right of the triangle. They had more emergency admissions and less day cases. As in previous years, there is very little spread of age. There is a good spread of diagnostic groups, but these appear to bear little relationship to mortality. This is due to the large amount of heterogeneity within the diagnostic groupings. The triangle plots were not rated very highly by those consultants responding to the questionnaire. Most found them clear and easy to understand, but few found them particularly useful. This presentation method appears to be surplus to requirements and could reasonably be ignored in future.

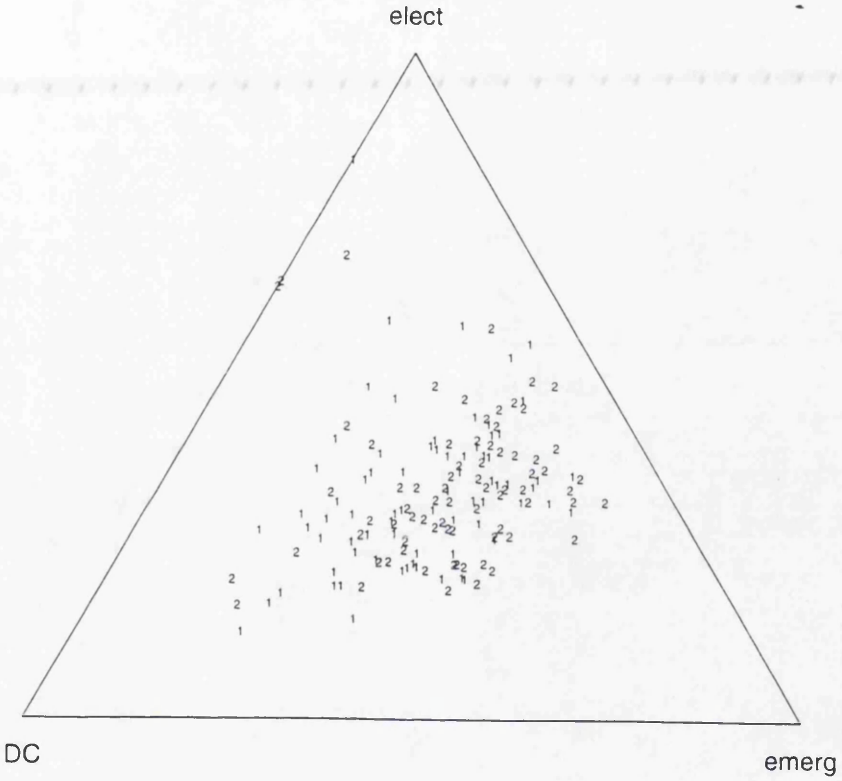


Figure 21(a): Triangle Plot Showing Admission Status Distribution for RCS 1993 Data

(1 = Mortality Rate Below Median, 2 = Mortality Rate Above Median)

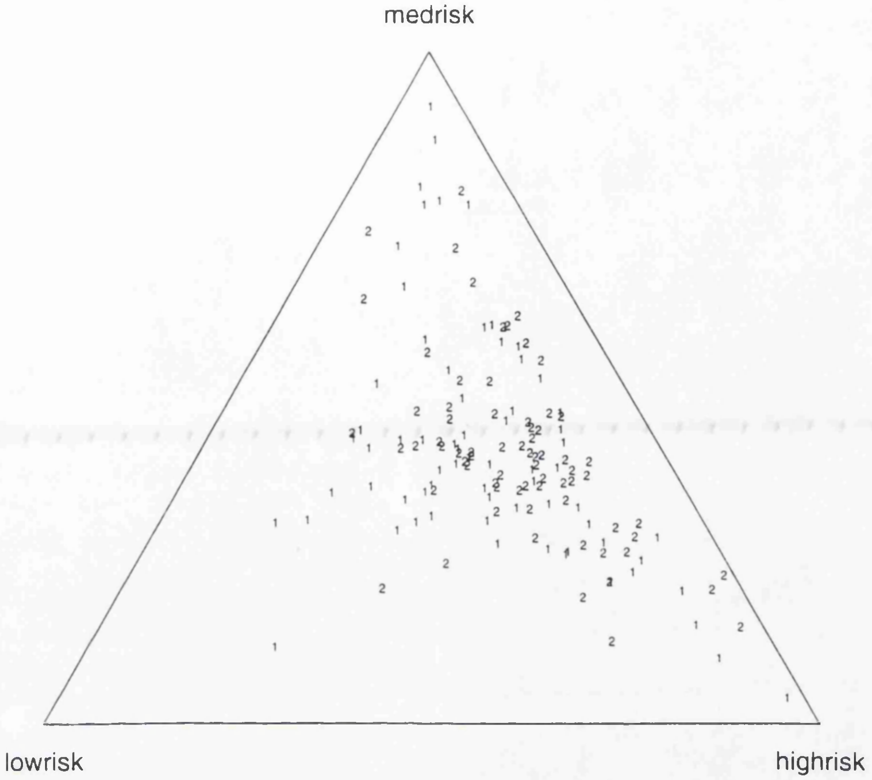


Figure 21(b): Triangle Plot Showing Diagnostic Group Distribution for RCS 1993 Data

(1 = Mortality Rate Below Median, 2 = Mortality Rate Above Median)

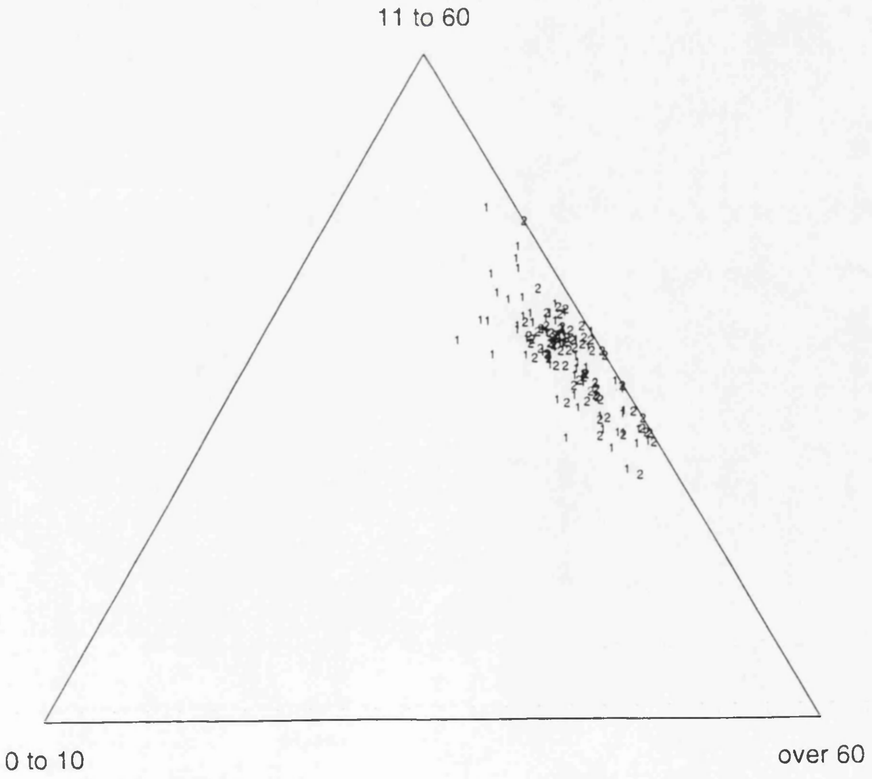


Figure 21(c): Triangle Plot Showing Age Group Distribution for RCS 1993 Data

(1 = Mortality Rate Below Median, 2 = Mortality Rate Above Median)

5.4 Confidence Interval Plots

The plots of ranked unadjusted and adjusted confidence intervals were shown separately. They have been combined in figure 22 to show the movement from the unadjusted to the adjusted value by means of arrows. The confidence intervals shown are the adjusted ones. Complete data were available for 128 consultants. The intervals have not been plotted for the three consultants with no deaths, although values were estimated for their individual printouts, as will be described in the next section. The confidence interval plots were regarded as slightly less clear than the triangle plots. Some consultants did not grasp the concept of "relative mortality", or how this could be related to case mix. However, the majority found them very clear. These plots were perceived as useful, and rated more highly than the triangle plots on this point.

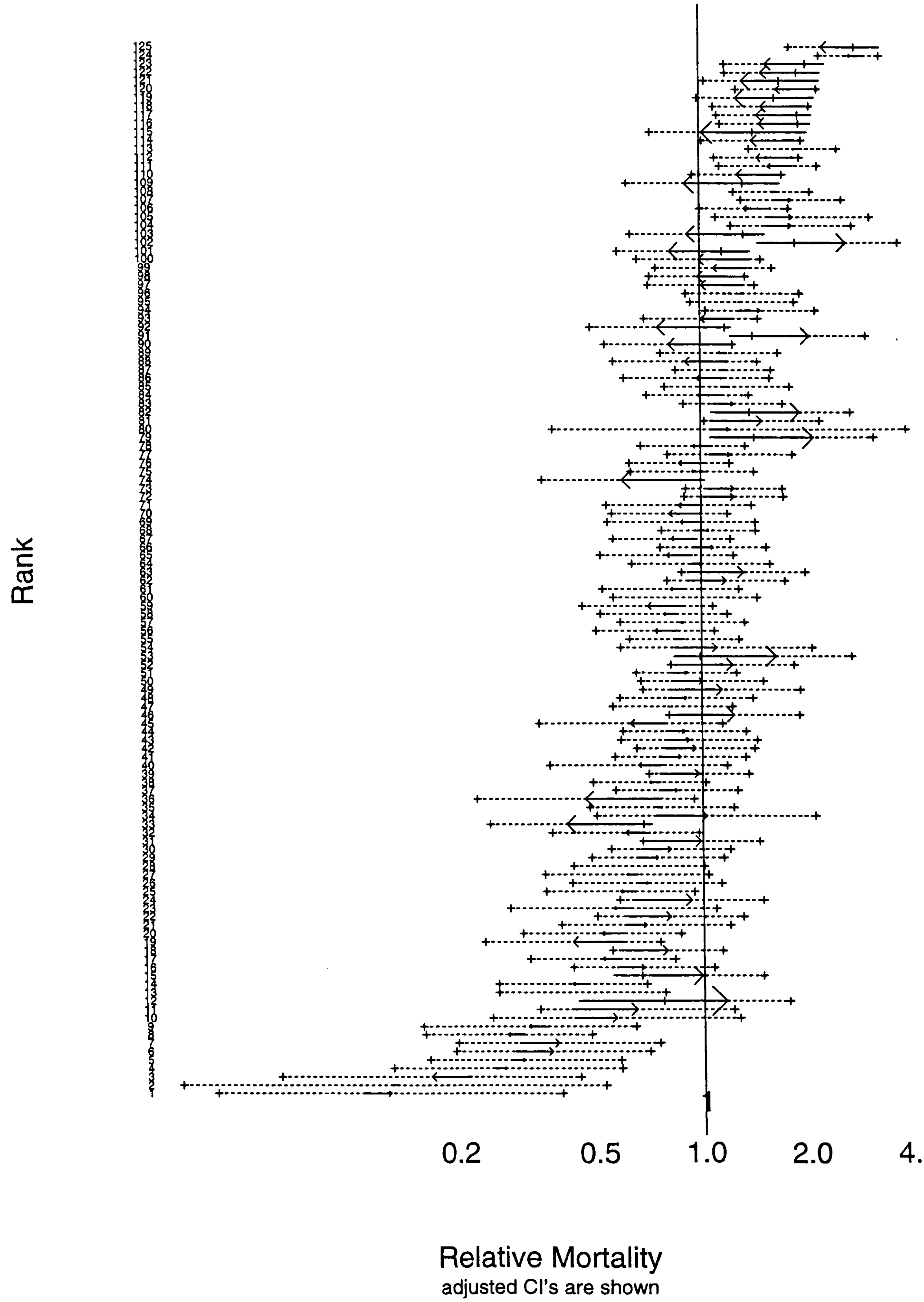


Figure 22: Change between Unadjusted and Adjusted Relative Mortality Confidence Intervals for Ranked Consultants (1993 RCS Data)

5.5 Personal Printouts

The 150 consultants who submitted mortality data were given a personalised printout of this experimental analysis. A specimen printout from 1994 can be seen in Appendix 3. The printouts for the 1993 data were similar, except the numbers of non-operative admissions and deaths were not included, and rankings were only given for adjusted relative risks for all the variables. The adjusted values are the ratio of the observed and predicted log odds, as was described in chapter 4. The unadjusted ranking was out of the entire 150, whereas the adjusted one was out of the 128 with all the necessary data. The adjustments for individual variables were calculated using all the available data for those variables. The logistic regression equations used to calculate the predicted log odds are as follows.

1.	**Age**	$0.6\times \text{age61pl} - 4.4\times \text{age0to10} - 3.9$
2.	**Admission Status**	$0.8\times \text{emerg} - 1.0\times \text{day case} - 3.9$
3.	**Diagnostic Risk**	$0.3\times \text{highrisk} - 0.4\times \text{lowrisk} - 3.9$
4.	**All Case Mix**	$1.1\times \text{age61pl} - 4.3\times \text{age0to10} + 1.4\times \text{emerg}$ $- 1.1\times \text{day case} + 0.5\times \text{highrisk} - 4.5$

where age61pl and age0to10 are the proportions of patients aged over 60 and under 11 respectively, emerg is the proportion of emergency admissions, day case the proportion of day case admissions and highrisk and lowrisk refer to the proportions of known diagnoses in each of the risk groups as previously defined.

The proportion aged under 11 has a very high weighting, as the very small numbers of patients in this age group lead to high variability. The standard error of this coefficient is 0.8 whereas the other variables have coefficients with standard errors of around 0.2. Where a consultant has a relatively large proportion of patients in this age group, it has a disproportionately large effect on the adjustment. A cut off point of 40 years for the middle age group would be preferable. All the coefficients have the intuitively correct signs, and so have the expected effect on adjustment. The coefficients for diagnostic risk are very low, so adjusting by diagnosis did not change the rank order as much as by the other variables. The effect of these adjustments could still be surprising to some consultants. For example, if they had performed a large number of haemorrhoid operations, these would be included as colorectal, and thus high risk .

The confidence intervals for those three consultants reporting no patient deaths were calculated using a binomial distribution result. If θ is the probability of death for a consultant admission, the 95% CI for θ goes from 0 to the root of $(1-\theta)^n = 0.05$. The upper limit for relative mortality was then calculated using the ratio of this solution to the overall mortality rate. The adjusted values were calculated by dividing by the predicted mortality rate obtained from the logistic regression equations.

The personal printouts were considered the easiest to digest and the most useful of the new methods by nearly all consultants. There was still some trouble with interpretation, with one person who did not understand what the confidence interval figures meant. Of the people answering the question, more than half said that they would perform some form of investigation in the light of their results, with all but one saying they would like to see the methods developed for future use.

5.6 The 1994 Data

The data received in 1994 were less complete than in previous years. Data were received from 128 consultants, of whom only 89 supplied data on numbers of deaths. Of these, 82 had data on admission status, age and diagnosis. The problem of categories not adding up to the total numbers of admissions had not diminished any from previous years. Several consultants had numbers in the three age groups summing to the number of admissions minus number of day cases, and others simply had data missing.

Individual printouts, as shown in Appendix 3, were presented at the meeting in June 1995, and were met with some enthusiasm by those in attendance. Improvements to the printout from the year before were adding the numbers of non-operative patients and deaths for information, and ranking the adjusted values for all the case mix variables.

This time, models calculated from 3 years' data (1992 - 1994) were used to adjust the relative risks. It was decided to use models from 3 years' data rather than 1 because of the smaller number of consultants with data in 1994. The data were assumed to be independent, so a consultant who had submitted data each year would be included three times. This is in contrast to when the 1991 and 1992 data sets were combined as then the same consultants' data were added together for the two years. This should not greatly affect the results, as most of the difference is caused by case mix, and increased numbers of consultants to model with is of benefit when calculating a model. The models are:

1.	**Age**	1.1×age61pl - 4.4
2.	**Admission Status**	1.0×emerg - 1.0×day case - 4.0
3.	**Diagnostic Risk**	0.4×highrisk - 0.3×lowrisk - 4.0
4.	**All Case Mix**	1.6×age61pl + 1.2×emerg - 1.2×day case + 0.7×highrisk - 4.9

Some of the adjusted relative risks had counter-intuitive values, given the particular case-mix. The reason for this is the use of a different data set to calculate the model, so the adjustments are relative to a different mean mortality rate, as described in chapter 4. This effect was particularly noticeable for the diagnostic group adjustments, which were very sensitive to this due to the small coefficients in the model. For example, one consultant had an unadjusted relative mortality of 100%. With his diagnostic group case mix of only 28% high risk diagnoses compared to the average of 40%, one would have expected his value adjusted for diagnostic risk would be higher than the unadjusted one. However, without making any correction it was 93%. The mean mortality rate of those having data on diagnoses in 1994 was 1.73%, whereas over the 3 years, the value was 1.96%. This means that the unadjusted relative risks were calculated relative to a lower mortality rate than the adjusted ones and thus the values are too high in comparison. Alternatively, we could say that the adjusted values are too low, because the predicted values are based on a higher mortality rate. We could therefore make a correction for this effect in two ways, either by changing the way we calculate the unadjusted risk, or by correcting the adjusted one.

The first of these approaches simply involves dividing the observed mortality rate by the mean predicted mortality rate instead of the mean observed mortality rate to obtain the unadjusted relative risk. This value does not depend on the particular case mix of the consultants in the study. The overall mean predicted mortality rate for the 1994 data using the 3 year model was 2.00%. Thus our example consultant, who had a mortality rate of 1.76% would have an unadjusted relative mortality of 88% instead of 100%, which makes sense when compared with the value adjusted for diagnostic mix.

Strictly speaking, we should have a different unadjusted value for each case mix variable as the different models give differing overall mortality rates due to the inclusion of different subsets of patients with available data. For example, the mean predicted probability of death considering only diagnostic casemix is 1.97%, which would give the above consultant an unadjusted relative risk of 89%. In most studies, however, there will only be interest in adjustment for overall case mix, so this technicality will be irrelevant.

The other approach, which in this study removes the inaccuracy caused by using different models without having to consider separate unadjusted values for the subsets of patients with data on each case mix variable, is to correct the adjusted values. This can be done by multiplying the predicted probabilities by the ratio of the mortality rates from the current data and the model data set, which in this case is 0.881. Thus the adjusted relative risk for consultant i becomes the exponential of

$$ln\left(\frac{d_i}{a_i}\right) - ln\left(\frac{0.881 \times p_i}{1 - 0.881 \times p_i}\right),$$

where p_i = predicted mortality rate for consultant i (obtained via above equation). Thus our example consultant would have a relative risk adjusted for diagnosis of 106%.

In fact, an approximation to this is far easier to calculate, and gives very similar results. The adjusted relative risk is simply multiplied by 1/c, where c is the ratio of the two relevant mortality rates (current data/model data). The values of c for the above data are 0.876 for age, 0.887 for admission status, and 0.880 for the overall model. Using the approximation rather than the above value means the relative risk differs by a factor of (1-p)/(1-cp), which is very small. When we quote the results to the nearest percent, it gives the same value in almost every case as multiplying the probabilities. However, if there was a very large difference between the mortality rates of the two samples, this might make more of a difference. It is, however, unlikely that one would wish to model two such greatly differing groups of consultants in the same way.

If using a pre-existing model such as POSSUM or APACHE, one is unlikely to know the mortality rate of the population on which the model was calculated, and the second method would be inappropriate. An obvious extension of the first method is that individual consultants who are not include in a comparative audit study, could calculate their own values using the model.

5.7 Discussion

The main criticism of the results by the consultants was the lack of differentiation between "inevitable" and "preventable" deaths. However, as stated in chapter 2, this type of definition is highly subjective, and the information is likely to be very sensitive. If the consultants were to judge which deaths were inevitable, it would defeat the purpose of the audit, and of attempting to model outcome. It may be more instructive only to consider postoperative deaths, although this could still excuse differences in surgical skill. At any rate the data as they are collected at present do not allow for these to be considered separately. This is because the case mix information applies to all admissions, and there is no break down of characteristics into those having operations or not. It would be possible to consider operative groups and postoperative mortality or morbidity, but we would have to make the unlikely assumption that the overall spread of age and admission category represented the spread of those patients receiving operations. Also, the data for operations are collected as numbers of procedures, so one patient admission could feature several times. A more appropriate classification to use would be the BUPA classifications (minor,.....,complex major) or the ASA grades. Most consultants are in agreement that diagnoses are inappropriate for this modelling. A change in data collection would be required in order that information was available separately for operative patients to make use of the BUPA classifications, and ASA values are very difficult for most units to collect. For the types of procedures, one is recorded per theatre visit per patient, so there is still the problem that numbers of operations for those receiving operations, exceed numbers of admissions. Considering operative procedures at present still excludes those patients attending for treatment such as chemotherapy or radiotherapy. Consultants are highly aware of the large differences in policies on admission of terminally ill patients, which depend on the adequacy of local hospice provision. Until some consideration is made of those patients having no operation, they will not have any confidence in the Comparative Audit results.

Most hospitals had some difficulty collating the data as requested by the RCS. The majority who responded to the questionnaire either had their own computer system or gathered the results from the HAA data. Some even did the work by hand, and many scraped the data together from a variety of sources. The most popular commercial system among respondents was Micromed. This could be because the original study was based on this package, so data collection is easier. Originally the Comparative Audit meeting was combined with the Micromed user group AGM, and this could have encouraged interested consultants to choose that particular package.

Of the consultants who answered, half said they thought the new presentation methods were more informative than the original method, and all the rest but one said that they were equally informative. There is a definite demand for these methods to be developed for future use, but there will need to be major changes in the data collection. A list of recommendations to the RCS can be seen in the following box. A review of data collection methods in individual hospitals is also required, and standardisation introduced.

Consultants were asked what information they would like to see presented in future, and if they had any suggestions or comments. These have been used in this discussion. A selection of quotes from the questionnaire are given in Appendix 4.

Recommendations for Royal College of Surgeons Comparative Audit future data collection, if methods are to be developed further.

- Collect case mix information separately for operative and non-operative patients.
- Diagnostic groupings should be comprehensive.
- Change diagnostic groups so that they are more homogeneous for risk
- Use age 40 as the cut off point for the low risk age group rather than 10.
- Include treatments which are not operative procedures (e.g. chemotherapy).
- Emphasise collection of data on BUPA classifications

6 INVESTIGATIONS OF AGGREGATE *VERSUS* PATIENT DATA

6.1 Introduction

The data collected by the Royal College of Surgeons Comparative Audit Service are in the form of totals for each consultant for the year. This means that there is no way of assessing the relationships between variables, for example it is unknown whether particular diagnoses are more common in patients in a certain age group, or if emergency admissions are restricted to particular types of operation. The associations between the case mix variables and outcomes can only be estimated using the proportions of patients in each category. Obviously, much information is lost from the original data on individual patients, and it would be informative to know the effect of this on the actual results of our analyses. This has been investigated in two ways, using an actual data set (Colorectal Cancer Study), and by simulating patient data.

6.2 Colorectal Cancer Study

We obtained the colorectal cancer data from the study organisers, who studied variability among surgeons in their publication (McArdle and Hole, 1991). The data were collected over six years at Glasgow Royal Infirmary, and had 10 year follow-up information. In the original publication, the 13 consultants were compared for patient survival using Cox's proportional hazards to adjust for various factors. Significant differences were found between the surgeons, with three of their hazard ratios being significantly different from 1.

While the colorectal cancer study used survival up to 10 years as the outcome measure, for this investigation we use 30 day survival, as it is similar to the mortality information in the RCS data. This outcome measure was calculated from admission and death dates. Also, so that all the data could be utilised, patients of "consultant 14" here are in fact all those patients for whom surgeon was unspecified.

In the proportional hazards analysis the patients were divided into groups depending on the type of operation they had (curative resection, palliative resection, palliative diversion) and the calculations done separately. We used all 645 patients in the study together for this investigation.

The unadjusted confidence intervals were calculated as with the RCS data, using the Logarithm method. These can be seen in Figure 23.

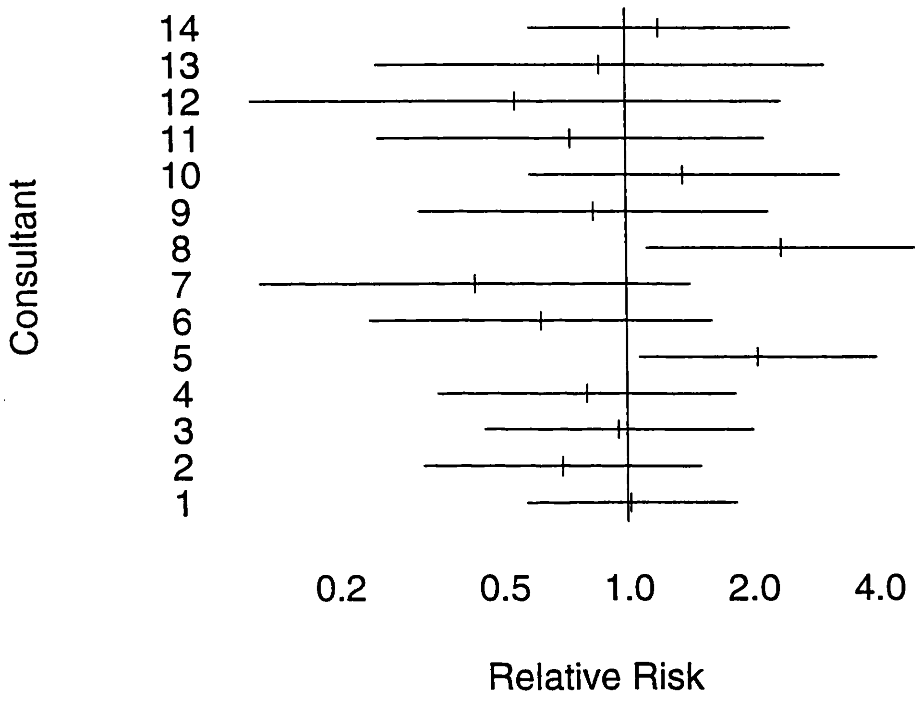


Figure 23: Unadjusted Relative Risk Confidence Intervals for Consultants (Colorectal Cancer Study)

The variables considered for inclusion in our model were those found to be significant for survival in the original study for any of the types of operation, that is sex, emergency admission, whether the patient was over 75, differentiation, Dukes' stage, local invasion of the tumour, the presence of distant metastases and pre-existing cardiac or respiratory disease. Also included was whether the treatment was curative or not. It could be argued that a curative resection may be attempted by some consultants on some patients while other less skilful or adventurous surgeons might opt for palliative care, and so should not be considered strictly as a case mix variable. However, since it is a strong prognostic factor, we shall include it for the purpose of this investigation. All of these variables were binary for the patient data, and were summed for each consultant to obtain a summary data set with numbers in each category. Separate stepwise logistic regressions were carried out for the two data sets.

With the patient level data, the significant factors were whether the patient was male or over 75, had an emergency admission or a poorly differentiated tumour, as well as whether the resection was curative. With the summary data, only curative was entered into the equation. The coefficients resulting from these stepwise regressions can be seen in Table 10.

term	Patient data	Aggregate data
curative	-1.676(0.264)	-3.208(0.209)
over 75	0.522(0.255)	0
emergency	0.694(0.232)	0
male	0.502(0.233)	0
poor differentiation	0.543(0.296)*	0
constant	-1.775(0.240)	0

* p=0.07

Table 10: Coefficients (and SE's) of terms in LR equations

It can be seen, as expected, that relationships between case mix and outcome are not so well defined with the use of summary data. In the analysis of patient level data, 5 variables and a constant came into the equation, whereas only one variable was significant with the summary data. However, "curative" explains much of the relationship between case mix and mortality. Had curative resection not been included as a candidate for stepwise selection in the logistic regression, no factor would have been significant with the summary data, and so the adjustments would not have been possible. With the patient data, the coefficients would have been as in Table 11. It can be seen that extra information is gained from the other variables, and most of the effect of "curative" is absorbed in the constant.

term	coefficient (SE)
over 75	0.59 (0.244)
emergency	0.86 (0.224)
male	-0.46 (0.241)
poor differentiation	0.50 (0.244)*
Dukes' C stage	0.50 (0.283)*
constant	-2.33 (0.231)

*0.05<p<0.08

Table 11: Coefficients of terms in LR equation for patient data, had curative not been included for selection

Looking at the adjusted plots for the Colorectal Cancer data (Figures 24(a) and (b)), it can be seen that in both cases all the intervals have moved to include 1. The values for consultants number 1 and 10 are on different sides of unity for the two types of data, but these are all values very close to 1 and the confidence intervals show considerable overlap. In no case would different conclusions be drawn about a consultant if aggregate data were used rather than patient data.

This study has given fairly good predictive models from both patient and summary data, and we would hope to be able to achieve a reasonable model with the Royal College of Surgeons' data. However, here only one type of procedure is being considered whereas in the RCS study we can be far less specific as we are dealing with the whole of general surgery. Also, the Colorectal Cancer study has a higher postoperative mortality rate than the one for general surgery (16% compared with 2%) and so the outcome is more relevant and thus easier to model. However, the RCS data involves much larger numbers of patients and consultants so it should be possible to achieve a reasonable model. We did not have access to data from any study with patient level data on as many consultants as we have from the Comparative Audit Service so it is necessary to create some data which are comparable with those of the RCS, to explore the effect of having no patient information on this scale.

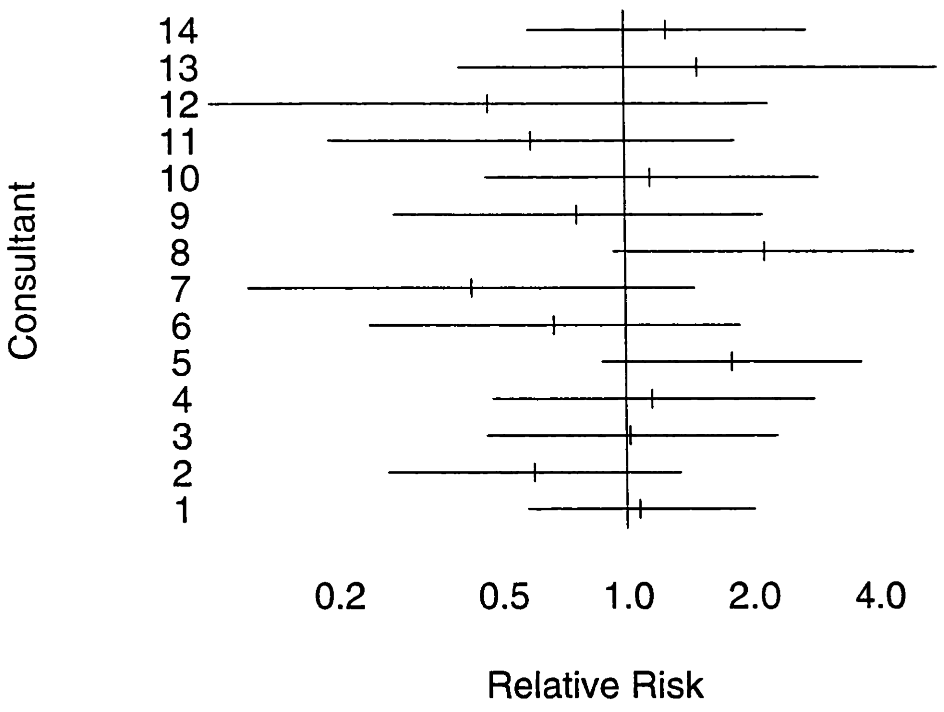


Figure 24(a): Adjusted Relative Risk Confidence Intervals - Individual Patient Data (Colorectal Cancer Study)

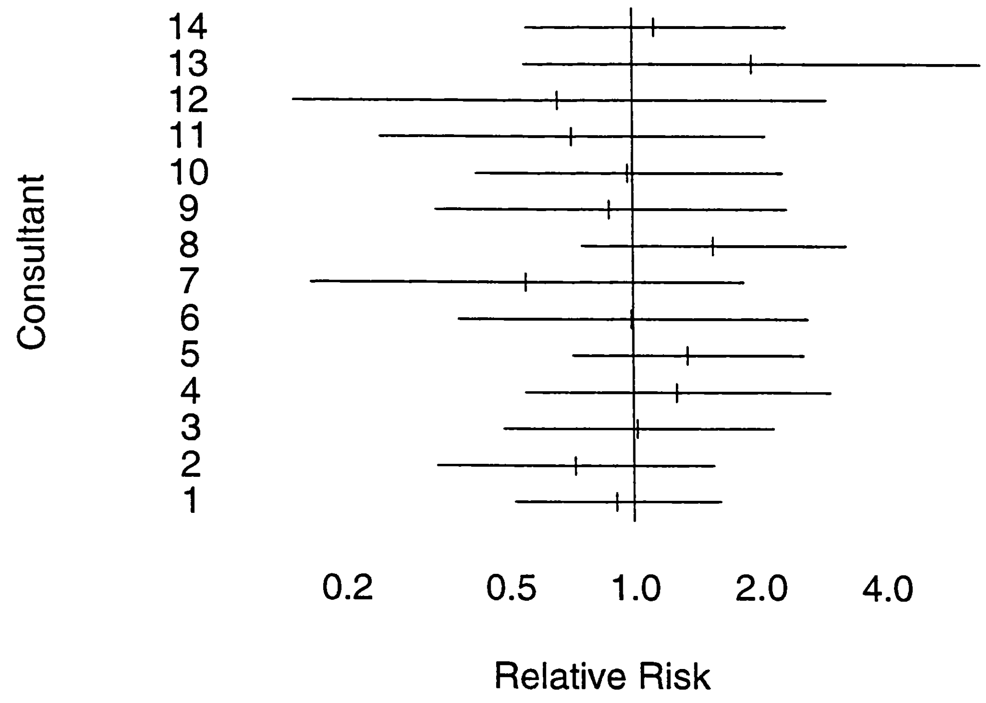


Figure 24(b): Adjusted Relative Risk Confidence Intervals - Summary Data (Colorectal Cancer Study)

6.3 Simulated data

We wished to produce simulated patients who would give similar totals in each category as were present in the RCS data sets. In order to do this, we required some method of parameterising the relationships between the factors, so that the joint probability distribution could be obtained from the marginals. This was approached in two ways, firstly by log-linear modelling, and secondly by a far more simplified method involving contingency tables. In this section, the various approaches to generating patient data that were considered while developing an efficient method will be described, as well as some interim analyses. When the method was finalised, more analyses were carried out, and the differences between the adjustments using patient and aggregate data explored, as well as the effect on these of differing correlation structures.

For these simulations, only two case mix variables were considered; age group and admission status. This is because it was far easier to visualise the inter factor relationships with two than with three variables. Initially, marginal totals in each category were generated from Normal distributions based on the distributions of the 1992 data. It was then decided to sample from the actual data and generate patients which could have given these totals.

6.3.1 Development of Methods

6.3.1.1 Log-linear modelling

With two factors (age and admission status) at three levels each, an ordinal log-linear model was fitted. For this each level of each factor is given a "score" of 1 to 3. So, for example, day case receives a score of 1, elective 2 and emergency admission 3, and similarly for the three age groups. The model is

$$\log(n_{ij}) = \theta + \lambda_{s_i} + \lambda_{a_j} + \beta(i-2)(j-2) \qquad i,j = 1,2,3$$

where n_{ij} is the number of patients a consultant has with age group i and admission status j. This gives six independent parameters $(\theta, \lambda_{s_1}, \lambda_{s_2}, \lambda_{a_1}, \lambda_{a_1}, \beta)$, and there are six known marginal totals $n_{i.}$ and $n_{.j}$. Thus six equations are obtained (as in Appendix 5), which are solved for the above using a numerical algorithm. From the parameter estimates, estimates of the number n_{ij} in each of the 9 categories can be obtained. It was difficult to choose initial estimates for the parameters which would create convergence, for the particular numerical algorithm which was employed. Straightforward log-linear modelling was carried out using BMDP 4F to give an idea of the magnitude of the parameters.

If three variables were to be used, the model would involve 11 independent parameters, but only 9 marginal totals would be known, so the equations could not be solved. In order to estimate the parameters for three variables in the same way, the two-way marginal totals would be required.

Patients were generated by this method using a FORTRAN program. Firstly, suitable marginal totals were generated and used to estimate the parameters of the above log-linear model, using the NAG routine C05NBF which solves n non-linear equations in n unknowns. Unfortunately, this routine did not converge on every occasion, and the generation of patients by this method proved rather time consuming. Much exploration into the relationships between parameters and the marginal totals was carried out, but convergence occurred for some data which were very similar to those for which the algorithm diverged.

From each set of six totals, a proportion of patients in each of the nine cells could be estimated for each consultant. These proportions were summed successively in order to give a set of probabilities in the interval (0,1) so that a uniform random variable would allocate a patient to a particular category. Once a patient's admission status and age group were determined, their probability of dying was calculated via the logistic regression equation

$$\ln(p/(1-p)) = -1.5\times\text{dc} + 1.25\times\text{emerg} - 0.25\times\text{age0to10} + 1.75\times\text{age60pl} -4.5$$

where dc, emerg, age0to10 and age60pl represent the binary variables for each patient. This equation was based on previous results, and gave reasonable death rates. Whether or not the simulated patient actually died was then determined by whether another random uniform variable was less than their predicted probability of dying.

A patient data file consisting of 1's and 0's for each variable, and a summed data file were produced. Several data sets were obtained by this method, using different numbers of consultants and patients. These were then analysed. Firstly, stepwise logistic regressions were carried out to find which factors were significant for mortality for both the patient and aggregate data. Table 12 shows the regression coefficients obtained from some of these data sets. Unadjusted and adjusted confidence intervals were calculated. For the aggregate data this was done in Minitab, and for the patient data, separate logistic regressions were performed in BMDP for each consultant, and their coefficients extracted manually. All of these confidence intervals were plotted, and it was evident that the adjusted intervals calculated from the aggregated data were often different from those calculated using the patient data. This effect was less noticeable when the number of patients and consultants were higher. Examples of the plots for the simulation with 70 consultants can be seen in figures 25(a) to (c).

Patient Data **Coefficients**

number consultants	average no. patients	dc	emerg	age0to10	age60pl	constant
10	200	-1.51	1.31	0	2.15	-6.02
10	900	-1.11	1.47	-0.67	1.67	-4.56
25	200	-1.48	1.23	0	1.95	-4.62
30	600	-1.25	1.41	-1.32	1.57	-4.49
50	500	-1.66	1.14	0	1.92	-4.55
70	1000	-1.46	1.26	-0.46	1.79	-4.54
100	250	-1.55	1.28	-0.62	1.69	-4.66
Actual Values						
Coefficients in program		-1.5	1.25	-0.25	1.75	-4.5
Aggregate Data						
10	200	0	2.56	0	0	-4.00
10	900	-2.85	0	-4.79	0	-1.78
25	200	0	0	0	3.79	-4.26
30	600	0	2.44	-1.37	1.29	-4.10
50	500	0	1.32	-1.01	0	-3.31
70	1000	-0.62	2.05	-1.13	1.00	-3.72
100	250	0	2.33	-1.61	0	-3.56

Table 12:Coefficients of variables for sample patient and aggregate data sets simulated by log-linear modelling method

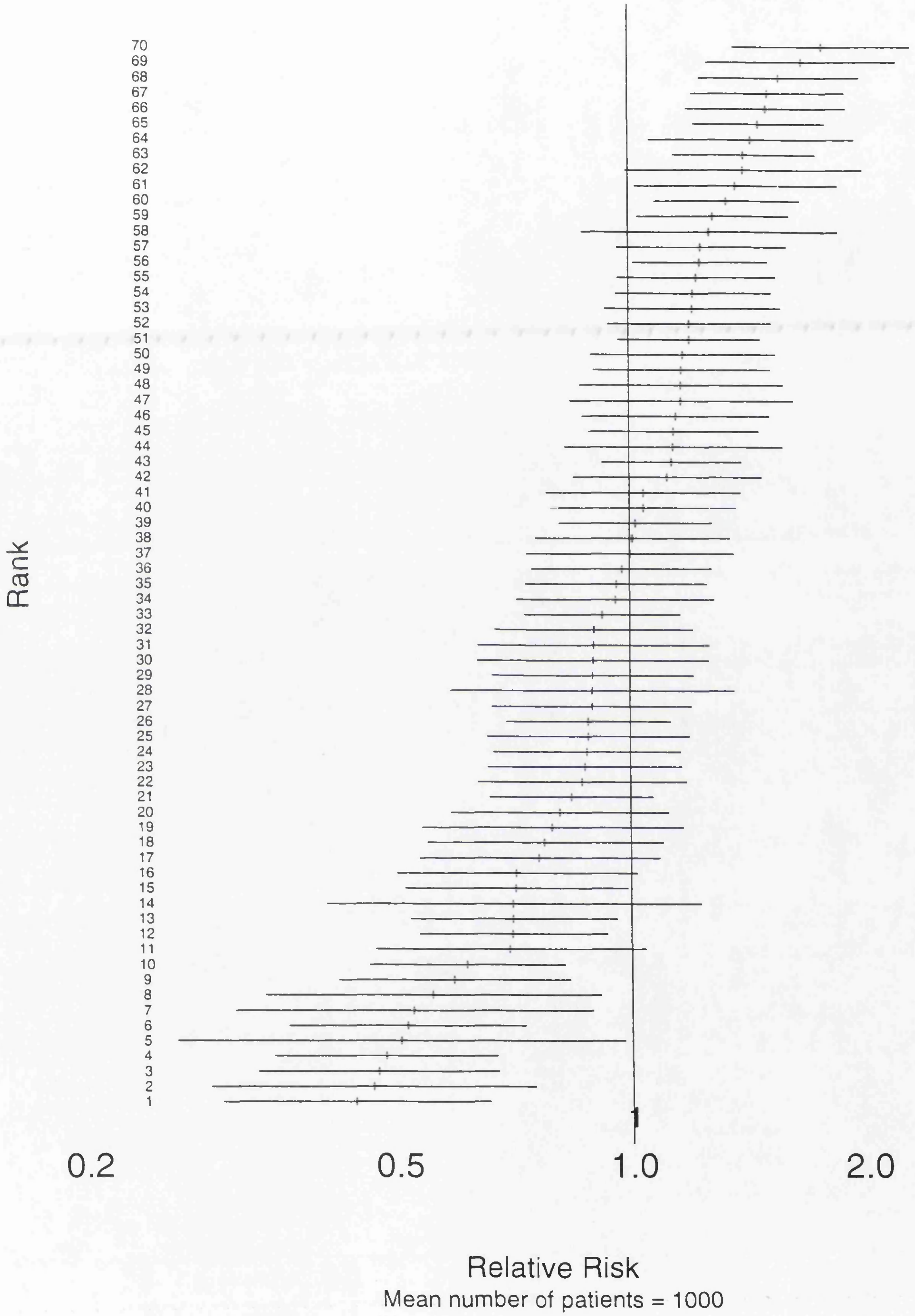


Figure 25(a): Unadjusted Relative Risk Intervals for Simulated Consultants

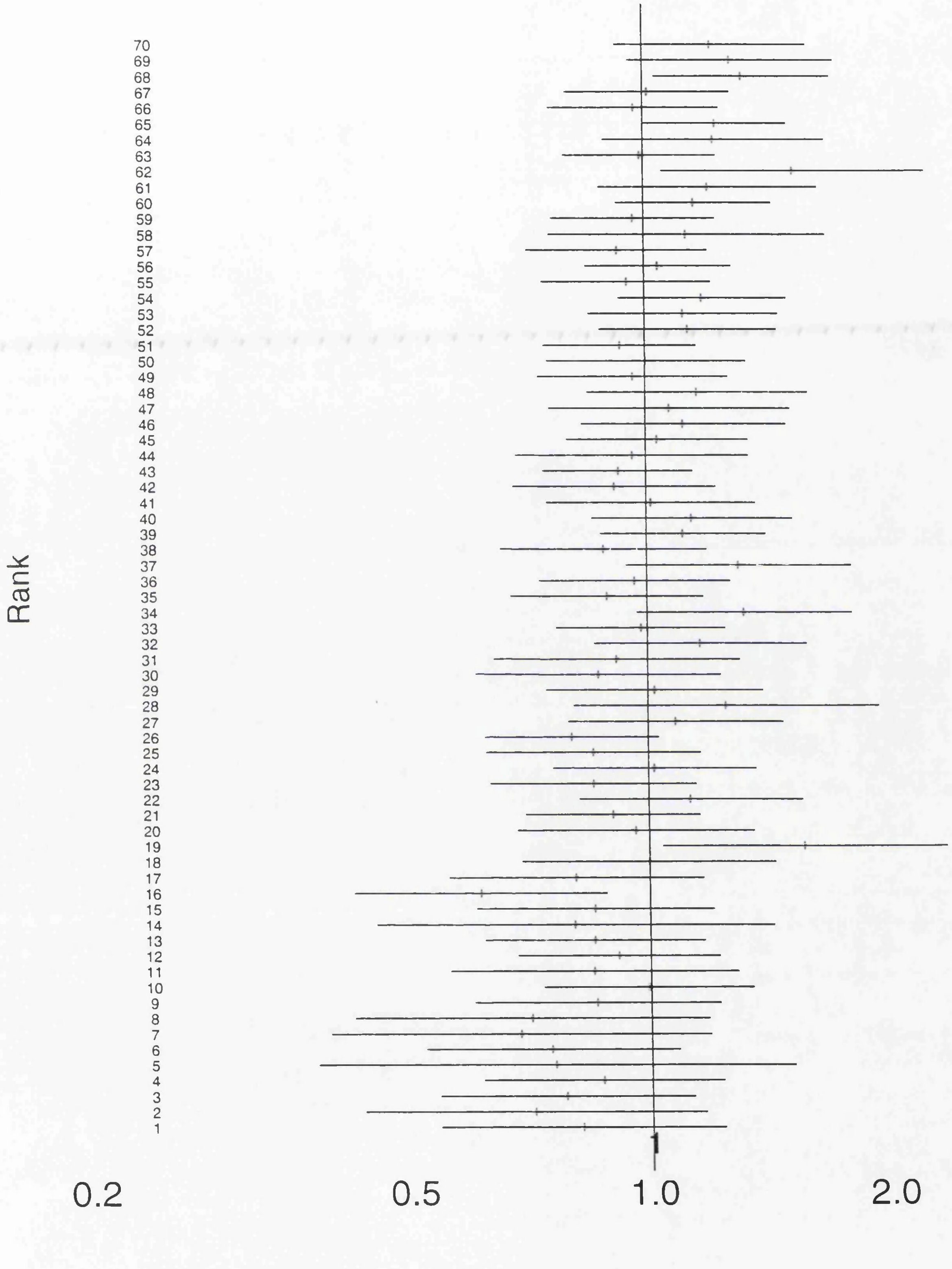
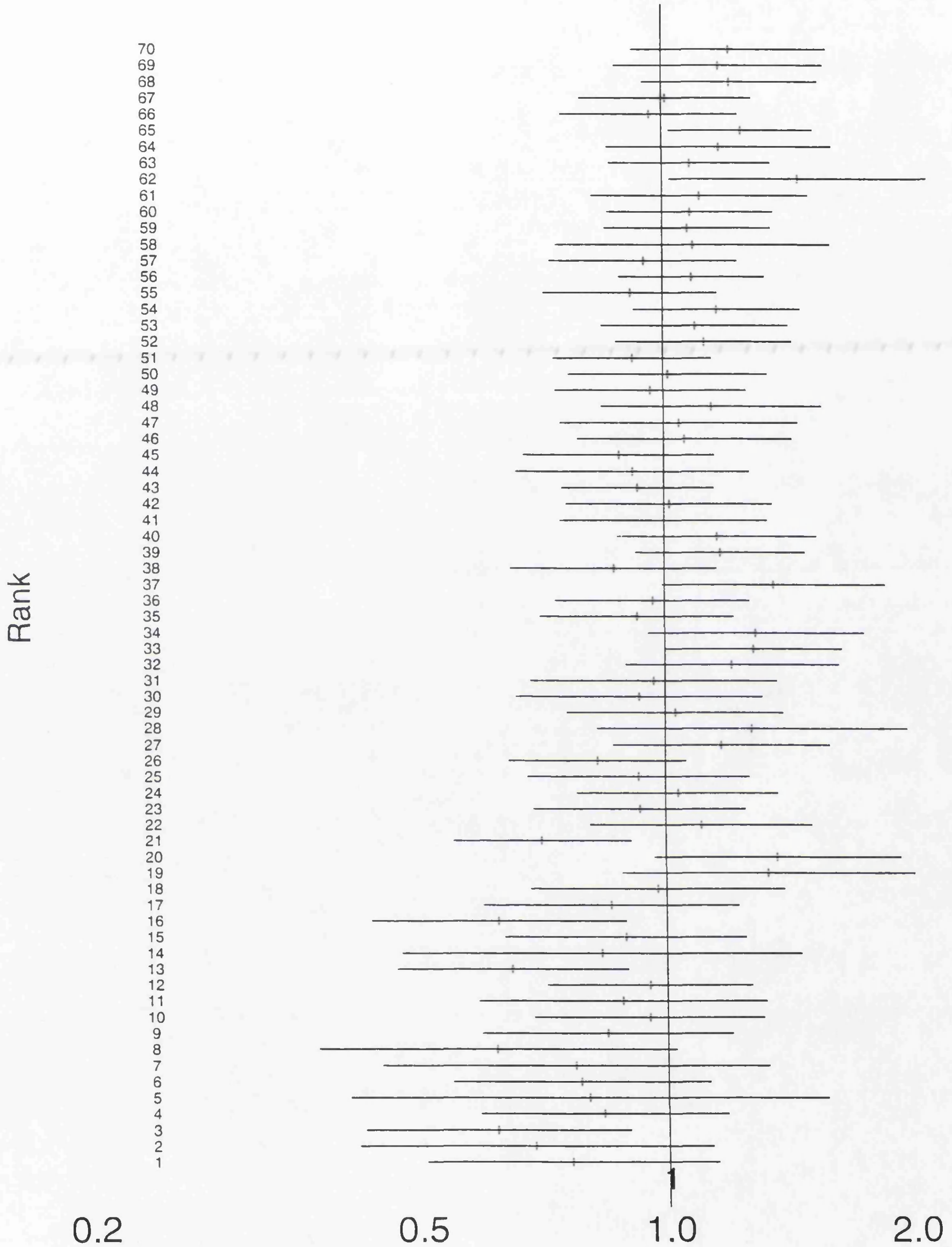


Figure 25(b): Adjusted Relative Risk Intervals for Simulated Consultants - Patient Data



Relative Risk
Mean number of patients = 1000

Figure 25(c): Adjusted Relative Risk Intervals for Simulated Consultants - Aggregate Data

For the patient data, it seems that the increased number of consultants is more important for accurate estimation of the regression coefficients than the increased patient numbers, as the data sets with 70 and 100 consultants gave values closest to those used for generation of death probability in the original program. With the aggregate data, the set with 70 consultants with an average of 1000 patients each was the only one which had all four factors significant. This size of data set is the most realistic of those considered above. However, the coefficients were markedly different from those originally used. As can be seen, much information has been lost by using the aggregated data. This is reflected in the confidence interval plots, which show the adjusted values for the aggregate data as quite different from those of the patient data in several cases. For example, one would have drawn differing conclusions from the two data sets for 4 of the above 70 consultants, as their confidence intervals are adjusted to include 1 with the patient data, but not with the aggregate data. (Figures 6.3(a) to (c)).

The log linear model approach was rather time consuming, and the relationship between the variables was dependant on the marginals rather than being controlled externally. The β value could have been specified, thus controlling the correlation, but this would have been rather difficult as β is not constrained and was usually estimated somewhere between -1 and 4. To combat these problems of data generation, a different approach was then adopted.

6.3.1.2 USING PROPORTIONS OF THEORETICAL MAXIMUM CORRELATION AND INDEPENDENCE CONTINGENCY TABLES.

This method was far simpler and faster to use than the log-linear approach, and advantageously did not rely on any external routines.

The idea is to generate from the marginal distributions a set of probabilities for each cell which would arise with maximum correlation (pcorr), and a set which would arise from complete independence (pind). The independence model is simply the product of the marginals, and the correlation probabilities are calculated from a contingency table, by putting the maximum possible values on the diagonal and filling in the rest on the off diagonals.

For example, the following tables show the independent and maximum correlation probabilities which would result from the marginal probabilities shown.

pind

0.025	0.03	0.045	0.1
0.15	0.18	0.27	0.6
0.075	0.09	0.135	0.3
0.25	0.3	0.45	

pcorr

0.1	0	0	0.1
0.15	0.3	0.15	0.6
0	0	0.3	0.3
0.25	0.3	0.45	

A coefficient, α, can then be specified, which gives the size of contribution of the independence model, and so the probability of any cell p(i,j) is calculated from

$$p(i,j) = \alpha pind(i,j) + (1-\alpha)pcorr(i,j) \qquad (0 \leq \alpha \leq 1).$$

Patients can then be allocated to cells using random variables as before. Initially, this type of simulation was performed using marginals generated from particular Normal distributions, as with the log-linear method, and the results of 4 simulations using 150 consultants each with their total number of patients from a distribution N(1200,400) can be seen in Table 13. The first two were generated using α=0.75, and the second two with α=0.25.

Patient Data

simulation	emergency	day case	under 11	over 60	constant
1	1.25	-1.40	-0.49	1.74	-4.49
2	1.25	-1.53	-0.55	1.79	-4.53
3	1.23	-1.36	-0.56	1.71	-4.45
4	1.21	-1.51	-0.55	1.77	-4.50

Aggregate Data

simulation	emergency	day case	under 11	over 60	constant
1	0.87	-0.69	0	2.36	-4.04
2	1.45	0	0	2.66	-4.42
3	0.55	0	0	2.70	-3.91
4	0.84	0	0	2.60	-3.99

Table 13: Coefficients of variables in LR equations for 4 data sets

Again, the patient data give fairly consistent estimates over the data sets, and they correspond well with the original equation whereas, even with these large numbers, the aggregate data lose the information, and the variables with positive coefficients are favoured over the low risk ones, as they are more common. These equations would not, then, make equivalent adjustments to the patient and aggregate data mortality rates. The effect of α on the adjustments must also be explored in more detail.

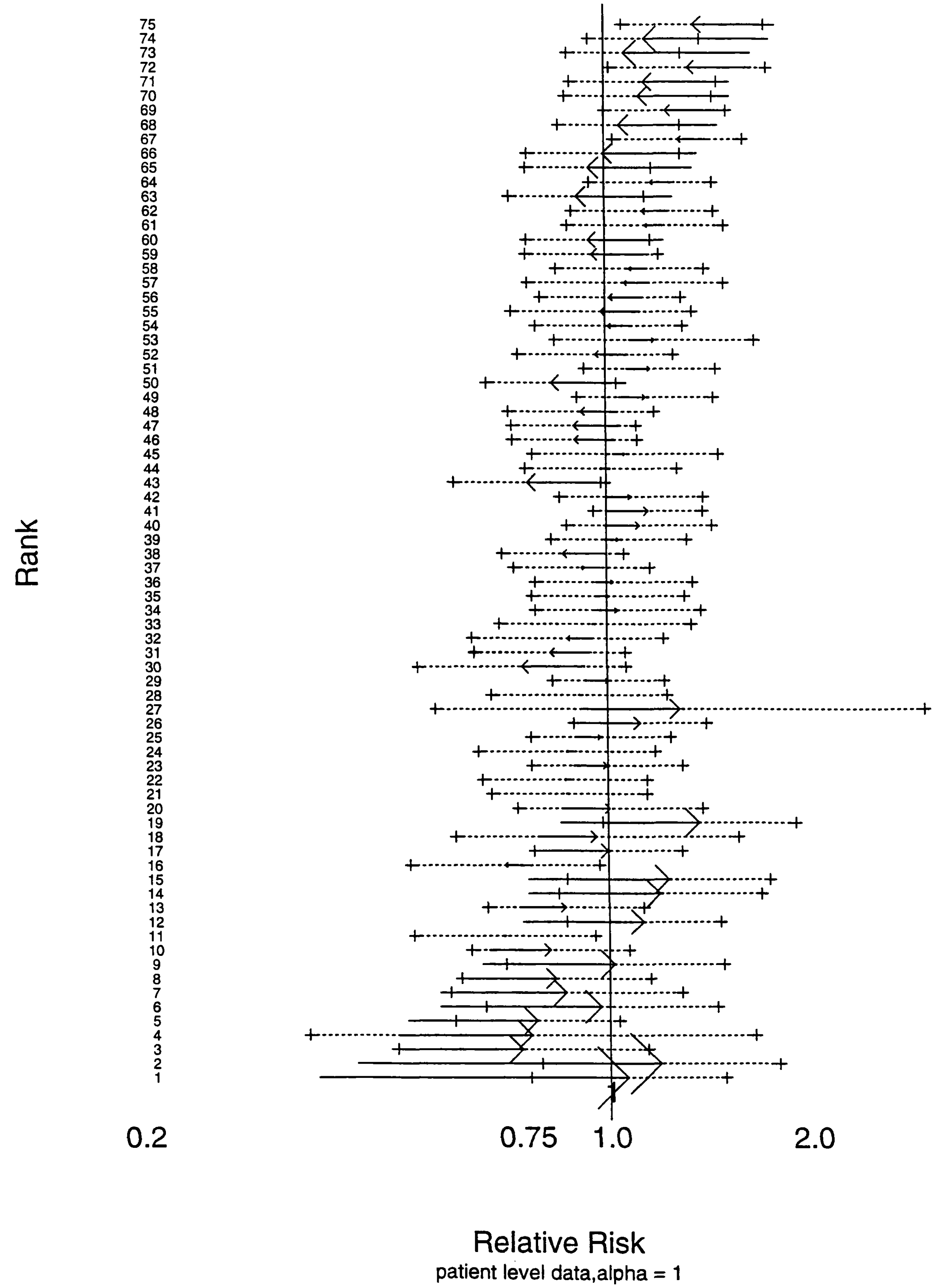


Figure 26: Example of plot showing change between unadjusted and adjusted relative risks for simulated consultants

6.3.2 Final Analyses

It was decided to base marginal totals on the actual data rather than generating them, and so for the definitive simulations random samples were taken from the 1992 data. Up until this point, the coefficients of each consultant variable had been extracted manually from the BMDP output of separate logistic regressions for each one. This was a rather tedious and time consuming process, but BMDP has no facility to extract the required coefficients and their standard errors. These could, however, be extracted from the summary of the output from the glm function with a binomial link and with the total number in the category as a weight in Splus. The coefficients were then written out to data files, and the required ones extracted using a FORTRAN program. This was a far more efficient method of acquiring the data required to plot the confidence intervals. The generation of the main data sets and analyses of them will now be described.

The number of consultants was kept at 75, and 5 sets of data were generated at each of 5 levels of α. For all of these, the changes in relative mortalities on adjusting were plotted in the form of arrows, along with the adjusted confidence intervals. An example of such a plot is given in Figure 26. The general pattern evident in these plots is that the intervals tend to move towards unity, but we would hope that the intervals behave the same for the patient and aggregate data. Since no inherent difference between consultants was introduced in data generation, approximately 95% of these relative risk intervals should contain 1. We could thus consider whether the interval for log R contains 0 as a Binomial variable with n=75 and p=0.05. Under this distribution, there is almost no probability of achieving a value greater than or equal to 10, and by chance we could expect about 3 or 4 of the adjusted intervals to lie away from 1 if the adjustments are working.

In order to summarise the results, and to see the effect of correlation between the variables, a z value was computed for each consultant, where

$$z = \frac{\log \hat{R}}{\text{s.e.}(\log \hat{R})}.$$

The number of confidence intervals for R not including 1 could easily be found by counting the number of the $|z|$ values greater than 1.96. The numbers of intervals in the five simulations at each level of α which do not contain 1 are summarised in Table 14. If there were no difference between the consultants, we would expect the z values to have a N(0,1) distribution. Since there has been no inherent difference between surgeons included in the simulation, one would expect this to be the case for the adjusted values.

	Patient data		Aggregate data	
α	unadjusted	adjusted	unadjusted	adjusted
0	26.2	6.4	26.2	7.6
0.25	26.8	4.6	26.2	6.4
0.5	22.6	6.0	22.2	6.8
0.75	23.8	4.0	23.8	2.8
1	24.6	4.6	24.6	3.8

Table 14: Average numbers of confidence intervals for R not containing 1 for simulated data sets

The first thing to observe from Table 14 is that the adjustments have removed a large proportion of the variability for patient and aggregate data. For the lower values of α (most correlation), more confidence intervals for R from patient data than aggregate data contain 1, whereas with the independent data, the aggregate data adjustments appear to be just as, if not more, effective. It is difficult to tell whether there is any trend, but numbers of intervals not containing 1 appears to decrease with decreasing dependence with the aggregate data, suggesting the model gained from these data may be more sensitive to correlation between the variables.

The sum of squared z values, as given in Table 15, would be expected to follow a χ^2 distribution with 75 degrees of freedom if the z values were distributed as N(0,1). The values of $\sum z^2$ should then be around 75, with an approximate 98% prediction interval in which $\chi^2(75)$ should lie being (70.1,106.4). It can be seen that all the adjusted $\sum z^2$ values are contained within this interval, whereas the unadjusted ones are much higher. The values calculated from the patient data are less variable, and closer to 75, suggesting that their z values follow more closely an N(0,1) distribution than those calculated from the aggregate data. However, the totals obtained from the aggregate data are not significantly different from this distribution, and as such it can be concluded that the adjustments are having the required effect overall.

	Patient data		Aggregate data	
α	unadjusted	adjusted	unadjusted	adjusted
0	386.97	77.61	384.35	101.30
0.25	426.30	78.89	423.35	93.34
0.5	306.54	82.25	304.73	99.66
0.75	313.04	76.27	314.80	70.43
1	352.29	81.32	347.25	77.94

Table 15: Mean values of $\sum z^2$ for simulated data

By exploring these standardised values, we have investigated the deviation from 0 of the log relative risks, and can see that the adjusted values are, as we would expect, not significantly different from 0 overall. However, an important effect of adjusting which has been overlooked here is the direction. The regression equations used to adjust the patient and aggregate data sets were quite different, but they could in general give the same results overall. If we look at the diagrams showing direction of movement from unadjusted to adjusted values, we can count how many move in opposite directions for the two types of data to give an idea of the similarity in adjustment effects. Table 16 gives the mean numbers of simulated consultants at each value of α where the differences between the unadjusted and adjusted risks had opposite signs for the patient and aggregate data. This does not take magnitude into account and some of the differences are very small.

α	mean
0	8.2
0.25	10.0
0.5	8.0
0.75	8.0
1	4.2

Table 16: Mean numbers of adjustments going in opposite directions for patient and aggregate data sets

The number of times out of 75 the adjustment goes in different directions appears from these simulations to be approximately halved when no correlation is present.

The above summaries have all suggested that aggregate data performs best when the factors are uncorrelated, but with only 5 data sets with each value of α the results are unclear. In order to try to identify a trend more clearly, the simulations have been repeated. The results from the next simulations are given in tables 17, 18 and 19. These are again for 5 data sets, each with 75 consultants at each of the 5 levels of α.

	Patient data		Aggregate data	
α	unadjusted	adjusted	unadjusted	adjusted
0	19.4	4.4	19.2	6.4
0.25	24.4	4.6	24.6	7.8
0.5	23.4	3.0	23.4	4.4
0.75	26.0	2.8	26.2	3.4
1	23.8	1.8	24.0	2.4

Table 17: Average numbers of CI's for R not containing 1 for second set of 5 simulated data sets

	Patient data		Aggregate data	
α	unadjusted	adjusted	unadjusted	adjusted
0	271.17	76.30	270.14	93.78
0.25	333.17	86.84	331.69	96.99
0.5	331.37	71.50	327.22	79.26
0.75	342.71	72.00	338.42	72.18
1	302.02	63.79	299.70	57.01

Table 18: Mean values of $\sum z^2$ for second set of 5 simulated data sets

α	mean
0	12.8
0.25	9.6
0.5	7.2
0.75	4.4
1	3.4

Table 19: Mean numbers of adjustments going in opposite directions for patient and aggregate data for second 5 simulated data sets

Similar results can be seen for the second set of 25 simulations as for the first, although these have shown that there is great variability present. The basic trend that is evident is that inter "surgeon" variation is reduced more by adjustment if there is less dependence between the explanatory variables. For $\alpha = 0$, there were less initial differences between the unadjusted values than for the rest of the values of α, so the adjusted values have a lower mean $\sum z^2$ than we might expect from the trend. With the average numbers of adjustments going in opposite directions for patient and aggregate data sets, a clear trend appears. When there is no correlation, the number is very small.

Looking at the averages over all 10 simulations (Tables 20 - 22), the pattern is more visible. The number of confidence intervals not containing 1 and the number moving in opposite directions decrease as the level of dependence between the variables decreases. Even with independence, there are still more than three adjustments on average over the 10 simulations which have opposite effects. Sometimes, however these adjustments are as small as to be negligible. It may be more instructive to look at the significant values only, and observe how many of them do not match for the patient and aggregate data. Table 23 shows the mean numbers of significant adjusted confidence intervals for three values of α. It can be seen that for $\alpha = 0$ (complete dependence) there were on average 4.5 consultants under the aggregate data whose confidence intervals excluded 1, compared with only 0.9 for $\alpha = 1$. In fact, 3 of the 10 simulations with $\alpha = 1$ had no consultants in this category, whereas for $\alpha = 0$ and $\alpha = 0.5$, all of the simulations had at least one consultant with the aggregate data confidence interval significant when the patient data one was not. This shows that when the variables are independent, the quality of information achieved from the aggregated data is almost as good as that from the patient data. The number of relative risk confidence intervals which have moved in opposite directions for the different data sets, and are significantly different from 1, is very small. On average only one tenth of a consultant per 75 in the simulated study has a confidence interval which is significant with one data set and not the other, and has moved in the opposite direction.

	Patient data		Aggregate data	
α	unadjusted	adjusted	unadjusted	adjusted
0	22.8	4.9	22.7	7.0
0.25	26.1	4.6	24.5	6.6
0.5	23.0	4.0	22.8	6.6
0.75	24.8	3.4	24.5	3.1
1	24.2	3.2	24.3	3.1

Table 20: Average numbers of CI's for R not containing 1 for all data sets

	Patient data		Aggregate data	
α	unadjusted	adjusted	unadjusted	adjusted
0	329.07	76.96	327.25	97.54
0.25	379.24	82.14	464.14	95.17
0.5	318.46	77.19	316.98	89.46
0.75	317.36	73.63	326.61	71.31
1	327.16	72.56	323.48	67.48

Table 21: Mean values of $\sum z^2$ for all simulated data

α	mean
0	10.5
0.25	10.0
0.5	7.6
0.75	6.2
1	3.8

Table 22: Mean numbers of adjustments going in opposite directions for patient and aggregate data sets

α	# significant with patient data only	# significant with aggregate data only	# significant with both data sets
0	2.4 *(0.3)*	4.5 *(0.7)*	2.5 *(0.3)*
0.5	1.5 *(0.1)*	3.1 *(0.3)*	2.5 *(0.3)*
1	1.0 *(0.1)*	0.9 *(0.1)*	2.2 *(0.2)*

Table 23: Mean numbers of significant confidence intervals for patient and aggregate data sets.(Number of adjustments going the opposite way)

6.4 Conclusions

The simulations have shown that adjustments of aggregate data are almost as efficient as those of patient data, provided the numbers of consultants and patients are large enough. This is not a problem where our particular data are concerned, as they are larger than those considered here. The results of the 10 simulations looking at different amounts of correlation between the explanatory factors show that independence gives more satisfactory adjustments, especially for aggregate data. If the value of α as used above could be measured in a data set, it would have to be greater than 0.75 to give reliable results. While the actual relationships between variables cannot be estimated using only aggregate data, it would seem that it is reasonable to assume only weak dependence. It is not obvious, for example, whether a young person is more likely to be admitted to hospital as an emergency than an old person.

While patient level data are obviously preferable for accuracy of achieving a model, and are usually to be recommended, the problems of working with a data set of several hundred thousand records should be remembered. It has been demonstrated by these simulations that surprisingly good results can be achieved from aggregate data using these methods, and so we can be fairly confident in our interpretation of the results, if we can find a good model for outcome. This is, of course more difficult with real data, as many different factors will affect outcome, including unknown ones. It seems, however, that although a more accurate model can be found using patient data, the ones from aggregate data do not seriously lead us to different conclusions about the surgeons.

7 INVESTIGATION OF POSSUM SCORES ON A LARGE DATABASE: THE DATA

7.1 Introduction

We have seen that, while useful results were gained with the RCS data, they are not entirely satisfactory, due to the collection of only aggregate data. If, in future, patient data were to be collected, they should be geared towards risk adjustment. Currently, POSSUM (Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity) is the most generally accepted and widely used predictive system for audit of general surgery, and its use has been suggested by the RCS. In chapter 2, we discussed some of the system's drawbacks. It is rather complex to calculate, and requires a large amount of data to be collected. It is not well calibrated for low risks, and so tends to overestimate mortality rates, thus potentially masking poor surgical performance. It also contains variables which could be interpreted as being a function of surgical skill, so does not meet some of the criteria for a model for use in audit. We were approached by consultants from Portsmouth at the 1994 meeting of the RCS Comparative Audit Service, where our presentation methods as described in the last chapter were presented. They were very interested in the ideas and wanted to further the research by considering POSSUM. They were collating a large amount of patient data on the variables required by POSSUM, which they have allowed us to use. We wish to use these data to test the POSSUM scoring system on a large set of general surgical data, which shall be described in this chapter.

We also wish to explore further improvements which could be made to the POSSUM system. These could be made by changing the logistic regression coefficients or the method of scoring. It may also be possible to find a simpler scoring system by discarding those variables which do not contribute substantially to the discriminatory power of the system. We investigate this using the POSSUM scores, and individual variables. We will also investigate the effect of missing data, and look at possible ways in which it can be treated. It is also of interest to assess whether possible surgeon-related variables are necessary. These questions will be addressed in chapter

8, and after selecting the "best" model, we shall use it to compare the consultants in this study in chapter 9.

7.1.1 The Portsmouth Data

The data from the Portsmouth NHS Hospitals Trust General Surgical Database consisted of 143 fields on 11231 patient episodes. Of these episodes, only 8123 involved a surgical procedure. For independence, it was decided to include each patient once only, and so the first episode for which there is an operation has been considered. This left 6714 patients. The data included admission, physiological, operative severity and outcome variables, as well as the calculated POSSUM scores and derived risks of morbidity and mortality. The POSSUM scores are calculated as stipulated in the original publication by Copeland *et al.* (1991). These are detailed in Appendix 6. We shall refer to the individual components of the Physiological and Operative severity scores as "weights". These weights can take the value 1, 2, 4 and 8. For the purpose of the initial analyses, a training data set consisting of a random sample of approximately half the patients has been used. This is so that any models developed can then be tested on the test data set of the remaining patients. The training data set consists of 3381 patients, with 101 postoperative deaths. The test set has 3333 patients and 85 deaths.

The data set has been used essentially as it was received, except for changes to very obvious data entry errors. For example, a value for pulse rate of 556 was considered to be a data entry error, and changed to 56. A pulse rate of 7 was changed to 70. For sodium concentration, two values of 12 and 14 were assumed to be 120 and 140. On several occasions, cells had obviously run into each other. For example, a value of sodium of 1384, followed by potassium concentration 0.2 were changed to 138 and 4.2. Other, less obvious errors have been left as they were, so could affect the results.

7.2 Preliminary Analyses

7.2.1 Outcome measure

The outcome measure of estimated 30 day mortality has been used. The data actually reported in-hospital mortality and there were not complete follow-up data. There is thus likely to be under reporting of deaths due to patients who were discharged and subsequently died within 30 days. Where patients had multiple episodes, it was possible to check whether they had died within 30 days of their first procedure, and to include this as a postoperative death. Patients who were known to have died after 30 days are considered to have survived for the purpose of these analyses. Complications data were considered too unreliable to be used, and there was no longer term follow-up information.

7.2.2 Variables

Firstly, we shall introduce the variables which are involved in the POSSUM score. The first stage was to study their relationship with 30 day mortality, in order to see how well the POSSUM weights appear to fit the data, and also to consider the form in which they might be included in a model based on the variables rather than scores.

7.2.2.1 Continuous Variables

These were divided into groups of similar numbers of patients, and the mortality rate plotted for each group. The patterns of these plots could then be studied. Missing data is a considerable problem, and will be discussed in more detail later.

Age

From the plot (Figure 27), it can be seen that the mortality rate generally increases with age, as one would expect. It would thus be reasonable to include age as a continuous variable in a logistic regression. The mean age of patients in the data set was 57, ranging from 14 to 100. Three patients had missing values for age. The POSSUM weights are shown on the plot, and in general are appropriate. A patient scores 4 if they are aged over 70, although from these data we would conclude that a more suitable cut-off point is age 75. The following table investigates the relationship of the POSSUM scores with mortality.

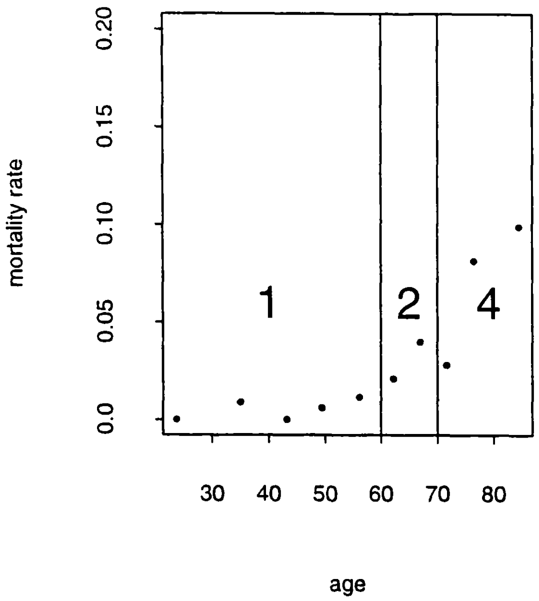


Figure 27

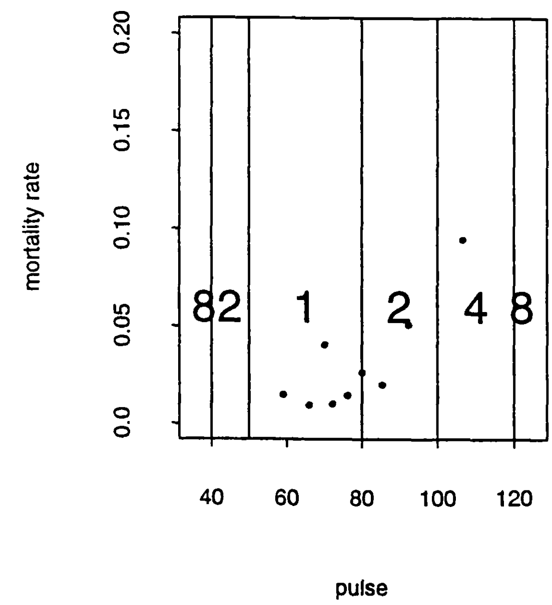


Figure 28

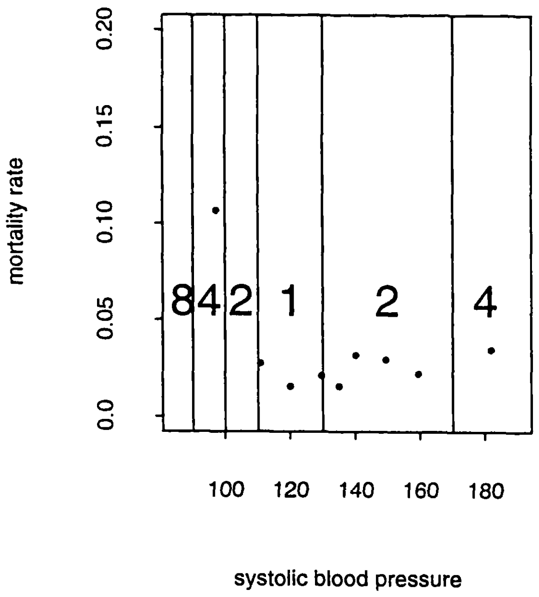


Figure 29

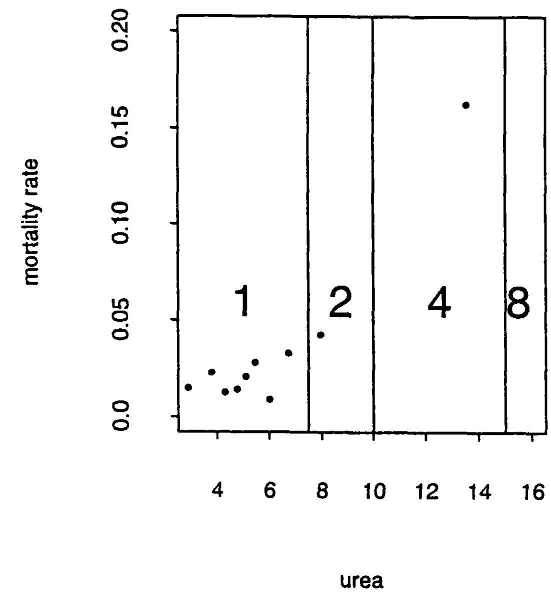


Figure 30

Figures 27-30: Mean mortality rates for grouped continuous variables. Sections show POSSUM Physiological Score. (POSSUM Study, training data set).

Age POSSUM Weight	Deaths(%) if cut-off 71	Deaths(%) if cut-off 75
1	11 (0.63)	11 (0.63)
2	19 (2.94)	32 (3.34)
4	71 (8.02)	58 (10.1)

Table 24: Death Rates by POSSUM and Experimental Weightings for age

The mortality rate for patients aged 71 to 74 was 4.18% (13/311). The decision as to where to categorise these patients is fairly arbitrary, but they lie closer to the lower age group than the higher one.

Pulse

Mortality rate also increases with an increase in pulse (Figure 28). There may also be an increase with very low values of pulse rate. It is difficult to tell, as this is fairly rare. The mean value was 79, with range 32 to 240 but with the bulk of patients having values between 60 and 100. As with age, we do not have any mean values occurring at the extreme values which achieve high risk POSSUM weights, so it is difficult to assess whether they are valid. If we look at mortality categorised by POSSUM weight, we get the following table.

Pulse POSSUM Weight	Deaths(%)
1	46 (1.97)
2	40 (4.50)
4	9 (8.33)
8	4 (40.0)

Table 25: Death Rate by POSSUM Weights for Pulse Rate

This appears to validate the scores, with the mortality rate increasing exponentially. While pulse could be included as a continuous variable, it would perhaps be more appropriate to use a categorical variable with cut-off 100 (and perhaps less than 50) or to transform the variable to take account of low values, using |pulse-70|. There were 30 missing values, which had a much higher than average mortality rate.

Systolic Blood Pressure

Figure 29 shows that by far the highest mortality rate occurs at very low blood pressures (below 100). Medical opinion, as suggested by the POSSUM scoring system would also attribute a higher risk to very high pressures. The plot suggests grouping the patients into above or below 100, but a better transformation may be |SBP-130|, as 130 appears to be in the centre of the groups whose mortality rates are all below average. The mean value was 135, ranging from 60 to 260. A table of POSSUM score and postoperative mortality is given below. From Table 26, it is apparent that not enough weight is given to the extreme values of blood pressure, and that there is not much to distinguish between the two lowest scores. This leads us to believe that perhaps two categories would be sufficient here.

SBP POSSUM Weight	Deaths(%)
1	33 (2.11)
2	43 (3.12)
4	15 (6.17)
8	6 (75.0)

Table 26: Death Rates by Systolic Blood Pressure POSSUM Weights

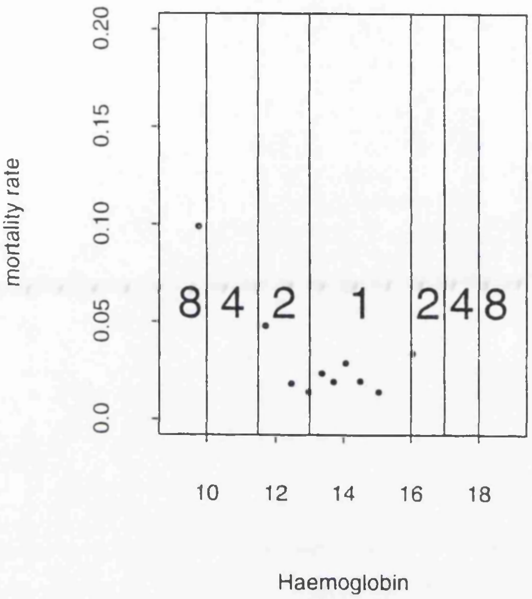


Figure 31

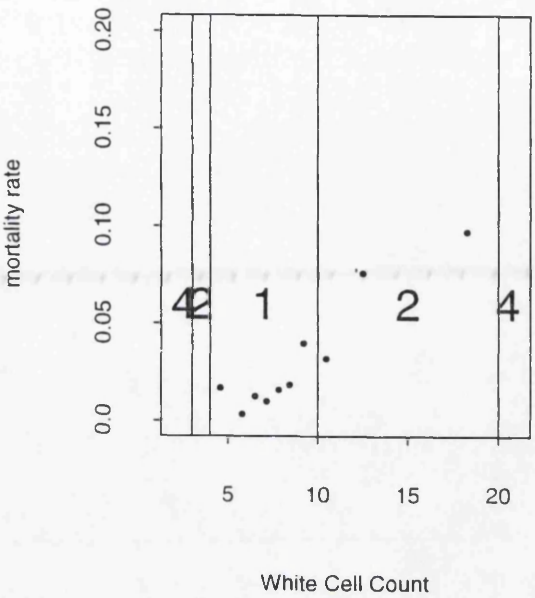


Figure 32

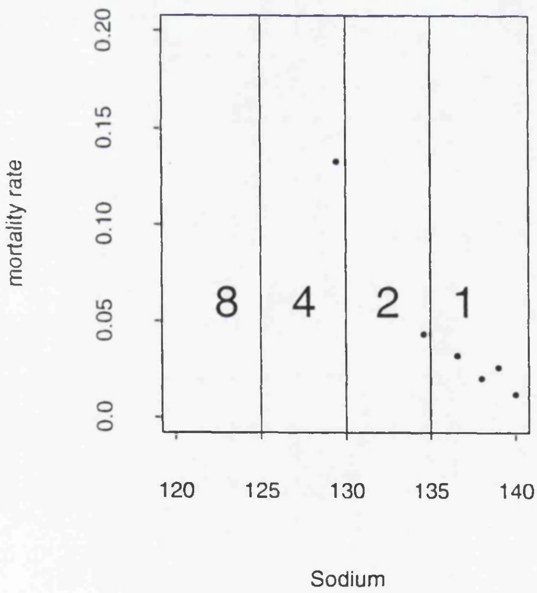


Figure 33

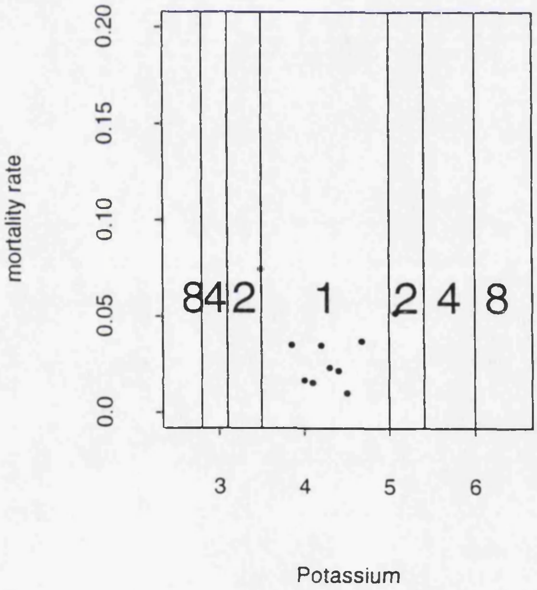


Figure 34

Figures 31-34: Mean mortality rates for grouped continuous variables. Sections show POSSUM Physiological Score. (POSSUM Study, training data set).

Urea

Figure 30 shows an increase in mortality rate with increasing urea concentration. The values fluctuate around a low value for urea less than about 7 mmol/l, and increase dramatically above a concentration of 10 mmol/l. This suggests using a categorical variable for above/ below 10. There were 647 missing values, which corresponded to a very low mortality rate. The mean value was 6, and the range was 1.4 to 41.8. The relationship between urea and mortality can be easily seen in the following table.

Urea POSSUM Weight	Deaths (%)
1	45 (1.99)
2	16 (5.82)
4	21 (16.41)
8	14 (21.87)

Table 27: Death Rates by Urea POSSUM Weight

Haemoglobin

The following table shows numbers of deaths for each POSSUM Haemoglobin weight value. There is no distinction with these data between 1 and 2, as could have been expected from Figure 31.

Hb POSSUM Weight	Deaths (%)
1	43 (2.26)
2	17(2.25)
4	17 (6.67)
8	20 (14.1)

Table 28: Death Rates by Haemoglobin POSSUM Weight

It can be seen from the plot that very low concentrations of Haemoglobin correspond to a marked increase in mortality rate. There is likely to be an increase in mortality rate for high values, so a transformation of the type used above would be appropriate. It is not obvious from the plot which value should be subtracted in the transformation. The median value of 13.5 could be used. However, since low values appear to have much higher mortality, a larger value, say 15 would perhaps be more appropriate. Alternatively, a value of 14.5 corresponds to the midpoint of the group in the original

POSSUM score with a weight of 1. In fact, 13.5 appears to give a slightly more significant coefficient in logistic regression than either of the higher values tried. Missing values, of which there were 334, corresponded to a low mortality rate. Given the information in Table 28, a categorical variable where "high risk" is less than twelve or greater than 16, and "low risk" from 12 to 16 may also be suitable.

White Cell Count

Table 29 shows a strong relationship between the POSSUM scores for White Cell Count and mortality. With a maximum weight of 4, perhaps not enough emphasis is given to this variable, as the mortality rate for this group is fairly high. Looking at Figures 30 and 31, we can see that a group scoring 2 here has a similar mortality rate to the group scoring 8 in the Haemoglobin plot (Figure 32).

WCC POSSUM Weight	Deaths (%)
1	35 (1.62)
2	48 (5.98)
4	14 (18.2)

Table 29: Death Rates for WCC POSSUM Weights

The plot of WCC against mortality rate suggests a logistic relationship, and that WCC would be included in a model as a continuous variable. Perhaps there is a slight increase in mortality for very low values below 4 which is difficult to see from this plot due to the small number of patients with very low counts. POSSUM gives a weight of 2 for values from 3.1 to 4, and 4 for counts of less than 3. To take this into account, we could use a transformation such as |WCC-7|. The data had a mean count of 7.05 and ranged from 0.4 to 47.8. There were 338 patients with no data for this variable, of whom 4 died.

Sodium

Mortality rate shows a steady reduction with increased sodium concentration (Figure 33). There is a marked increase for very low values, and a categorical variable of less than or greater than 132 appears to describe the situation. There are 137 patients with a value below this threshold. The average value was 138, with a minimum of 93 and a maximum of 192. Looking at the following table, we can see that the relationship between the sodium weight and mortality is a strong one.

Na POSSUM Weight	Deaths (%)
1	53 (2.35)
2	29 (7.67)
4	10 (11.6)
8	4 (23.5)

Table 30: Death Rates by POSSUM Weights for Sodium Concentration

Potassium

The plot of mortality rate for grouped potassium data (figure 34) shows a marked increase in mortality for high and low values, with fluctuation about the mean in the middle of the range. The transformation $|K-4.2|$ has also been plotted (Figure 35), and shows an increase in mortality with increased values of this transformed variable above 0.6. There is fluctuation for smaller values, and a categorical variable which takes the value 1 if the potassium concentration is between 3.5 and 5, and 2 otherwise seems appropriate. There were 661 missing values, corresponding to a very low mortality rate. The mean value of potassium in this group of patients was 4.2, ranging from 2.4 to 7.7.

K POSSUM Weight	Deaths (%)
1	72 (2.89)
2	13 (8.39)
4	8 (13.3)
8	2 (13.3)

Table 31: Death Rates for Potassium POSSUM Weight

The above table of POSSUM weight by mortality shows that any abnormal value (outwith 3.5 - 5) constitutes a higher risk, providing evidence for the above categorical variable and backing up the tight "V" shape of the plot. This suggests that the POSSUM scoring for this variable could be less detailed.

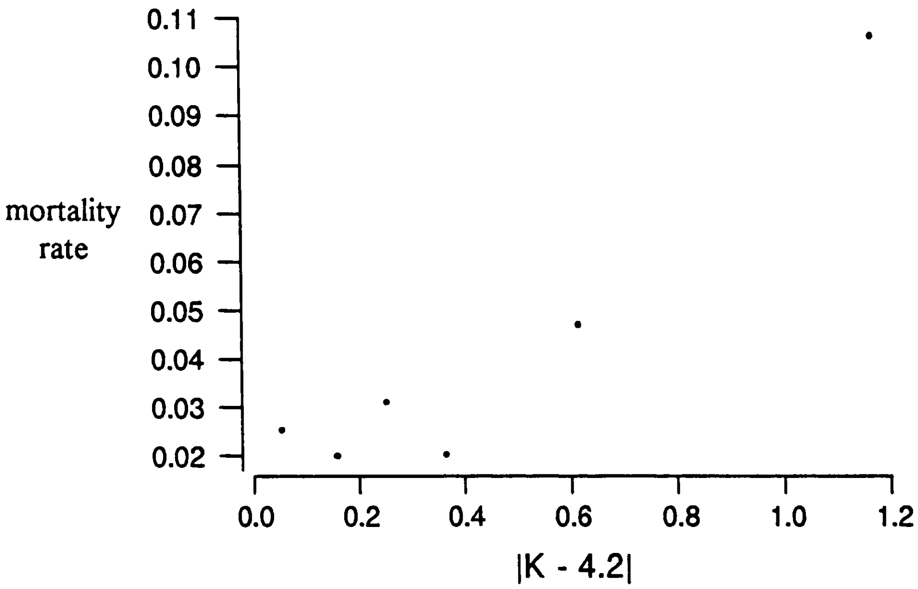


Figure 35: Transformed Potassium Data versus Mean Mortality Rates (POSSUM Study, Training Data Set)

Glasgow Coma Score

The data on GCS were not particularly helpful, as only 18 of the patients had scores of less than the maximum possible score of 15. Of the 5 patients with the lowest scores none died. It is likely that most patients with a low GCS will be in Intensive Care, and so this variable is not of use for modelling in a general surgical setting.

7.2.2.2 CATEGORICAL VARIABLES

Dyspnoea

The data for this variable are summarised in Table 32. The POSSUM weights contribute to the "respiratory history" section of the Physiological score. Since patients with dyspnoea on exertion and those with limiting dyspnoea do not differ in mortality rate, perhaps 1, 2 and 4 would be more appropriate weights, with the middle two groups both scoring 2. Combining the three groups having dyspnoea gives a mortality rate of 6.2% for 655 patients.

	no dyspnoea	dyspnoea on exertion	limiting dyspnoea (one flight)	dyspnoea while at rest	missing
POSSUM score	1	2	4	8	1
Number of patients	2604	471	146	38	108
Mortality rate	1.9%	5.7%	4.1%	13.2%	11.5%

Table 32: Mortality rates for levels of Dyspnoea

Cardiac Drugs

In calculating the POSSUM weight for "Cardiac signs", any therapy for cardiac problems is taken into account. Their contributions to this score are shown in Table 33.

Drug	POSSUM weight	Number of patients	Mortality Rate
Diuretic	2	248	8.06%
Digoxin	2	91	10.99%
Antianginal	2	128	5.47%
Hypertensive	2	320	3.44%
Warfarin	4	28	7.14%

Table 33: Mortality Rates for Cardiac Drugs

There were 140 patients receiving two drugs, the most common combination being diuretic and hypertensive therapies, 33 patients received three drugs and 1 four. It was relatively uncommon for digoxin to be administered with any other treatment.

If we carry out χ^2 tests of association on these data, two of the drugs are significantly related to mortality. The drug which is most highly related to postoperative mortality is digoxin. Diuretic is also significantly related. Based on these data, one might increase the weights for diuretic and digoxin to 4.

Oedema

The mortality rate for those without oedema was 2.9%, compared with 4.9% for those with. This difference is only just significant at the 5% level (χ^2=4.017). Strangely, 9 of the 67 with missing data died within 30 days, giving a rate of 11.8%. This also contributes to the POSSUM "Cardiac signs" score, providing a weight of 4 if present. Based on these mortality rates, this is rather a high weight.

Jugular Vein Pressure

Those patients experiencing JVP had a mortality rate of 7.7% (7/72) compared with 2.6% for those who did not. Again, the missing data have a high mortality rate associated with them of 12.2% (10/82). This variable is also part of the "Cardiac signs" score, gaining a weight of 8 if present. Again, this appears rather high given these data.

ECG

The data on electrocardiograms were fairly sparse, as over two-thirds of patients either did not have these carried out, or the data were not recorded. The patients who had an ECG done had a higher mortality rate than those who did not (4% versus 2.3%, p<0.005). The data are summarised in Table 34.

ECG Reading	POSSUM Score	Number of patients	Mortality Rate
Normal	1	625	2.4%
AF 60-90	4	31	7.7%
AF >90	8	16	18.8%
≥5 ectopics/min	8	18	5.8%
Q or ST/T wave	8	164	3.7%
Other abnormal	8	176	5.7%

Table 34: ECG Data, POSSUM scores and mortality rates

These data suggest that the weights given to abnormal ECG readings are too high, and that they should be assigned 4 instead of 8, except perhaps for an atrial fibrillation rate of greater than 90. Since all these categories contain few patients, when considering variables, we shall construct a binary summary variable for ECG with normal versus abnormal.

Chest X-ray

For chest X-ray, again over two-thirds had no information. Those who had an X-ray had a higher mortality rate than those who did not (4.7% versus 1.7%). Heart X-rays contribute to the cardiac score and lung X-rays to the respiratory history score. The information is summarised in tables 35a and 35b.

	POSSUM Score	Number of patients	Mortality Rate
Normal	1	831	3.5%
Borderline Cardiomegaly	4	93	5.4%
Cardiomegaly	8	64	10.9%
Missing/ not done	1	2393	2.5%

Table 35(a): Heart X-rays, POSSUM weights and mortality rates

	POSSUM Score	Number of patients	Mortality Rate
Normal	1	811	2.8%
Mild COAD	2	55	3.6%
Moderate COAD	4	25	16.0%
Fibrosis/ Consolidation	8	57	10.5%
5	1	29	13.8%
Missing/ not done	1	2393	2.6%

Table 35(b): Lung X-rays, POSSUM weights and mortality rates

For the heart X-ray data, again weights of 1, 2 and 4 may be more appropriate than 1, 4 and 8 as the mortality rates are not particularly high. The “5” in the lung X-ray data has not been explained, but has been given a weight of 1 for the purpose of the POSSUM analyses. When considering lung x-ray as a categorical variable, normal versus abnormal, “5” has been included as abnormal due to the high mortality rate.

Malignancy

This makes up part of the Operative Severity rather than Physiological POSSUM score. The data are summarised in the following table.

	POSSUM Score	Number of Patients	Mortality Rate
None	1	2642	2.6%
Primary only	2	335	2.7%
Nodal Metastases	4	173	2.3%
Distant Metastases	8	69	10.1%
Missing	1	162	7.4%

Table 36: Malignancy, POSSUM Weighting and Mortality Rates

From these data, it appears that only the presence of distant metastases carries an increased risk of postoperative mortality. Thus perhaps primary tumour and nodal metastases should be weighted the same as no malignancy. For consideration of variables, it seems appropriate to split this into two groups: distant metastases versus no distant metastases.

Mode of Surgery

The mode of surgery is whether the operation was elective, urgent (within 24 hours of admission) or emergency (within 2 hours). Mortality rate increased through these categories, with 1.2% of elective patients, 5.4% of urgent patients and 37.7% of emergency patients dying within 30 days. The POSSUM weights for these three groups are 1, 4 and 8. There were 500 urgent patients and 85 emergencies, with 135 patients having their mode of surgery unrecorded. The mortality for this missing group was 7%. There is not a significant difference between the elective and urgent operations when other factors are accounted for, so these will be grouped together in our analysis of individual variables.

Multiple Procedures

The patients were categorised as having one, two or more than two procedures. The corresponding mortality rates in these categories were 2.7%, 3% and 4.6% respectively. The POSSUM weights for these categories are 1, 4 and 8, so they do not fit these data very well. A combined category of two or more procedures has a mortality rate of 3.2%. This variable does not discriminate well for survival. The 125 patients with missing data had a mortality rate of 7.4%.

Operative Severity

The POSSUM system has four categories for operative severity: minor, moderate, major and major+. Thus we get the following table of mortality rates by POSSUM categories. With these data, this does not appear to be the most useful method of grouping.

POSSUM score	Deaths (%)
1	17 (2.23)
2	6 (0.45)
4	20 (3.05)
8	52 (7.81)

Table 37: Death rates for POSSUM groupings of operative severity

The data for operative severity are summarised in the following table.

Category	Patients	Deaths	Mortality Rate
minor	761	17	2.23%
intermediate	1320	6	0.45%
major	655	20	3.05%
major +	367	30	8.17%
comp major D	50	1	2.00%
comp major C	65	7	10.77%
comp major B	46	14	30.43%
comp major A	2	0	0.00%
(comp major total)	163	22	13.50%
missing	115	6	5.22%
Total	3381	101	2.99%

Table 38: Mortality rates by Operative Severity category

Perhaps a more effective way of assigning the weights, based on these data, would be to give 1 for minor or intermediate procedures, 2 for major, 4 for major+ and 8 for complex major. If considering a categorical variable, one would be inclined to have 2 categories with major included in the lower risk category, and major+ and complex major together. Alternatively, a 3 category variable may be suitable with major+ and complex major as two separate categories.

Blood Loss

As discussed in chapter 2, this is a dubious variable to include in a model for audit as it could depend on surgical skill as much as the severity of the procedure. However, it is very highly related to postoperative mortality, which increases as volume of blood lost increases, as shown in the following table.

Volume	Patients	Deaths	Mortality rate
<100 ml	2565	43	1.68%
101-500 ml	487	18	3.70%
501-999 ml	93	8	8.60%
>1000 ml	93	20	21.51%
missing	143	12	8.39%
Total	3381	101	2.99%

Table 39: Death Rates for Blood Loss Groupings

POSSUM gives weights of 1, 2, 4 and 8 respectively to the above four volume categories. For inclusion in a model it could be kept as 4 categories, or split into 2 (less or greater than 0.5 litres). This variable is very highly correlated with operative severity, as can be seen in Table 40. Calculating the correlation between the values gives 0.63, and a χ^2 test of association gives a test statistic of 4608 on 32 degrees of freedom, which is very highly significant.

	Blood loss (ml)					
Operative Severity	missing	<100	101-500	501-999	>1000	Total
missing	115	0	0	0	0	115
minor	8	748	4	0	1	761
intermediate	3	1214	95	6	2	1320
major	9	423	195	19	9	655
major +	5	156	148	36	22	367
comp major D	0	9	12	9	20	50
comp major C	2	10	23	18	12	65
comp major B	1	3	10	5	27	46
comp major A	0	2	0	0	0	2
Total	143	2565	487	93	93	3381

Table 40: Relationship between Operative Severity and Blood Loss

Peritoneal Soiling

This is another variable which can depend on surgical skill and which we would thus rather not include. There was an increase in mortality rate with increased severity of peritoneal soiling, from 1.8% with none to 6.2% for minor, 9.2% for local pus and 13.7% for free bowel contents, bile or pus. The POSSUM weights for the four categories are 1, 2, 4 and 8, which do not fit these data too well. Combining the last 3 categories gives a mortality rate of 8.1% for any peritoneal soiling. Of the 143 with missing data here, 12 (8.4%) died.

The preceding variables are all required to make up the POSSUM Physiological and Operative Severity Scores. The following two variables are not included in the score, but are available in the data set, and are very useful.

Emergency

This refers to the type of admission, and patients were either emergency admissions or not. Note that this is not the same as mode of surgery, and is not strongly related to it. It is strongly related to postoperative mortality, with a rate of 8% for emergencies and 1% for others. Approximately 28% of the admissions were emergency. There were no patients with missing data for this variable.

Myocardial Infarction

This was in three categories: those who had never had a myocardial infarction, those who had had one over 6 months ago, and those who had had one in the preceding 6 months. Mortality rate increased over these 3 groups from 2.4% to 8.9% to 17.9%. The 103 patients with this information missing had a mortality rate of 9.7%. Since there were only 28 patients with an infarction in the last 6 months, this group could be combined with those having one before.

Having described the data set and the individual components of POSSUM, we go on in the next chapter to use them for modelling. We will use the scores, individual weights and the actual variables in models, as we attempt to improve on the present POSSUM system.

8 MODELLING THE POSSUM DATA

8.1 POSSUM Scores

The Physiological and Operative Severity scores which make up POSSUM are calculated by adding the weights assigned to the twelve physiological and six operative severity variables separately. We would expect an increase in both of these scores to be associated with an increase in mortality. Table 41 shows that this generally is the case. The scores have been grouped together for easy presentation, and also because many of the combinations of scores do not contain any patients. The 17 patients with no data for any of the operative severity score categories are assumed to have a value of 1 for each, and so are included as 6.

			Operative Severity Score						
			6	7,8	9	10,11	12-16	17-34	total
	12	n	215	47	80	31	32	7	412
		m.r.	0	0	0	0	0	0	0
	13	n	241	90	97	57	42	14	541
		m.r.	0	0	0	0	0.024	0.071	0.004
Physiological Score	14	n	161	47	54	52	47	13	374
		m.r.	0.006	0	0	0	0	0.077	.005
	15,16	n	241	62	90	62	67	40	562
		m.r.	0.004	0	0	0	0.015	0.125	0.013
	17-19	n	199	67	51	71	65	35	488
		m.r.	0.005	0	0	0.014	0.031	0.057	0.012
	20-22	n	152	38	52	27	53	41	363
		m.r.	0.026	0.026	0	0	0.075	0.171	0.044
	23-28	n	149	48	38	61	73	55	424
		m.r.	0.027	0.021	0.026	0	0.082	0.236	0.059
	29-62	n	84	25	19	24	35	30	217
		m.r.	0.119	0.04	0.105	0.125	0.343	0.5	0.198
	Total	n	1442	424	481	385	414	235	3381
		m.r.	0.015	0.007	0.006	0.010	0.063	0.187	0.029

Table 41: Mortality Rates by POSSUM Scores

The two scores were included in a logistic regression model. This gave the model

$$\mathit{ln}\{R/(1-R)\} = 0.16\times P.S + 0.17\times O.S.S - 8.98.$$

The published report (Copeland *et al.*, 1991) gives the model

$$\ln\{R/(1-R)\} = 0.13 \times P.S + 0.16 \times O.S.S - 7.04.$$

The minimum possible predicted mortality obtainable from the published model is 1.08%, which as has previously been pointed out is very high for many general surgical patients. The minimum predicted probability of death from the above equation calculated from the Portsmouth data is 0.24%. Thus we would hope it would be better calibrated. We can compare the models using Receiver Operator Characteristic (ROC) curves, which plot the true positive rate (sensitivity) against the false positive rate (1–specificity). Figures 36(a) and 36(b) show the ROC curves when the above models are used to predict outcome on the training and test data sets respectively. The predicted values from the test set for the two equations have been plotted against each other in Figure 37. It can be seen that the published equation consistently estimates higher probabilities than the one calculated here. This means that more patients are predicted to die at every cut-off point, and the false positive rates are higher. Thus the ROC curves for the published model have points further to the right than the ones for our calculated model. ROC curves are not the ideal method of judging models when there are very low mortality rates, as we are dealing with here, but they give an idea of the relative performance of the models.

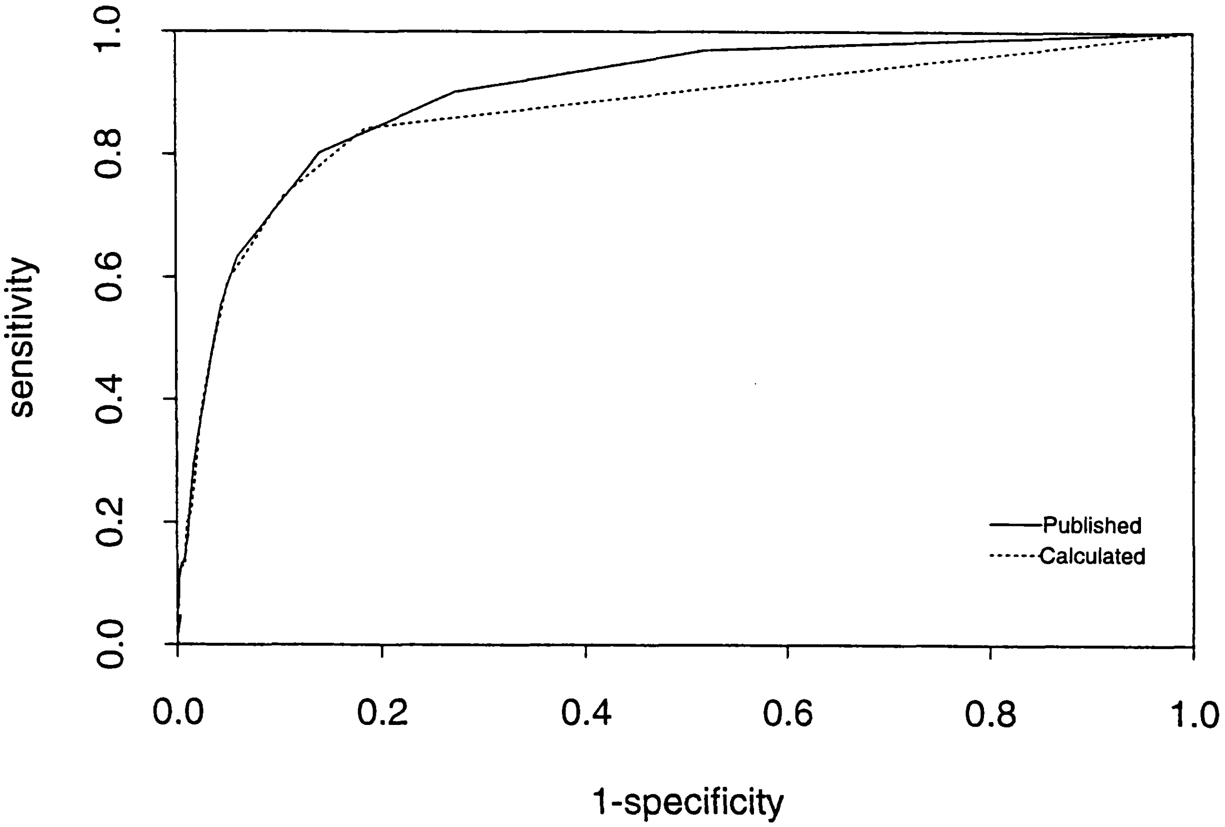


Figure 36(a): ROC curves comparing Published and Recalculated POSSUM Models (Training Data)

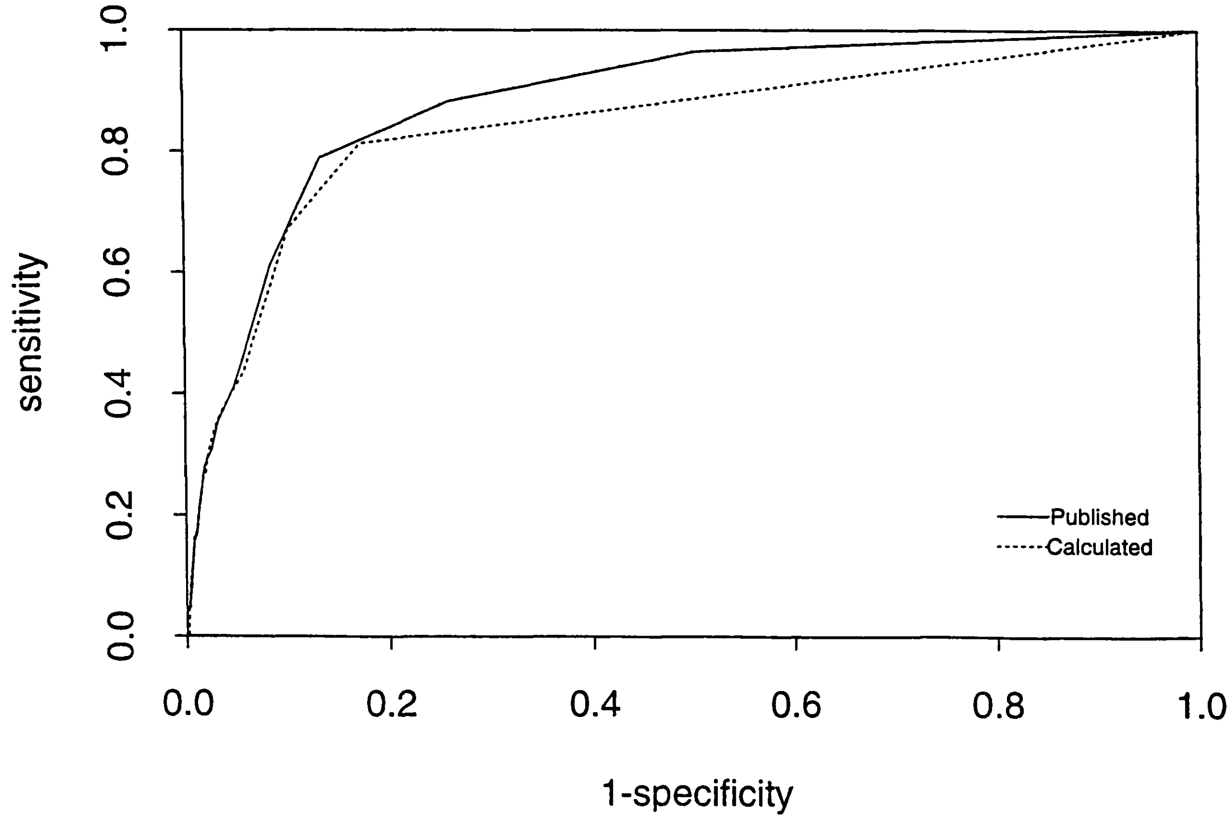


Figure 36(b): ROC curves comparing Published and Recalculated POSSUM Models (Test Data)

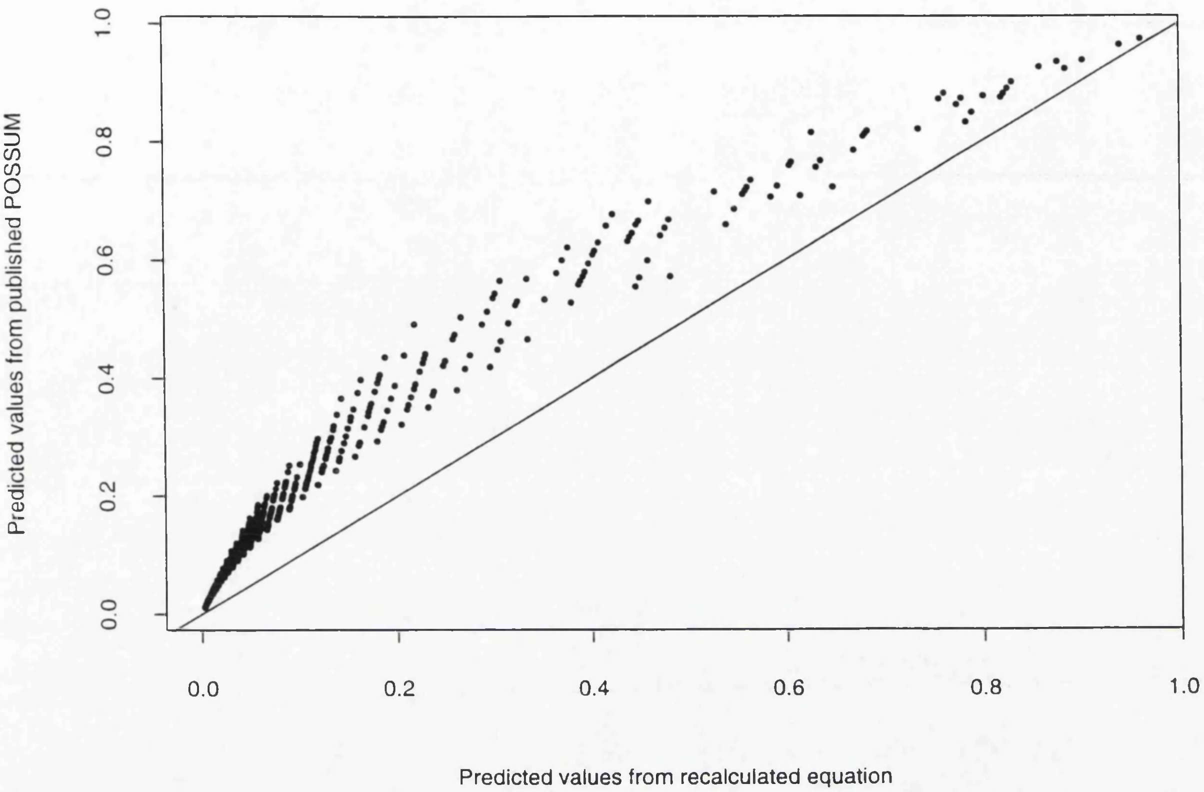


Figure 37: Comparison of Predicted Probablilities of Mortality from Two POSSUM Models (Test Data Set)

A measure of the efficiency of a model is the average quadratic, or Brier, score. For Logistic Regression, the quadratic score for each patient is calculated from

$$Q = 2(\text{observed} - \text{predicted})^2,$$

where "observed" is either 0 or 1 depending on whether the patient died, and "predicted" is the predicted probability of death from the regression equation. (Titterington *et al.*, 1981). This assesses both the discrimination and the calibration of the model. The averages of these have been calculated for the models discussed here, and are presented in Table 42. Note that a "baseline" value of this score, calculated by assuming that every patient has a probability of death equal to the mean mortality rate, is 0.0580 for the training data set., and 0.0497 for the test set. Thus it is desirable to have values smaller than these. Very small differences can mean a great improvement in the model.

Model	Data set	Q
Published POSSUM	training	0.0547
	test	0.0591
Calculated POSSUM	training	0.0488
	test	0.0493

Table 42: Average Quadratic Scores for POSSUM equations

We can see, then, that the published POSSUM model is performing worse than the null model when we assess it in this way. The model we have calculated is only a very slight improvement. However, if the model were performing as poorly as the null model, the ROC curve would simply be the line $y = x$. Since this is far from the case, we would not get the entire picture by simply considering the quadratic score. It does however give a good summary measure to use for comparisons.

8.1.1 Experimental POSSUM Scores

Keeping in mind the criticisms of the individual POSSUM scores in section 7.2.2, some changes have been made to fit in with the evidence from the data. The experimental adjustments of the scores from the initial POSSUM values can be seen in Table 43. The original scores are as given in Appendix 6.

If new physiological and operative severity scores are calculated from the scores changed as in Table 43, we obtain the model

$$\ln \{R/(1-R)\} = 0.18\times P.S + 0.17\times O.S.S - 8.11.$$

The minimum possible predicted probability for this model is 0.27%. The values of Q obtained are 0.0468 for the training data set and 0.0478 for the test data, so we have made a substantial improvement on the original POSSUM scores. It can be seen that when the model with experimental scores is tested on the test data set its ROC curve lies slightly outside the one for the original scores (Figure 38).

Physiological Score

Age	Score 4 for ≥ 75 instead of ≥ 71
Cardiac Signs	Score 4 instead of 8
	Score 4 for diuretic or digoxin therapy
	Score 2 instead of 4 for oedema
Respiratory History	Score 2 instead of 4 for limiting dyspnoea
	Score 4 instead of 8
Blood Pressure	Score 8 for <100 or >170, else 1
Haemoglobin	Score 4 instead of 8
	Score 1 instead of 2
White Cell Count	Score 8 instead of 4
Potassium	Score 4 if outwith 3.5-5
Electrocardiogram	Score 4 instead of 8

Operative Severity Score

Operative Severity	Minor/ Intermediate scores 1
	Major scores 2
	Major + scores 4
	Complex major scores 8
Multiple Procedures	More than 1 scores 2
Peritoneal Soiling	Minor scores 4
Malignancy	Primary or nodal metastases scores 1

Table 43: Changes to original components of POSSUM Scores for "experimental" analysis

We can gain an idea of how well these models are performing by looking at how many patients actually die within ranges of predicted mortality. These are summarised for the three models considered so far in Table 44 for the training data and Table 45 for the test data.

	Published POSSUM			Calculated POSSUM			Experimental POSSUM		
Predicted probability of death	Number of patients	Number of deaths	Mortality Rate	Number of patients	Number of deaths	Mortality Rate	Number of patients	Number of deaths	Mortality Rate
<0.25%	0	0	-	215	0	0.00%	0	0	-
0.25%-0.5%	0	0	-	1052	2	0.19%	1341	2	0.15%
0.5%-1%	0	0	-	756	4	0.53%	744	1	0.13%
1%-2%	1247	2	0.16%	555	6	1.08%	516	9	1.74%
2%-3%	639	3	0.47%	178	6	3.37%	186	4	2.15%
3%-5%	507	5	0.99%	206	9	4.37%	227	12	5.29%
5%-10%	446	10	2.24%	195	14	7.18%	155	13	8.39%
10%-20%	284	17	5.99%	99	22	22.22%	105	21	20.00%
20%-50%	194	41	21.13%	96	26	27.08%	74	20	27.03%
≥50%	64	23	35.94%	19	12	63.16%	33	19	57.58%

Table 44: Comparisons of Predicted and Actual Mortality Rates for 3 POSSUM Models: Training Data

	Published POSSUM			Calculated POSSUM			Experimental POSSUM		
Predicted probability of death	Number of patients	Number of deaths	Mortality Rate	Number of patients	Number of deaths	Mortality Rate	Number of patients	Number of deaths	Mortality Rate
<0.25%	0	0	-	226	0	0.00%	0	0	-
0.25%-0.5%	0	0	-	1059	3	0.28%	1352	3	0.22%
0.5%-1%	0	0	-	762	2	0.26%	744	0	0.00%
1%-2%	1263	3	0.24%	548	6	1.09%	528	8	1.52%
2%-3%	648	1	0.15%	173	7	4.05%	166	6	3.61%
3%-5%	502	6	1.20%	180	10	5.56%	202	12	5.94%
5%-10%	419	8	1.91%	164	20	12.20%	136	13	8.56%
10%-20%	263	26	8.89%	102	8	7.84%	95	15	15.79%
20%-50%	153	17	11.11%	80	15	18.75%	66	13	18.70%
≥50%	85	24	28.24%	39	14	35.90%	44	15	34.09%

Table 45: Comparisons of Predicted and Actual Mortality Rates for 3 POSSUM Models: Test Data

From Table 44, we can see that the published model is performing worst, with only one category of predicted probability of death having the observed mortality rate contained within it. It looks as though the model recalculated from the original POSSUM scores performs slightly better than the one from the experimental scores because it has most patients predicted as very low risk, where they should be. Looking at the test data in table 45, there does not appear to be much to choose between the original and new scores.

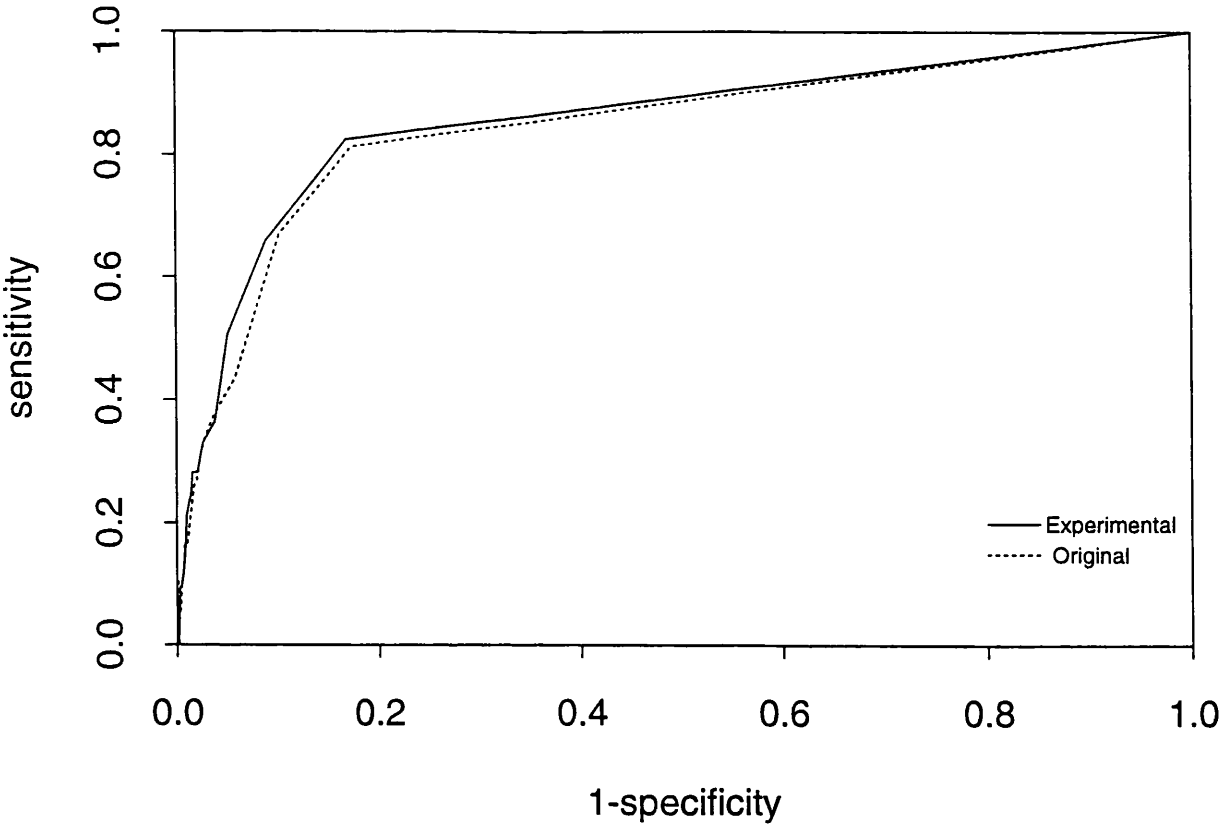


Figure 38: ROC Curve for Model containing experimental POSSUM Scores (Test Data)

8.2 Individual POSSUM components

Two of the problems with POSSUM are that there are many variables to consider and that some of them are associated with surgical skill. Although there are 18 components of the score, these rely on several more clinical variables as some scores e.g. 'cardiac signs' require several pieces of information. Using a stepwise selection procedure on the individual variable weights, it can be shown that not all of these are necessary for prediction.

Logistic regression was used in Splus. The criteria used for assessing a model are as follows:

- Initially, the stepwise selection procedure minimises the value AIC = D+2p, where D is the deviance and p the number of parameters in the model. This tries to achieve a balance between a good fit and parsimony. We assume here that there is no over-dispersion.
- We aim for a small value of the residual Deviance
- The coefficients should be significant, as should their Deviance contributions. However, with such a large data set, conventional significance levels will probably bring too many variables into the model, and one should pay more attention to measures of performance.
- The proportion of the total sum of squares which can be accounted for by the adjustment. This (from Smith 1994) is SSTa = SSTy-SSTd, where SSTy = $\Sigma\left(y_{ij}-\bar{y}\right)^2$ and SSTd is the sum of the squared differences between the observed values of y (0 or 1) and those predicted under the model. In order to satisfy the assumption that $\bar{p}=\bar{y}$, the predicted probabilities p_{ij} are multiplied by $\bar{p}/\bar{y}$ for the purpose of these calculations.
- Receiver Operator Characteristic Curve (ROC curve) for the fitted values from the model.
- Quadratic Score (Q)
- Testing the model on the test data set.

We use the same assumption as the original POSSUM here, that missing values are "normal" and are coded as 1. The first term to be dropped in the stepwise regression is the weight for "multiple procedures". This is not surprising given the very weak relationship in the data between this and mortality. The second term to be dropped is blood loss, which is rather fortunate since this is a variable one would prefer not to include. After this we lose Glasgow Coma Score, which did not look good for discrimination as 98.5% of the patients had the maximum score of 15. This is followed by the weight for the other variable about which we had misgivings, peritoneal soiling. Thus, after potassium and ECG are dropped from the model, we are left with 12 of the original 18 POSSUM variables. This model can be seen in Table 46. None of the cardiac drug variables are significant when included with the other information.

Note that the weights have been fitted as continuous rather than categorical variables. If we were to fit them as categorical variables, we would lose the information given in the coding. However, if we do fit these, the same seven variables turn out to be most significant as when they are treated as continuous. Furthermore, the terms for blood loss and peritoneal soiling are still dropped from the model.

Term	Coefficient	Std. Error	t value
Constant	-8.74	0.54	-17.32
Mode Surgery	0.42	0.05	8.53
Age	0.62	0.10	6.04
Op Severity	0.26	0.05	5.48
Pulse	0.29	0.09	3.08
WCC	0.39	0.14	2.71
SBP	0.24	0.09	2.64
Hb	0.13	0.05	2.58
Cardiac	0.13	0.06	2.18
Urea	0.14	0.06	2.45
Malignant	0.12	0.07	1.64
Respiratory	0.09	0.06	1.49
Na	0.14	0.09	1.46

Table 46: Model for Individual POSSUM weights achieved by stepwise selection

The ROC curves for this model are shown in Figure 39(a). It can be seen that the predictions are fairly accurate, with some predictive ability being lost, as we would expect, when the model is applied to the test data. Reduced models have been fitted, with 9 weights (removing the bottom 3 above), with 5 weights and then 3, cut off at the dotted lines in the table. An analysis of Deviance of these 4 models shows that the difference between the model containing 12 variables' weights and the one containing 9 is 7.12. This is less than χ^2 (3;0.95) so there is no evidence of a better model with malignant, respiratory and sodium left in. However, removal of the next 4 terms does significantly increase the Deviance, by a further 28.4. Likewise, removing pulse and WCC from the model makes a highly significant difference in Deviance. However we should judge the models also by their performance in prediction. Their ROC curves can be seen in figures 39(b) to 39(d). We can see that the curves deteriorate as the number of variables decreases, and that the test data always give a curve inside the one for the training data. The curves for twelve and three scores appear closer together than for five or nine scores. The average quadratic scores (Q) are given in Table 47, along with the amount of the total variation among patients which is accounted for by the models. It can be seen from these scores that, as we reduce the number of variables in the model, the values of Q for the training and test data sets become closer together. For the training set, better predictions are made with more variables, and as we remove variables, Q increases. However, for the test set, the values decrease, suggesting that a simpler model is more easily transferred to new data. This effect was investigated by Murray (1977), who concluded that larger numbers of variables in discriminant analysis do not necessarily reduce the error rate, due to the bias associated with the choice of subset which happens to give the best discrimination for a given data set. The ROC curves for the four models on the test data set are shown in Figure 40. These show a decreasing predictive ability with reduction in the number of terms in the model.

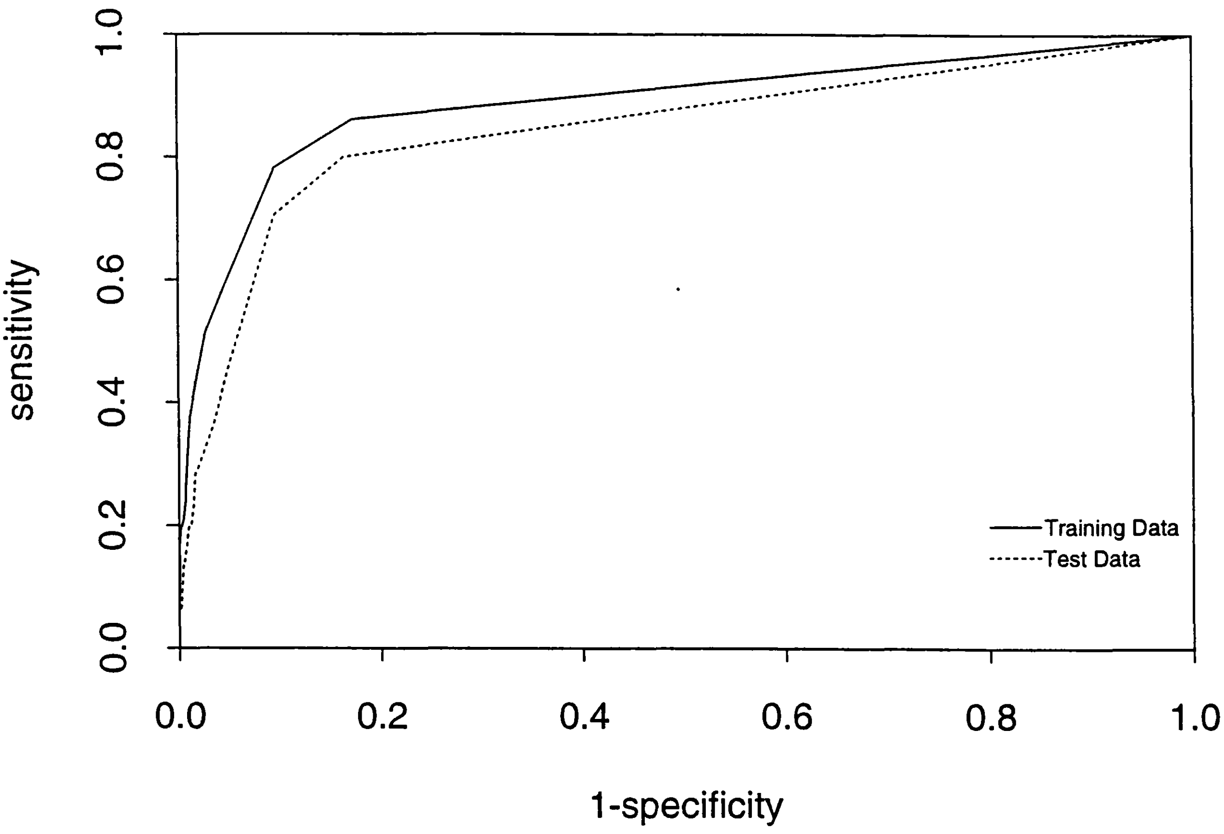


Figure 39(a): ROC Curves for Model containing 12 Individual POSSUM Weights

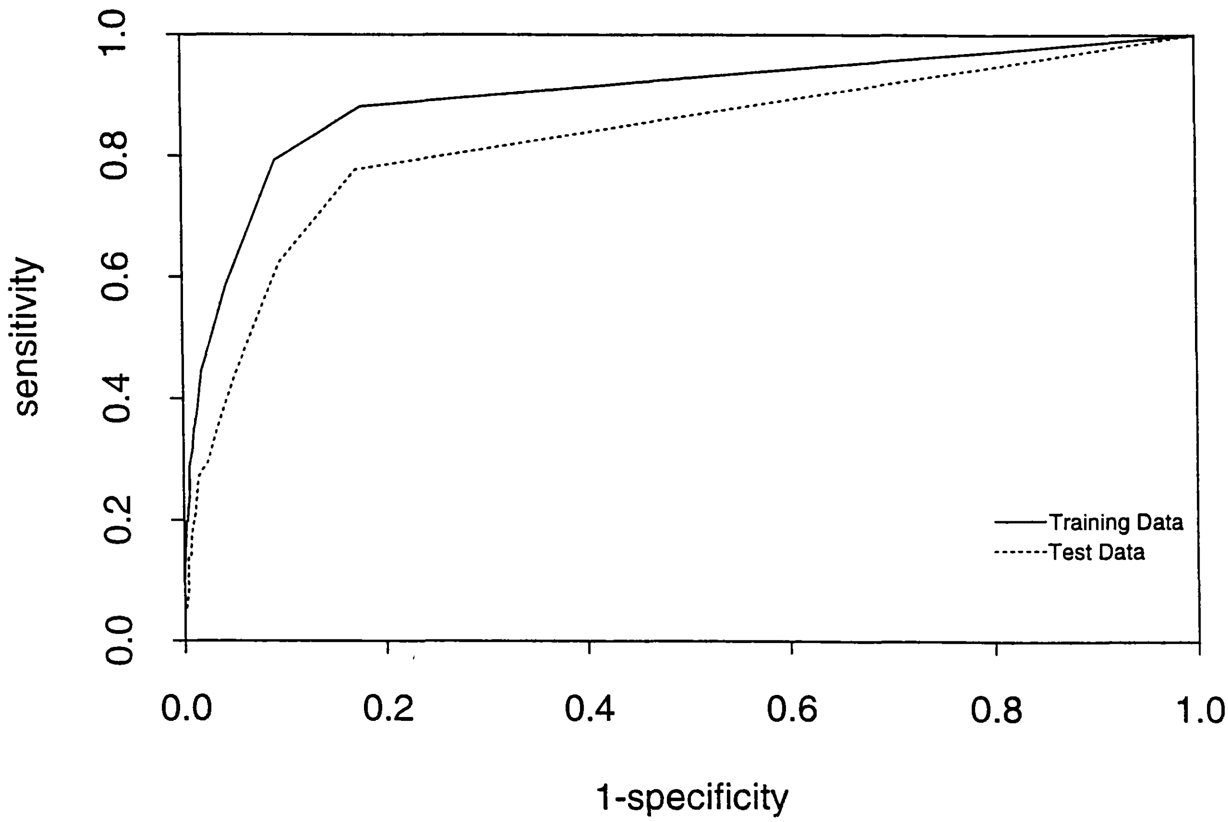


Figure 39(b): ROC Curves for Model containing 9 Individual POSSUM Weights

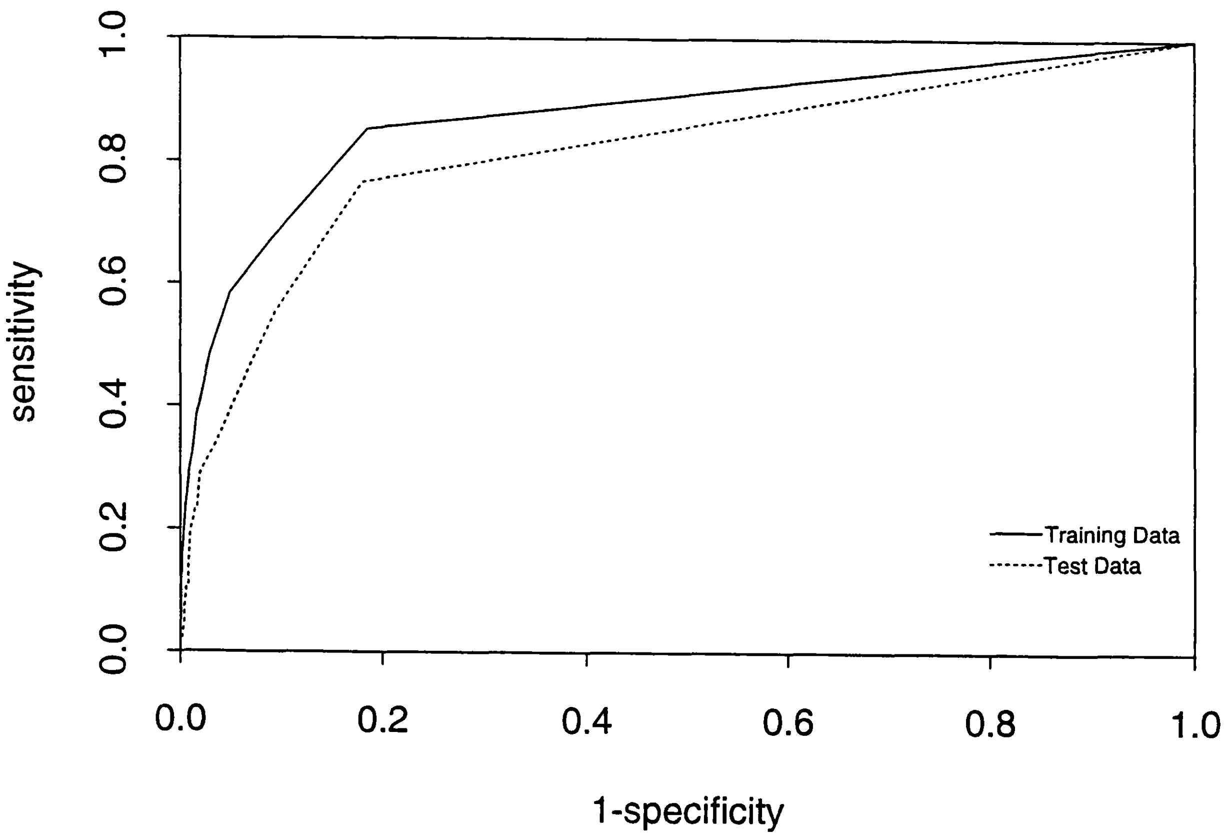


Figure 39(c): ROC Curves for Model containing 5 Individual POSSUM Weights

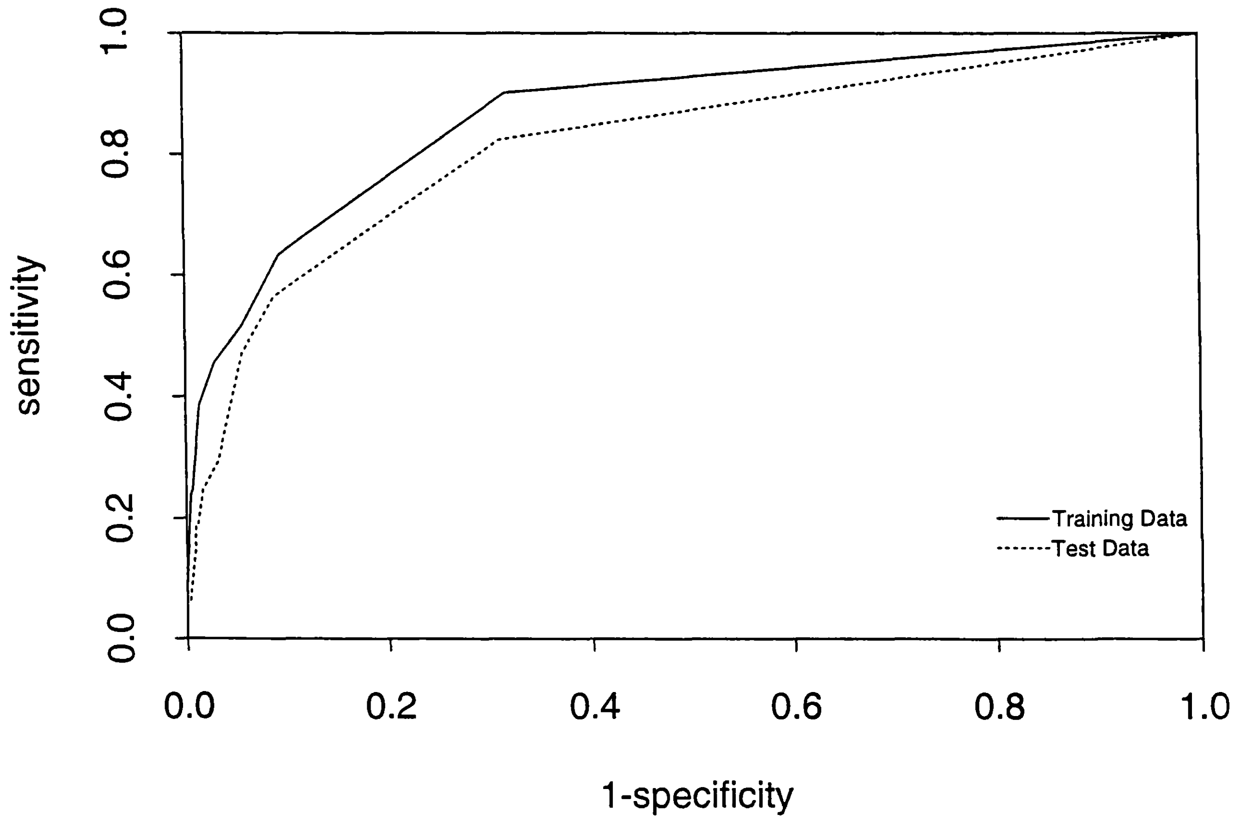


Figure 39(d): ROC Curves for Model containing 3 Individual POSSUM Weights

Model	Data set	Q	SSTa (% of SSTy)
12 weights	training	0.0437	24.39 (24.9)
(Mode surg - Na)	test	0.0487	6.21 (7.7)
9 weights	training	0.0443	23.47 (24.0)
(Mode surg - Urea)	test	0.0494	5.08 (6.3)
5 weights	training	0.0455	21.38 (21.8)
(Mode surg- WCC)	test	0.0507	3.37 (4.2)
3 weights	training	0.0461	20.56 (21.0)
(Mode surg-Op Sev)	test	0.0489	3.89 (4.8)

Table 47: Summary Performance Measures for Models based on Individual Weights

It would be informative here to look at the actual mortality rates within regions of predicted mortality, to see whether the extra predictive power gained by using 12 variables' weights is substantial enough to merit using this rather than the simple model with 3 terms which had the second smallest Q value for the test data. The values for the test data set only are given in Table 48.

	Three Term Model			Twelve Term Model		
Predicted probability of death	Number of patients	Number of deaths	Mortality Rate	Number of patients	Number of deaths	Mortality Rate
<0.25%	0	0	-	1150	1	0.09%
0.25%-0.5%	1267	2	0.16%	500	4	0.80%
0.5%-1%	557	9	1.62%	504	3	0.60%
1%-2%	427	4	0.94%	469	5	1.07%
2%-3%	567	11	1.94%	191	6	3.14%
3%-5%	183	11	6.01%	152	6	3.95%
5%-10%	113	8	7.08%	168	21	12.50%
10%-20%	146	19	13.01%	94	12	12.77%
20%-50%	33	8	24.24%	60	11	18.33%
≥50%	41	13	31.71%	45	16	35.56%

Table 48: Comparisons of Predicted and Actual Mortality Rates for two models containing individual POSSUM weights: Test Data

From the table, it appears as though the model incorporating 12 weights is performing substantially better than the one containing 3. This model accounts for 24.9% of the total variation in the training data set and for 7.7% in the test set.

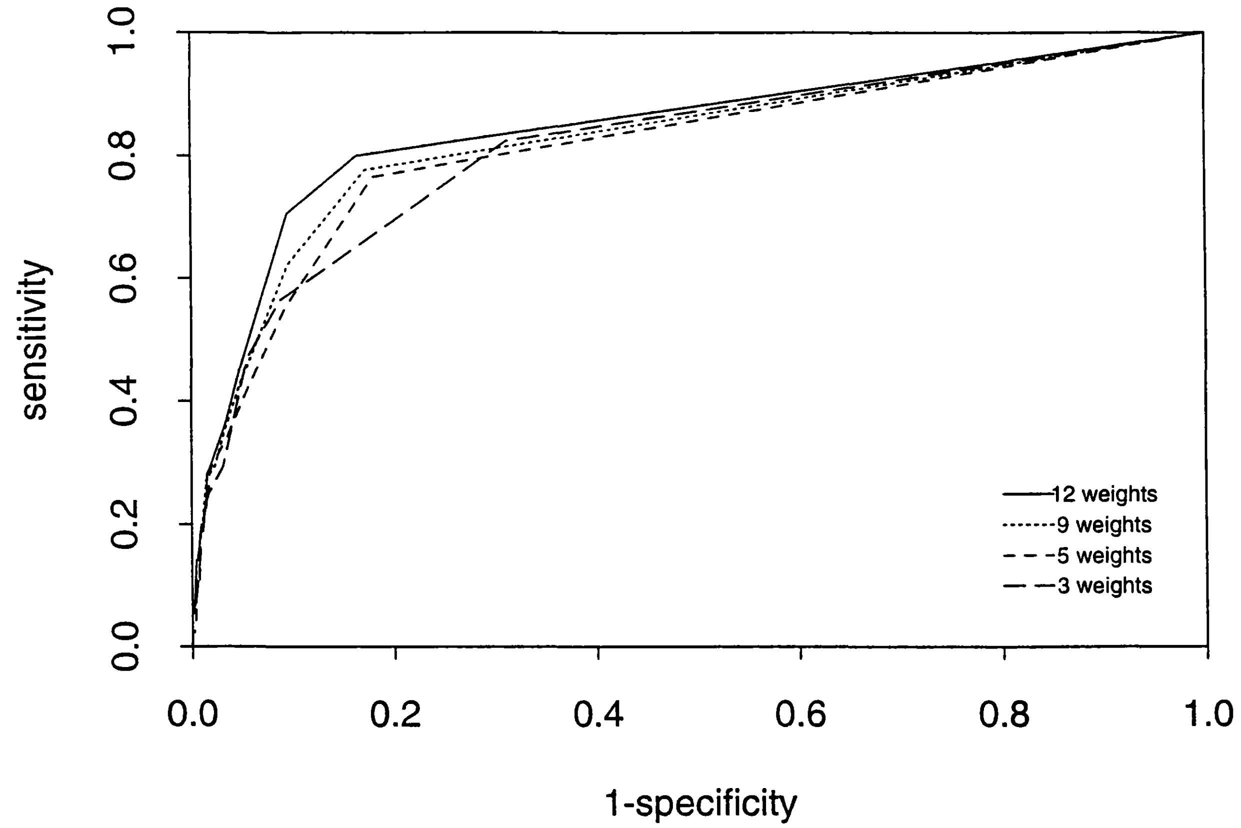


Figure 40: ROC Curves for Four Models containing Individual POSSUM Weights (Test Data Set)

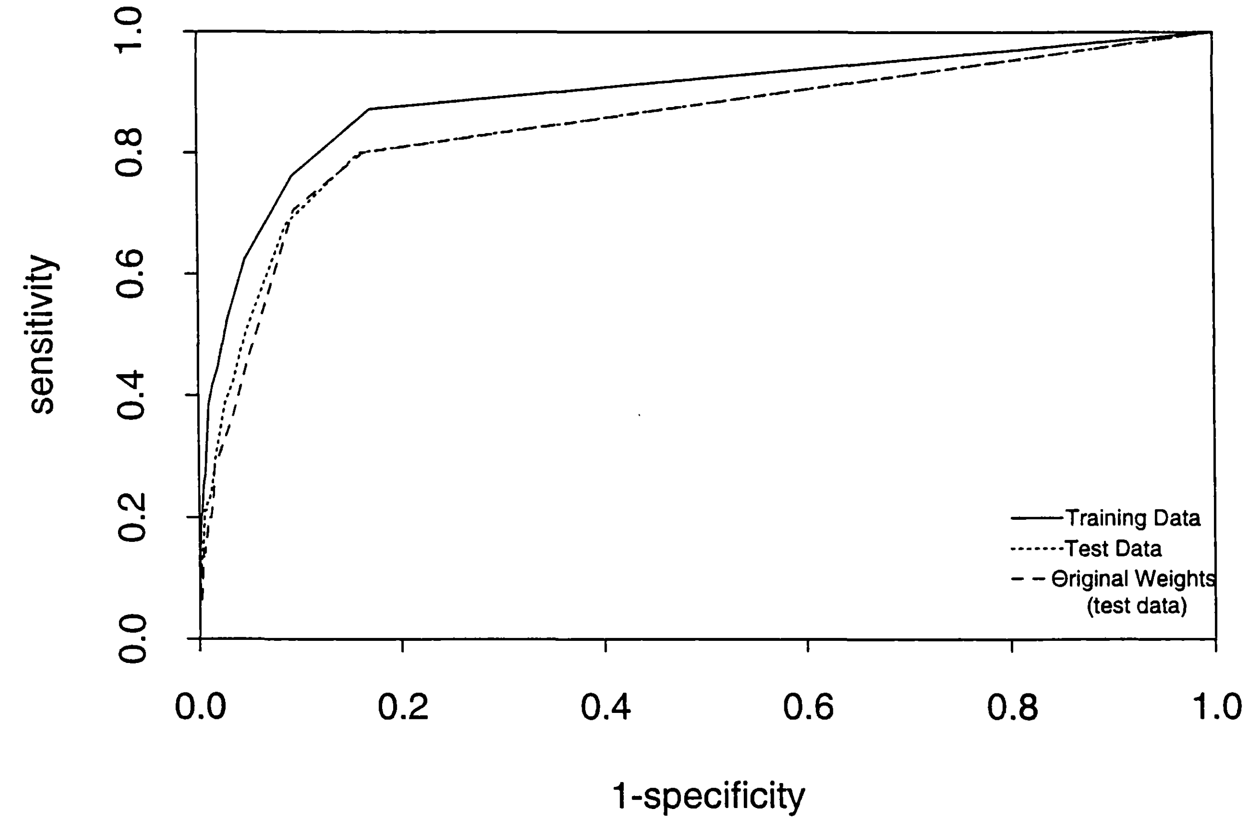


Figure 41: ROC Curves for Model Containing Experimental POSSUM Weights for 12 Variables

8.2.1 Stepwise selection of experimental weights

If the modified scores as described in Table 43 are included in a stepwise logistic regression, the model obtained is very similar to the one obtained from the original scores. The residual Deviance is reduced by 4.3 from before. The modified age score is slightly less significant, whereas the operative severity one is improved. Modified potassium score now stays in the model instead of sodium. White blood cell count contributes less to the deviance than before, but haemoglobin and cardiac are more important, as are respiratory history and presence of malignancy. The model achieved by stepwise regression is given in Table 49.

Experimental Weight	Coefficient	s.e.(coeff)	t
(Intercept)	-8.55	0.55	-17.41
Mode Surgery	0.42	0.05	8.60
Age	0.59	0.10	5.87
Operative Severity	0.29	0.05	6.07
Pulse	0.31	0.09	3.39
Cardiac	0.26	0.10	2.72
Hb	0.13	0.05	2.66
WCC	0.13	0.06	2.19
SBP	1.11	0.05	2.42
Urea	0.12	0.06	2.27
Respiratory	0.29	0.13	2.19
Potassium	0.22	0.11	2.12
Malignant	0.12	0.07	1.92

Table 49: Model from stepwise logistic regression of experimental POSSUM weights

The coefficient for malignant is not significant at the 5% level, but we shall leave it in the model at present. The ROC curves for this model of 12 weights are plotted in Figure 41 for the training and test data sets. The curve for the model containing 12 original weights is also included for comparison. From this it appears that the model containing experimental scores is performing only very slightly better than the one containing the original scores. The values of Q obtained are 0.0464 for the test data and 0.0432 for the training set. These are the lowest values obtained for any model so far, and suggest that a model using the modified scores may be best. The model accounts for 11.7% of the variation in the test data set, which is the highest for any model we have explored. We should again look at the predictive accuracy of the model by categorising the probabilities. This can be seen in Table 50.

	Training Data			Test Data		
Predicted probability of death	Number of patients	Number of deaths	Mortality Rate	Number of patients	Number of deaths	Mortality Rate
<0.25%	925	0	0.00%	953	0	0.00%
0.25%-0.5%	737	3	0.41%	983	4	0.41%
0.5%-1%	538	3	0.56%	571	2	0.35%
1%-2%	435	4	0.92%	441	9	2.04%
2%-3%	190	7	3.68%	167	5	2.99%
3%-5%	172	7	4.07%	167	7	4.19%
5%-10%	167	14	8.38%	143	15	10.49%
10%-20%	108	18	16.67%	79	10	12.66%
20%-50%	70	20	28.57%	73	14	18.18%
≥50%	39	25	64.10%	46	19	41.30%

Table 50: Comparisons of Predicted and Actual Mortality Rates for experimental POSSUM Score Model (twelve variables' weights)

This model performs well, although again, the predictive accuracy in the test data is not nearly as good as in the training data. We now investigate what happens when we use various subsets of these experimental scores. We will separate the variables at the dotted lines in Table 49. This was based on the analysis of deviance as well as the significance of the coefficients. The least significant term was removed individually at each stage. We consider four models, with eleven scores, seven scores, five scores and three scores. The ROC curves for the training and test data sets for each of these four models are given in Figure 42. From these we would conclude that the 11 score model is preferable. This can be seen more clearly in Figure 43, where the curves for the test data set for each of the four models are plotted. Table 51 shows the values of Q and of SSTa for each of these models when used to make predictions on both the training and test data sets.

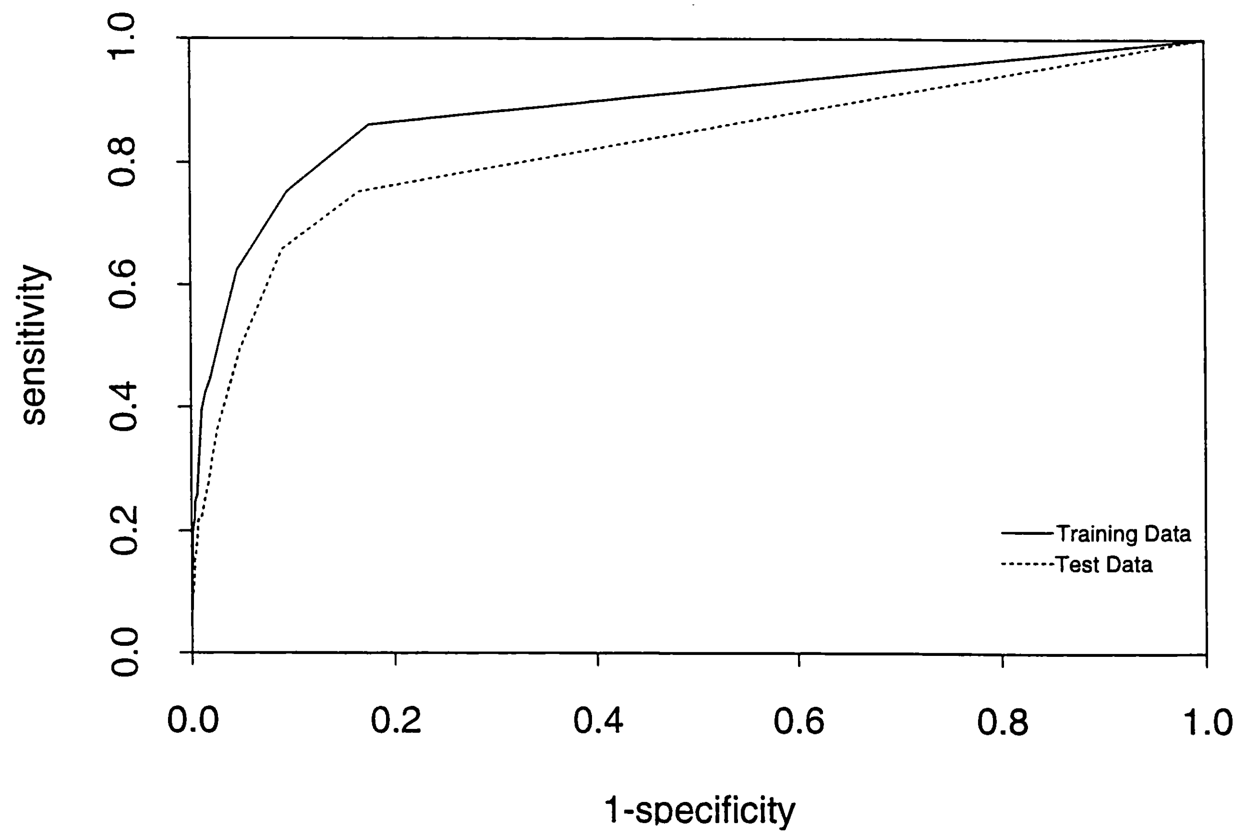


Figure 42(a): ROC Curves for Model containing 11 experimental weights

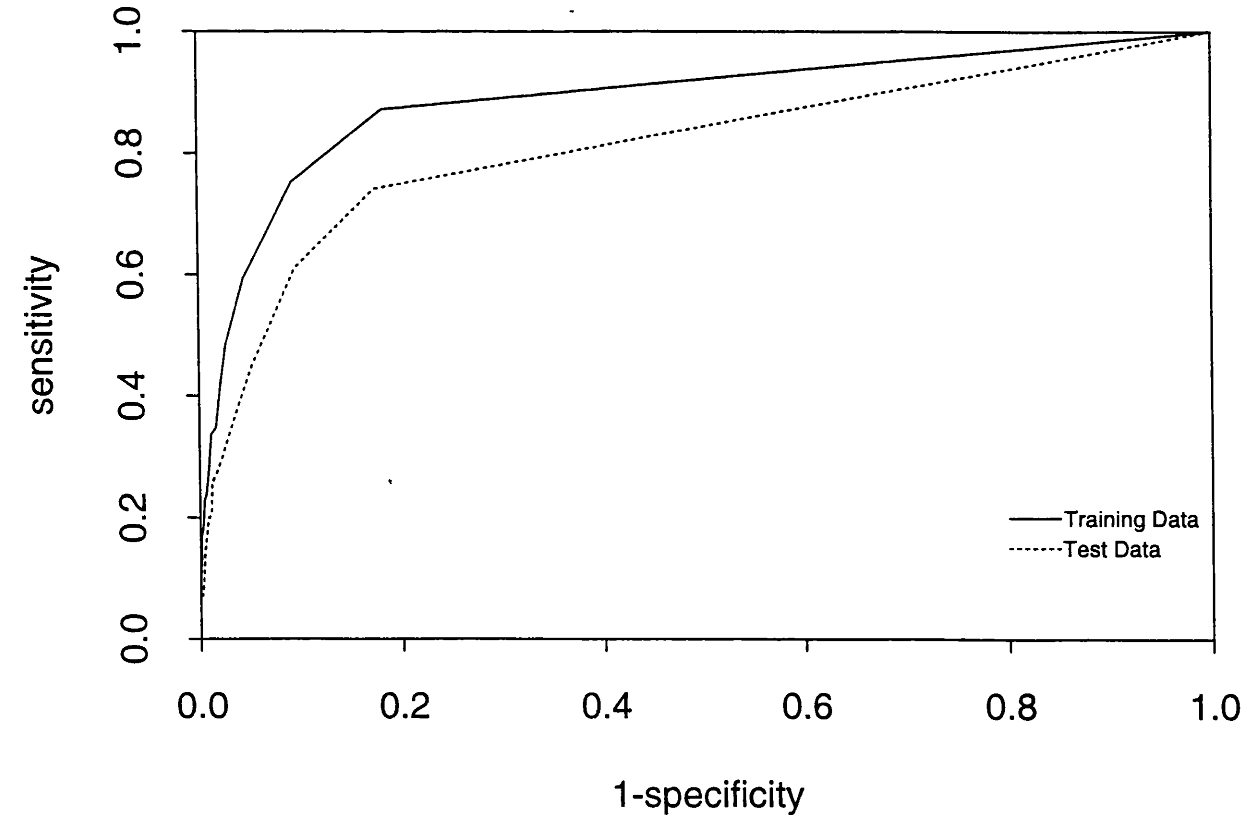


Figure 42(b): ROC Curves for Model containing 7 experimental weights

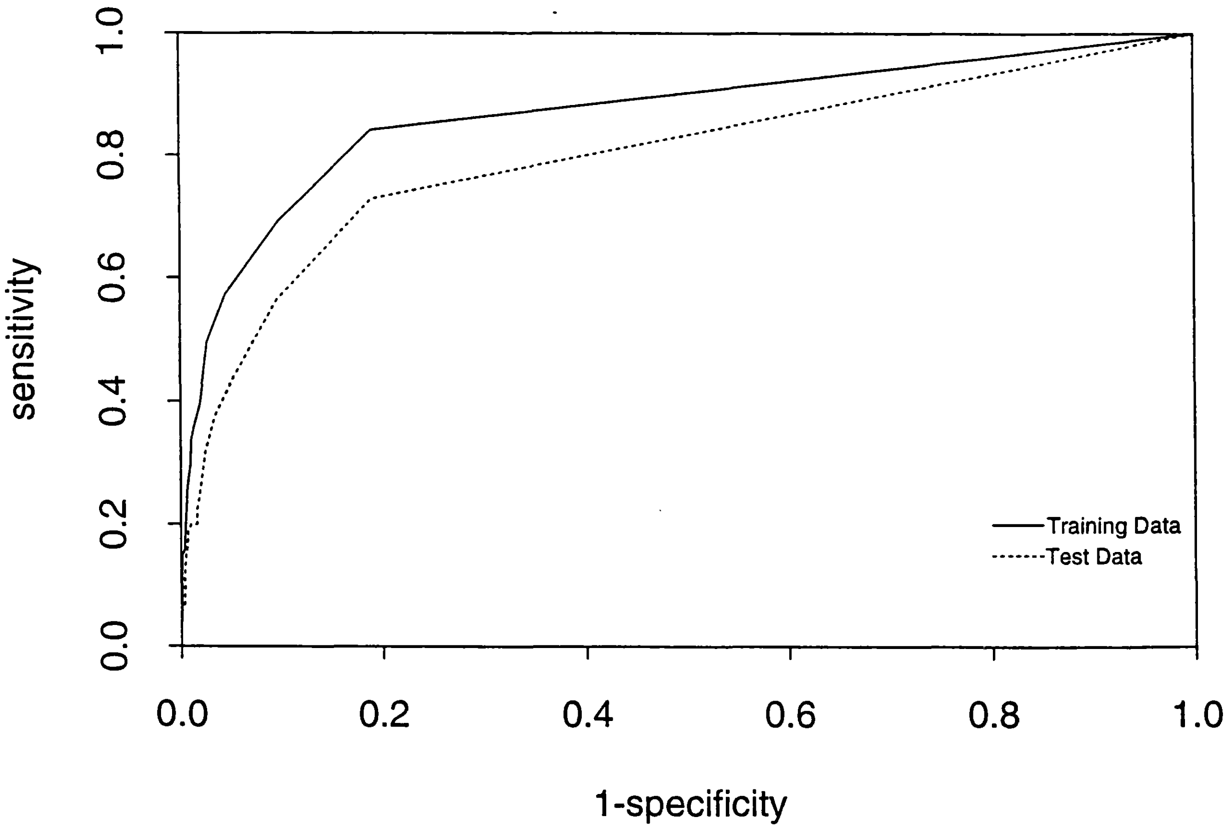


Figure 42(c): ROC Curves for Model containing 5 experimental weights

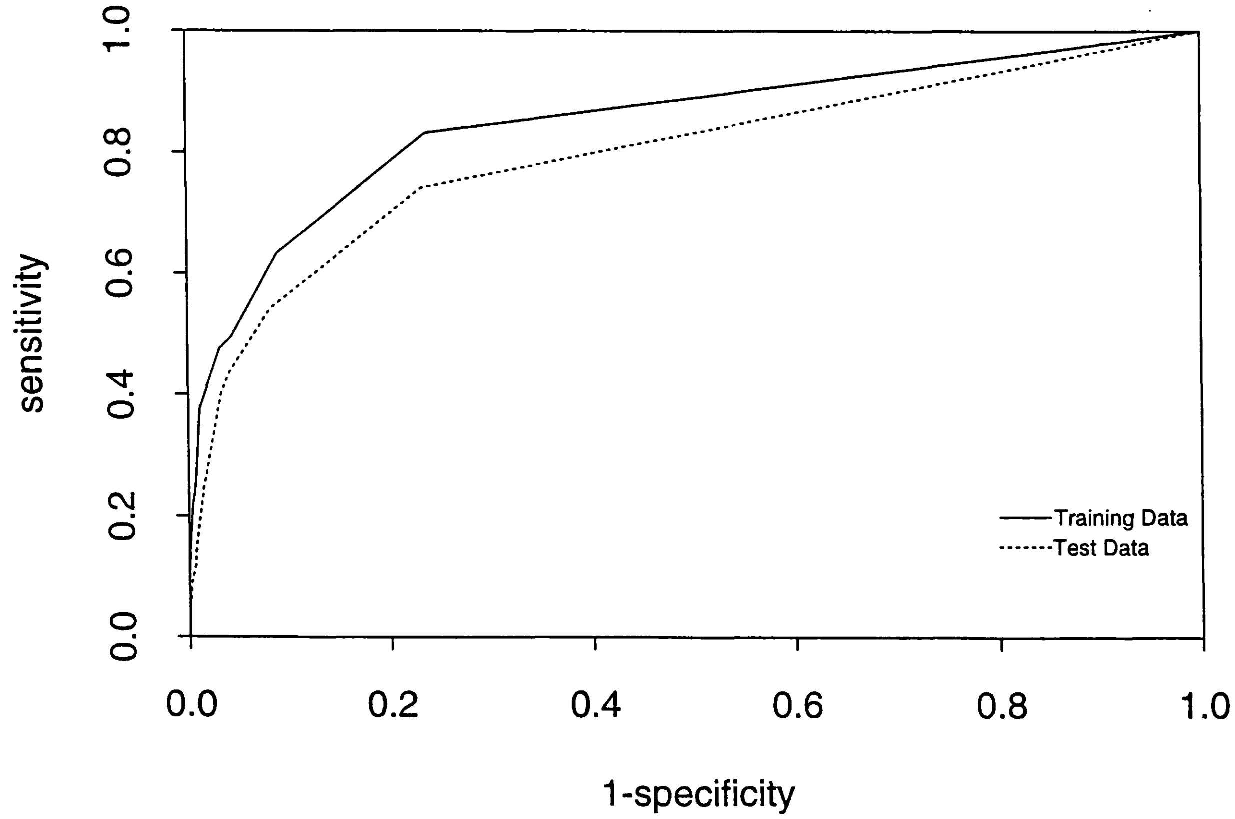


Figure 42(d): ROC Curves for Model containing 3 experimental weights

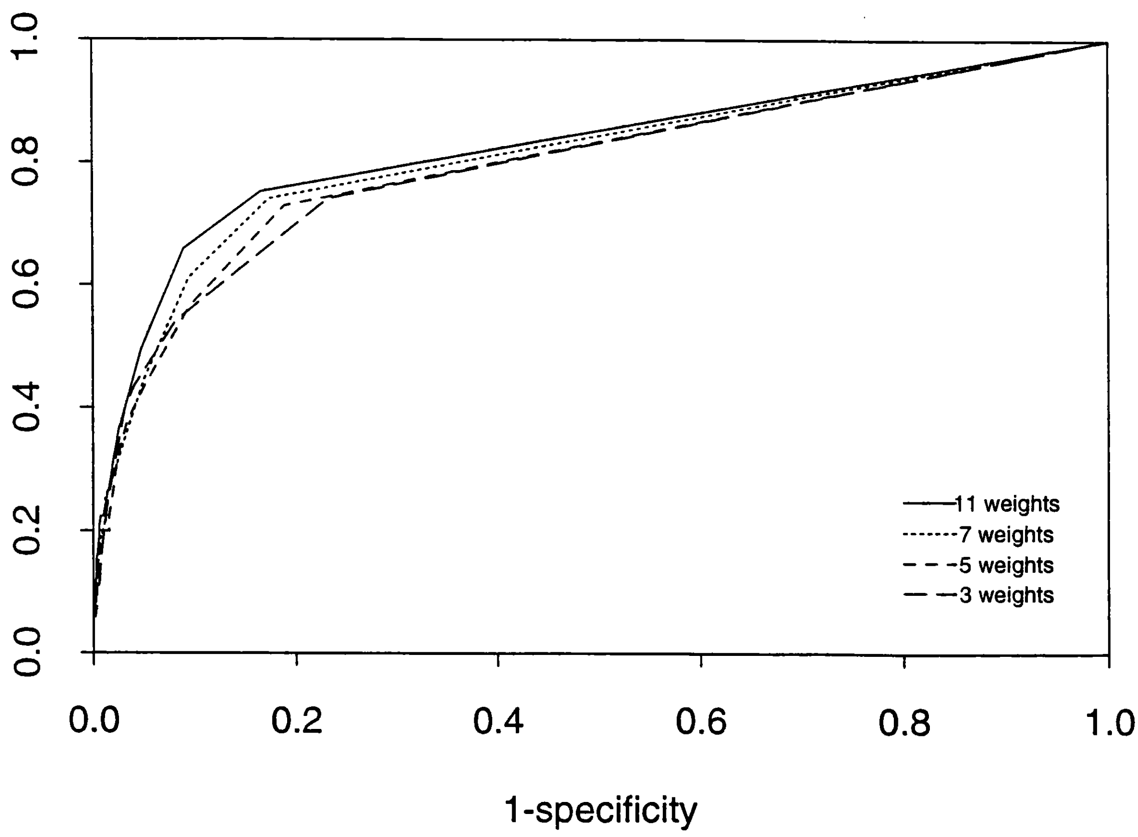


Figure 43: ROC Curves for four Models containing experimental weights (Test Data Set)

Model	Data Set	Q	SSTa (% of SSTy)
11 scores	training	0.0430	25.61 (26.1)
	test	0.0470	8.65 (10.7)
7 scores	training	0.0451	21.95 (22.4)
	test	0.0485	6.83 (8.4)
5 scores	training	0.0456	21.43 (21.9)
	test	0.0494	5.28 (6.5)
3 scores	training	0.0462	20.44 (20.9)
	test	0.0467	6.47 (8.0)

Table 51: Summary Statistics for comparison of Four Experimental Weight Models

From table 51, one would choose either the model with 11 weights or the one with 3 weights. The Q values are all lower than the corresponding ones achieved from the original POSSUM scoring (Table 47). The percentage of variation accounted for by the model in the test data set is also improved by using these changes in scoring method. In order to compare these two models further, we again examine the predicted probabilites. From table 52, it is apparent that the eleven scores give superior predictions. The three variables do not place anyone in the lowest category, and the actual mortality rates do not increase with predicted. Thus we can conclude that the "favourite" model so far is the one with 11 experimental POSSUM scores.

	Eleven Scores			Three Scores		
Predicted probability of death	Number of patients	Number of deaths	Mortality Rate	Number of patients	Number of deaths	Mortality Rate
<0.25%	958	0	0.00%	0	0	-
0.25%-0.5%	647	4	0.62%	1267	2	0.16%
0.5%-1%	599	4	0.67%	652	12	1.84%
1%-2%	443	11	2.48%	561	8	1.43%
2%-3%	176	5	2.84%	36	0	0.00%
3%-5%	165	5	3.03%	506	17	3.36%
5%-10%	148	14	8.46%	142	9	6.34%
10%-20%	84	11	13.10%	100	16	16.00%
20%-50%	66	12	18.18%	32	8	25.00%
≥50%	47	19	40.43%	37	13	35.14%

Table 52: Comparisons of Predicted and Actual Mortality Rates for two experimental POSSUM Score Models: Test Data

8.3 Modelling actual or transformed variables

It may be the case that by giving weights to the variables we are losing valuable information, so now, rather than model the POSSUM weights, we include actual measurements of the variables, or transformations of them, as were described in section 7.2.2. Categorical variables are fitted as "treatment" contrasts. This means that each level of the variable is compared with the first level. This, in effect sets the coefficient of the first level to zero. This is sometimes known as the "Corner Point Constraint" and is the default method used by BMDP LR. Missing values of continuous variables are given the mean value of the data set, and transformed variables are assumed to be 0. For categorical variables, patients with missing data are assumed to belong to the low risk or normal category. Again, a stepwise procedure has been used in order to select the most significant variables.

The model selected is shown in table 53. Malignant and Haemoglobin were also selected, but with very low deviance values they did not make a significant contribution to the model and could be removed. The analysis of deviance considers the terms added in the order they are given in Table 53. This model has a residual deviance of 543.64 on 3367 degrees of freedom. This is significantly lower than the value of 568.46 obtained from the model using the POSSUM weights. Two reduced models have also been fitted, with cut off points for the variables at the dotted lines in the Table 53.

ROC curves have been plotted for all three of these models on the training (Figure 44(a)) and test data sets (Figure 44(b)). From these plots, the 8 variables model appears to perform best on the training data set, but there does not appear to be much difference between the curves of the 11, 8 and 4 variable models when they are applied to the test data set. The average quadratic scores and amounts of variation accounted for by the models are given in Table 54. The models for variables are doing worse on the test data set than the null model, and much worse than any of the models using scores. The predictions on the training data set are rather good, with lower Q values and more of the variation accounted for than any of the score models. However, the models with variables do not transfer as well to new data as those with scores. This is likely to be because models with variables are very specific to the variability in the training data set and so do not generalise well.

Term	Coefficient	Std. Error	t	Deviance	(Pr(Chi))
Constant	-10.03	1.15	-12.79		
Age	0.06	0.01	5.92	108.53	(0.000)
Emerg	1.29	0.28	4.69	93.45	(0.000)
Mode Surg	2.21	0.34	6.54	64.91	(0.000)
Op Severity: Major+	1.51	0.28	5.32	41.91	(0.000)
Op Severity: Comp Major	1.78	0.35	5.12		
|WCC - 7|	0.06	0.02	2.82	11.95	(0.001)
Myocard	1.04	0.36	2.87	8.45	(0.004)
|Pulse - 70|	0.03	0.01	3.29	8.64	(0.002)
Urea >10	0.83	0.28	2.92	10.43	(0.001)
K < 3.5 or >5	0.77	0.30	2.53	5.66	(0.017)
JVP	1.24	0.51	2.42	4.28	(0.039)
|Sys BP - 130|	0.01	0.01	2.38	5.28	(0.021)

Table 53: Model for variables selected by stepwise logistic regression

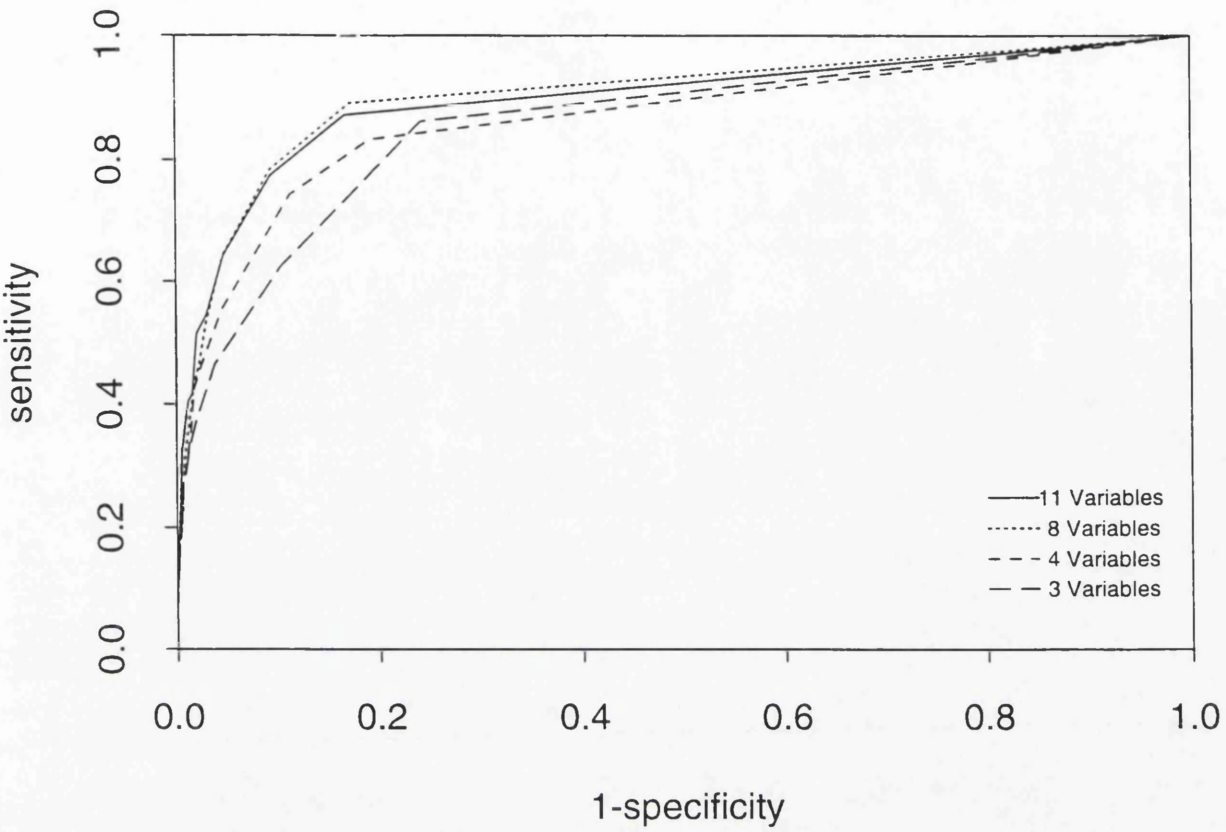


Figure 44(a): ROC Curves for Models from Variables (Training Data)

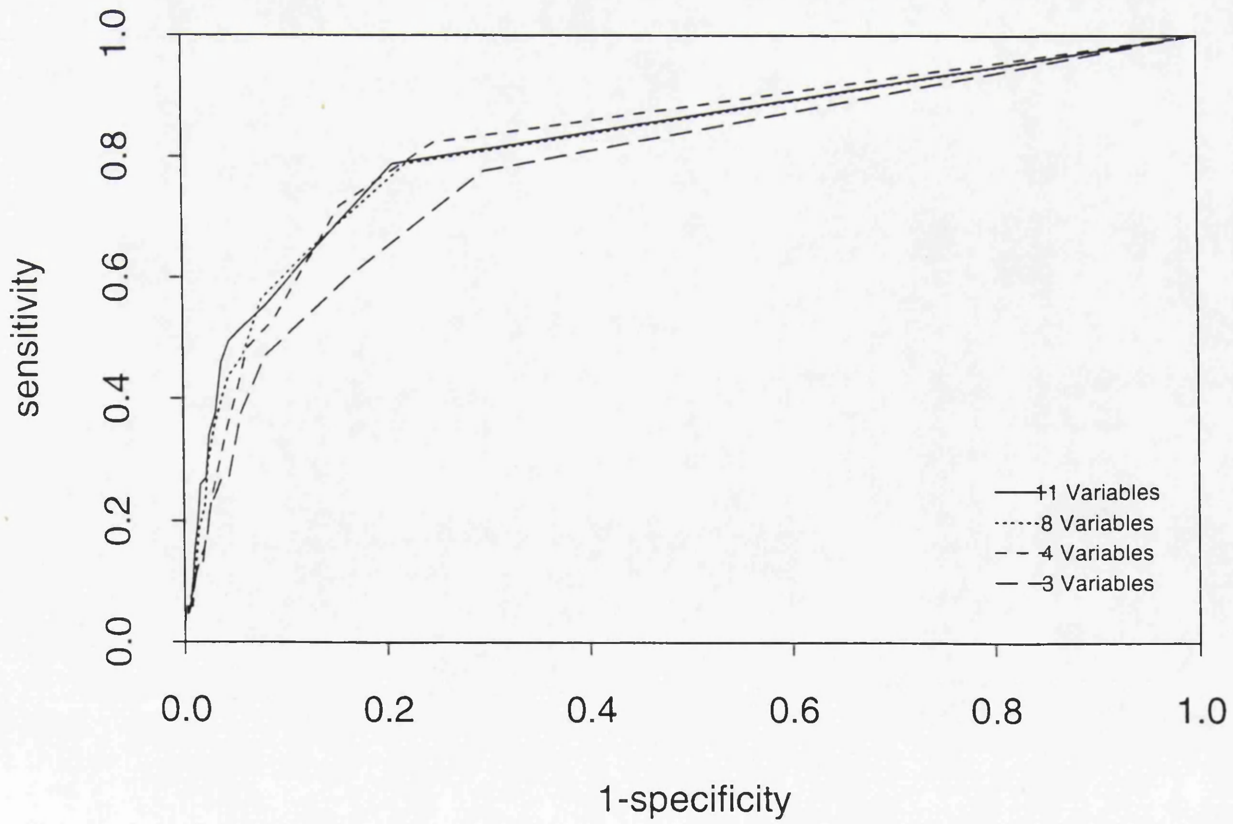


Figure 44(b): ROC Curves for Models from Variables (Test Data)

Model	Data set	Q	SSTa (% of SSTy)
11 variables	training	0.0418	27.61 (28.2)
(Age - Sys BP)	test	0.0531	7.11 (8.8)
8 variables	training	0.0433	25.22 (25.7)
(Age - Urea)	test	0.0536	5.5 (6.8)
4 variables	training	0.0454	21.79 (22.2)
(Age - Op Sev)	test	0.0562	4.27 (5.3)

Table 54: Summary measures of Model Performance for Variables Models

In order to assess the effect of the extra variable for emergency admission available here, but which is not part of the POSSUM scoring system, a model has been fitted using the data for age, mode of surgery and operative severity. This is a far poorer fit than the model containing emergency, as can be seen in the ROC curves for "3 variables" in figures 44(a) and (b), and as shown by the increased average quadratic scores of 0.047 for the test data set and 0.058 for the training data.

8.4 The Best Model

We have seen that weights provide more robust models that perform more satisfactorily on the test data set than do the raw variables. The experimental POSSUM weights provided more accurate predictions than the original ones, and the individual weights in a model were superior to using totals in the form of operative severity and physiological scores. Thus we should be able to choose an improved model which requires less data than the published POSSUM system. A variable which has no missing values, is highly significant and yet is not incorporated into the POSSUM score is whether the admission is emergency. We have seen that this variable made a great improvement to the predictive ability of a model containing other variables. Perhaps we could include a weight for this when modelling the other weights. We could assign 1 to elective admissions and 8, say, to emergency in keeping with the POSSUM style of weighting. Of course the choice is arbitrary since we are dealing with a variable which can take only 2 values, and its value will simply be reflected in the coefficient in the model. If we add our weight for emergency admission to the previous model of experimental weights, we find it is the next most significant term after operative severity, mode of surgery and age. Potassium is no longer significant when emergency is included. A model with 11 terms is given in table 55. Again we shall consider models of subsets of the weights, split at the broken lines in the table.

Term	Coefficient	s.e. (coeff)	t
(Intercept)	-8.55	0.54	-17.60
Operative Severity	0.32	0.05	6.58
Age	0.58	0.10	5.82
Mode Surgery	0.32	0.06	5.75
Emergency	0.21	0.06	3.47
Pulse	0.26	0.09	2.77
Cardiac	0.26	0.10	2.76
Haemoglobin	0.14	0.05	2.91
SBP	0.10	0.05	2.22
White Cell Count	0.12	0.06	2.06
Urea	0.13	0.06	2.29
Respiratory	0.31	0.13	2.35

Table 55: Model with new POSSUM Weights, including emergency admission

To compare these four models we can look at the summary statistics in table 56, and the ROC curves. Surprisingly, the model with four terms has the lowest value of Q for the test data. In fact, this is the only model where the value of Q for the test data set is lower than the one for the training data. Again we have a choice between including a fairly large number of terms and a minimal number. Figure 45 shows the ROC curves for the training and test data for these four models. The predictive power in the training set decreases with the number of variables in the model, and the curves appear to move closer together. This is confirmed if we compare the ROC curves for the models applied to the test data in Figure 46, which overlap. The effect of adding emergency score to the model can be seen in Figure 47, where the ROC curves for the two best models here, and without emergency are plotted. The model with 4 weights including emergency is performing almost as well as the one with 11 without emergency, but is not as good as the model with 11 including emergency.

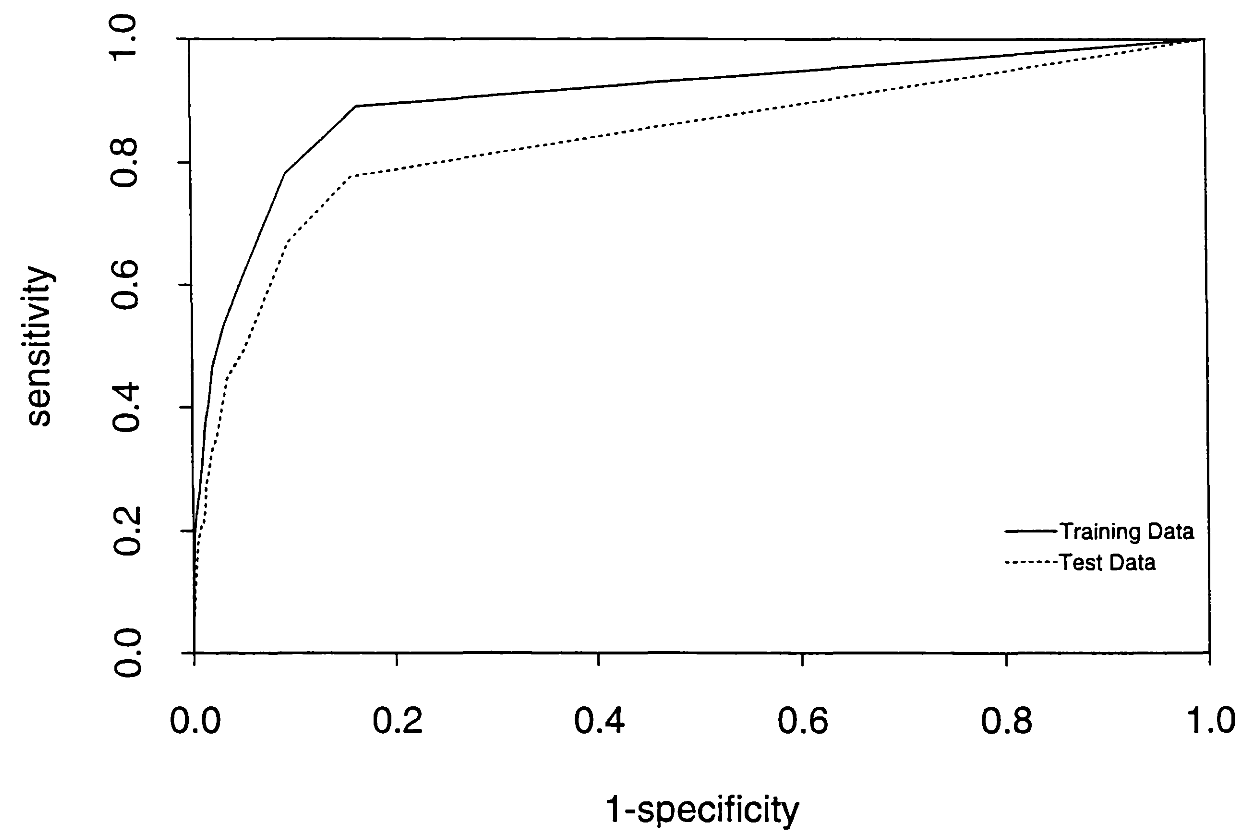


Figure 45(a): ROC Curves for Model Containing 11 Experimental Weights Including Emergency Admission

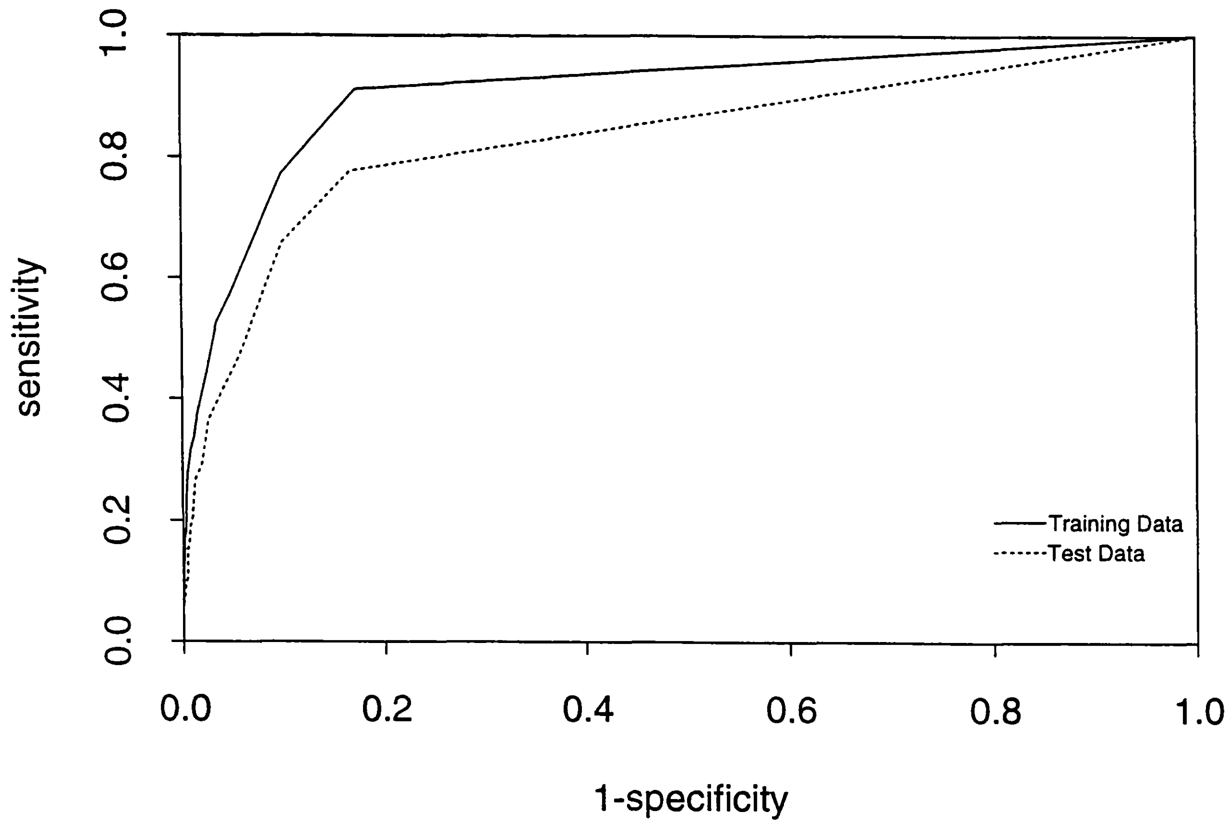


Figure 45(b): ROC Curves for Model Containing 9 Experimental Weights Including Emergency Admission

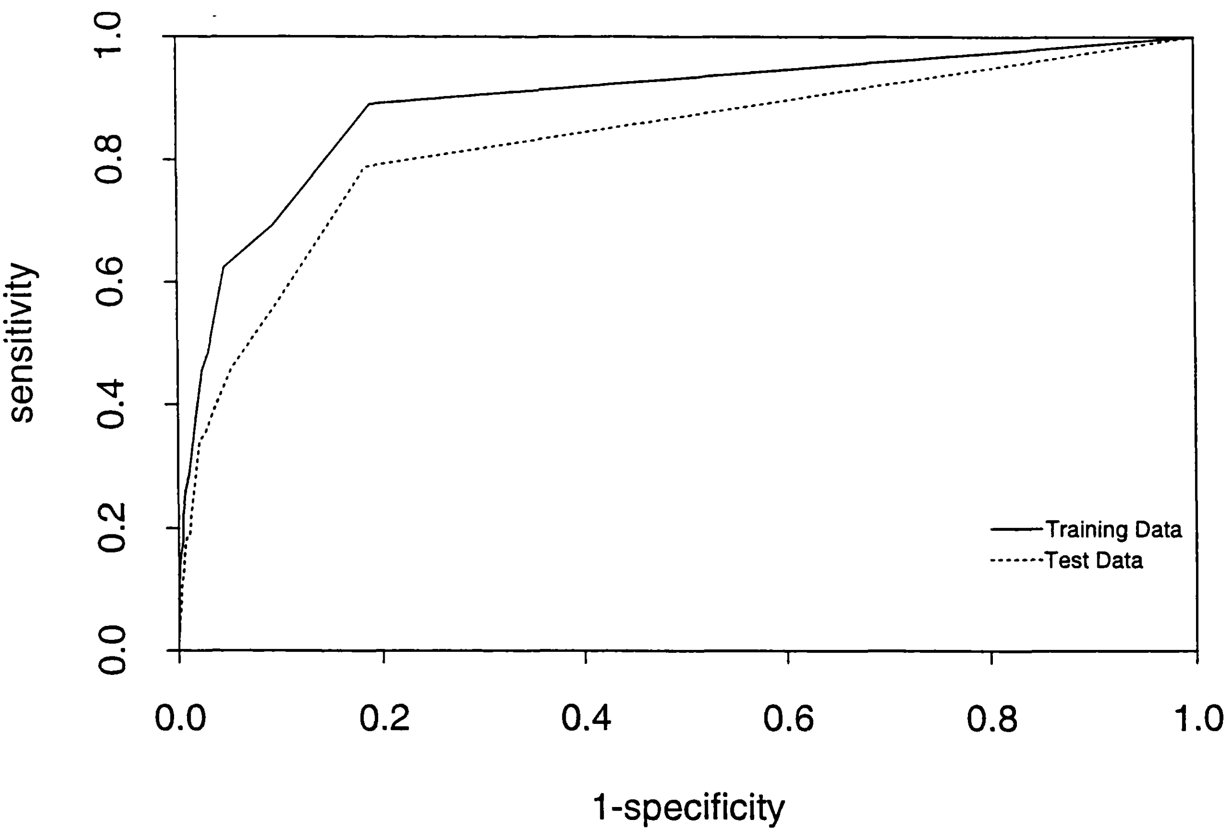


Figure 45(c): ROC Curves for Model Containing 6 Experimental Weights Including Emergency Admission

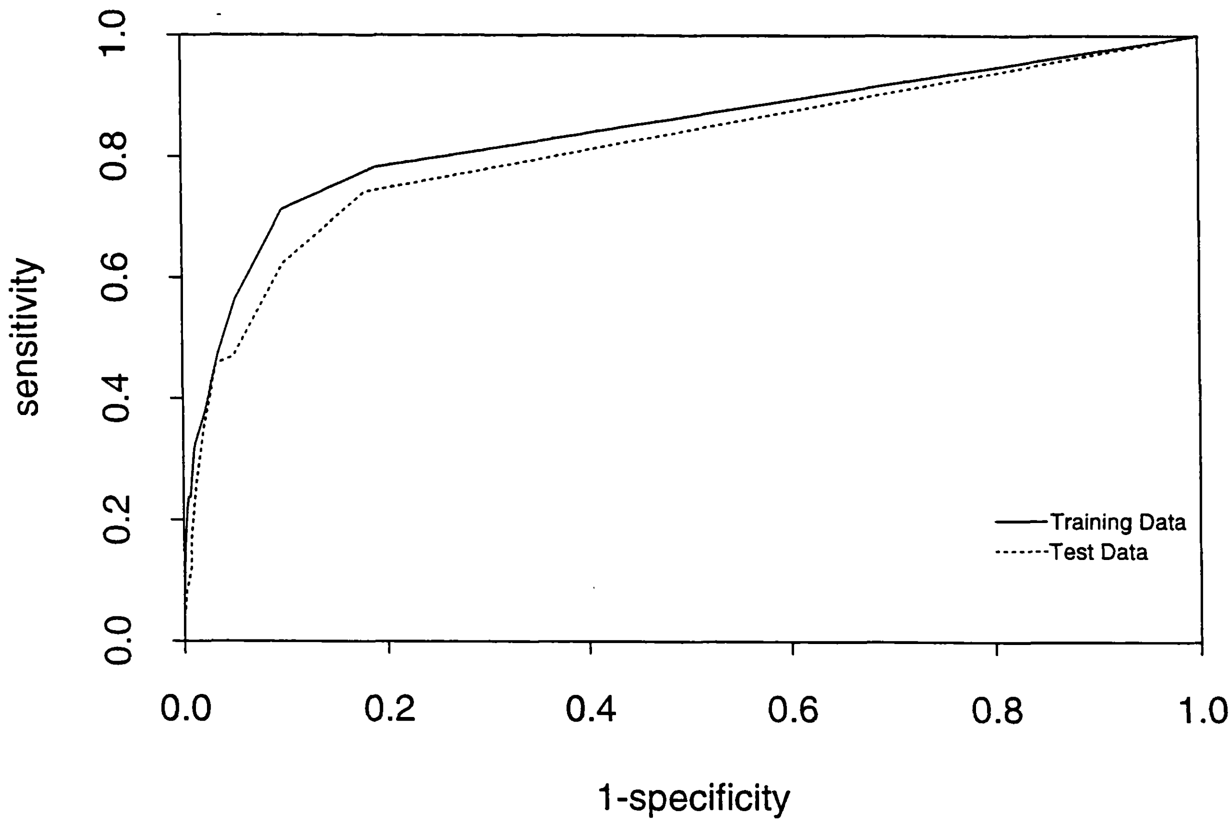


Figure 45(d): ROC Curves for Model Containing 4 Experimental Weights Including Emergency Admission

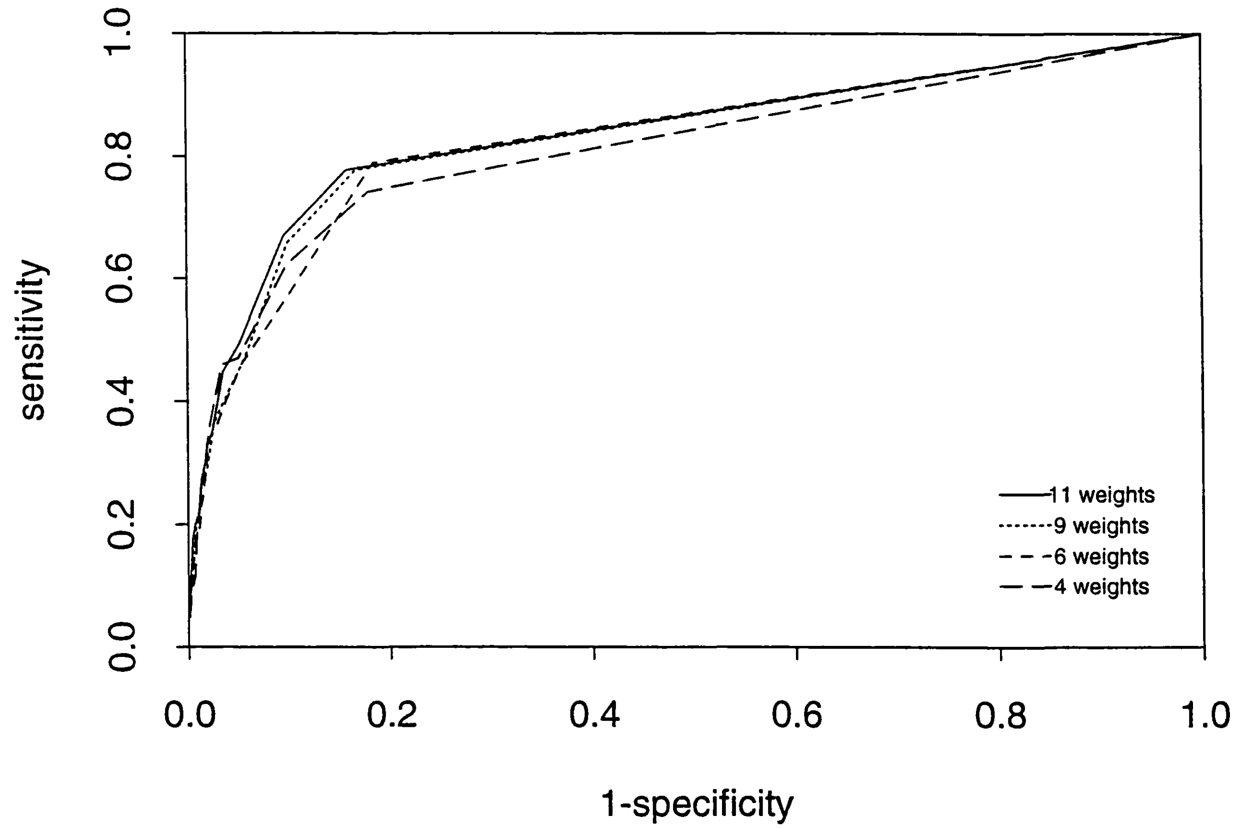


Figure 46: ROC Curves Comparing Models Containing Experimental Weights Including Emergency Admission (Test Data)

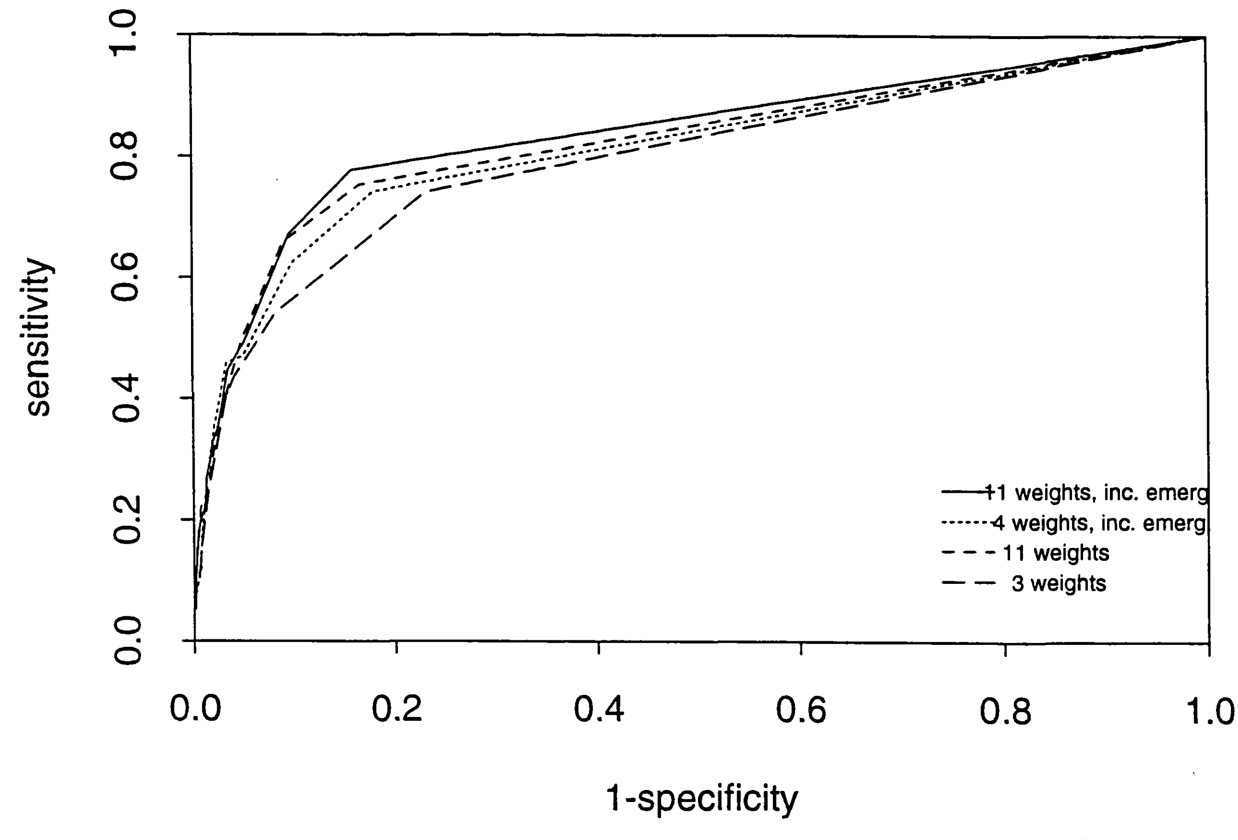


Figure 47: ROC Curves Comparing Models Containing Experimental Weights With and Without Emergency Admission (Test Data)

Model	Data	Q	SSTa (% of SSTy)
11 scores	Training	0.0436	24.75 (25.3)
	Test	0.0464	8.13 (11.3)
9 scores	Training	0.0442	23.48 (24.0)
	Test	0.0477	7.73 (8.6)
6 scores	Training	0.0454	21.66 (22.1)
	Test	0.0484	6.69 (8.3)
4 scores	Training	0.0460	20.77 (21.2)
	Test	0.0457	8.06 (10.0)

Table 56: Summary statistics for models with experimental POSSUM Weights, including emergency admission

We need to balance the loss in predictive ability with the increased simplicity of the model. In order to compare the two models further, we again study the mortality rates in categories of predicted probability of death. These can be seen for these two models in table 57(a) for the training data and 57(b) for the test data. The model with four scores has the observed mortality rate contained within the category of predicted rate 7 times for the training data, compared with 5 times for the model with 11 scores. However, the observed values in the smaller model tend to fluctuate, whereas the ones in the larger model increase monotonically throughout the categories, suggesting superior calibration. The same is true of the test data set, although here, both models have five categories containing the observed mortality rate. It makes sense here to choose the simplest model, as it does not perform considerably worse than the one with eleven variables.

	Eleven Scores			Four Scores		
Predicted probability of death	Number of patients	Number of deaths	Mortality Rate	Number of patients	Number of deaths	Mortality Rate
<0.25%	1146	1	0.09%	839	0	0.00%
0.25%-0.5%	575	1	0.17%	719	2	0.28%
0.5%-1%	490	4	0.82%	410	2	0.49%
1%-2%	446	4	0.90%	504	16	3.17%
2%-3%	174	2	1.15%	372	4	1.08%
3%-5%	168	10	5.95%	145	5	3.45%
5%-10%	156	17	10.90%	167	15	8.98%
10%-20%	114	15	13.16%	119	19	15.97%
20%-50%	77	24	31.17%	72	16	22.22%
≥50%	35	23	65.71%	34	22	64.71%

Table 57(a):Comparisons of Predicted and Actual Mortality Rates for two experimental POSSUM Score Models including emergency admission: Training Data

	Eleven Scores			Four Scores		
Predicted probability of death	Number of patients	Number of deaths	Mortality Rate	Number of patients	Number of deaths	Mortality Rate
<0.25%	1201	2	0.17%	899	1	0.11%
0.25%-0.5%	521	1	0.19%	706	5	0.71%
0.5%-1%	478	6	1.26%	346	3	0.87%
1%-2%	469	7	1.49%	484	12	2.48%
2%-3%	155	6	3.87%	396	5	1.26%
3%-5%	145	6	4.14%	130	6	4.62%
5%-10%	157	15	8.55%	172	13	7.56%
10%-20%	98	12	12.24%	105	10	8.52%
20%-50%	63	12	18.05%	57	17	28.82%
≥50%	46	18	38.13%	38	13	34.21%

Table 57(b):Comparisons of Predicted and Actual Mortality Rates for two experimental POSSUM Score Models including emergency admission: Test Data

The "best" model is thus as follows.

Operative Severity:	Minor/Intermediate	→ 1
	Major	→ 2
	Major+	→ 4
	Complex Major	→ 8
Age:	Under 60	→ 1
	61-74	→ 2
	75 and Over	→ 4
Mode of Surgery:	Elective	→ 1
	Urgent	→ 4
	Emergency	→ 8
Admission:	Not emergency	→ 1
	Emergency	→ 8

These scores are then incorporated in the model

$$\ln\{R/(1-R)\} = 0.31 \times \text{Operative Severity Score} + 0.74 \times \text{Age Score} + 0.33 \times \text{Mode Surgery Score} + 0.19 \times \text{Admission Score} - 7.71,$$

where R is the expected risk of mortality.

The minimum expected risk from this model is 0.21% and the maximum is 87%. We can compare this new model tested on the test data set with the published POSSUM model by looking at the proportions dying in categories of predicted probability of death (Table 58). It is clear that the new model is a great improvement, with 5 categories containing the observed values compared with none for the published model.

	Published POSSUM			"Best" Model		
Predicted probability of death	Number of patients	Number of deaths	Mortality Rate	Number of patients	Number of deaths	Mortality Rate
<0.25%	0	0	-	899	1	0.11%
0.25%-0.5%	0	0	-	706	5	0.71%
0.5%-1%	0	0	-	346	3	0.87%
1%-2%	1263	3	0.24%	484	12	2.48%
2%-3%	648	1	0.15%	396	5	1.26%
3%-5%	502	6	1.20%	130	6	4.62%
5%-10%	419	8	1.91%	172	13	7.56%
10%-20%	263	26	8.89%	105	10	8.52%
20%-50%	153	17	11.11%	57	17	28.82%
≥50%	85	24	28.24%	38	13	34.21%

Table 58: Comparisons of Predicted and Actual Mortality Rates for published POSSUM Model and "best" model (Test Data)

8.5 Missing Values

Missing values are a great problem in this type of work, as ideally, one would wish to calculate a risk for all patients and not just those for whom a complete data set was available. In this data set almost one third of patients have data missing for some variable that is required to calculate the POSSUM score. To include these patients, one must make certain assumptions. The easiest one is that a value which is not recorded is assumed to be "normal". This is the approach used by POSSUM, where missing measurements score 1, and the above analyses have all been carried out including missing values with normal. When considering data rather than scores, if a continuous variable is missing it is assigned the mean value of the data, and with categorical data the patient is included in the first category, i.e. the condition is assumed to be absent. For the continuous variables which were transformed by subtracting some value, missing observations were assigned the value 0, i.e. they were assumed to be "normal". The problem with this approach is that, for several of the variables, notably pulse, myocardial infarction, dyspnoea, oedema and JVP the missing values correspond to high mortality rates. This information could lead us to include missing values for certain variables in a higher risk category rather than in the normal one. This strategy, however, does not appeal, as it would assume patients had a fairly rare characteristic rather than that they were similar to the majority. It is important to consider why the data may be missing. It could be the case that the data were collected but never entered into the database, or the information may never have been recorded or measured. If they were never recorded, this could point to overworked staff who did not have time to write things down, or a particularly ill patient at an inconvenient time who was a drain on resources. It could also mean that they were never measured, which may point to defects in the quality of care.

We shall explore ways of accounting for missing data in a model, to gain an insight into the effect of missing data. However, incorporating a term for "missing" in a predictive system is likely to limit the applicability of that system to centres other than the one from whose data is devised. This is because if data are missing systematically due to the collection methods and clinical routine, it is highly unlikely that practice will be identical in another centre under different management, where a different bias will arise.

One could attempt to impute the missing values using conditional expectations based on other known values. With these data, however, values tend to be missing for several variables at once. As can be seen in table 59, only 131 patients have only 1 piece of data missing. The most common number of items missing is 3. This is because many patients have no information on sodium, potassium and urea. Almost as frequently, 5 items of data are missing, mainly due to patients also having no information on Haemoglobin and White Cell Count. Similarly, for the POSSUM Operative Severity Score data, all the patients who have no information on severity of the operation have none on any other aspects of the score either.

↓Number missing

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
2364	133	108	327	21	266	88	29	11	11	2	8	0	2	2	6	2	1

↑ Number of patients

Table 59: Numbers with missing data

8.5.1 Investigations of Missing data

8.5.1.1 INCLUDE MISSING VALUES IN A SEPARATE CATEGORY

This approach to missing values involves treating all variables as categorical, and to include "missing" as a separate category. This gives a coefficient to use for missing data, but is still generalising, in a different way from before, about the type of patient who has no data recorded. The variables were categorised, with "high risk" as 2, "low risk" as 1 and "missing" as 0, based on the exploratory analyses described previously. Operative severity had 3 categories plus missing and blood loss had 4 categories plus missing. Table 60 gives the analysis of deviance of the model resulting from a backwards stepwise elimination process.

Variable	df	Deviance	Pr(Chi)
Age	2	77.76	0.0000
Admission	1	86.03	0.0000
Mode Surgery	2	76.65	0.0000
Operative severity	3	57.47	0.0000
WCC	2	11.42	0.0033
Pulse	2	7.85	0.0197
Myocard	2	11.95	0.0025
Urea	2	16.10	0.0003
JVP	2	8.36	0.0153
Sys BP	2	11.37	0.0034
Chest X-ray: Lung	2	6.28	0.0433
Malignant	2	14.01	0.0009

Table 60: Model with all variables categorical, with 'missing' a separate category

This model was more significant than the one obtained when missing values were included as normal, with a residual deviance of 522.8. Thirteen variables were entered into the model, but Potassium has been discarded as it did not contribute substantially to the Deviance. The four most important variables were the same as previously, but there were some differences with the less significant terms. For example, malignant is the next most significant variable if missing is included as a separate category, but was discarded from the model when the missing values were assumed to be non malignant. This is due to the high mortality rate of the patients with this information missing. The other significant variables were as before, but chest X-ray of the lungs was also kept in the model. The ROC curves for this model can be seen in Figure 48. The area under the curve is large for the training data, suggesting a very good fit. For the test data, however, there is not much improvement on the curve of the model with the missing values coded as normal for the test data. This model is by far the best fit to the training data set, with a Q value of 0.0393 and accounts for 32.9% of the total variation present. For the test data, Q is 0.0476 and 11.3% of the variation is accounted for. Table 61 gives the classified predicted and actual mortality rates. This suggests that the treatment of missing data as separate provides a model which is too specific to a particular training data set. Moreover, it is unlikely that such a model would travel well between centres.

	Training Data			Test Data		
Predicted probability of death	Number of patients	Number of deaths	Mortality Rate	Number of patients	Number of deaths	Mortality Rate
<0.25%	1323	2	0.00%	1350	3	0.22%
0.25%-0.5%	465	1	0.22%	433	2	0.46%
0.5%-1%	630	5	0.79%	652	6	0.92%
1%-2%	335	3	0.90%	322	9	2.80%
2%-3%	91	3	3.30%	106	10	8.43%
3%-5%	187	8	4.28%	153	7	4.58%
5%-10%	131	12	8.16%	115	9	7.83%
10%-20%	99	14	14.14%	77	7	8.09%
20%-50%	71	17	23.94%	73	11	15.07%
≥50%	49	36	73.47%	52	21	40.38%

Table 61: Predicted and actual mortality rates for model where all variables are categorical, with separate categories for "missing".

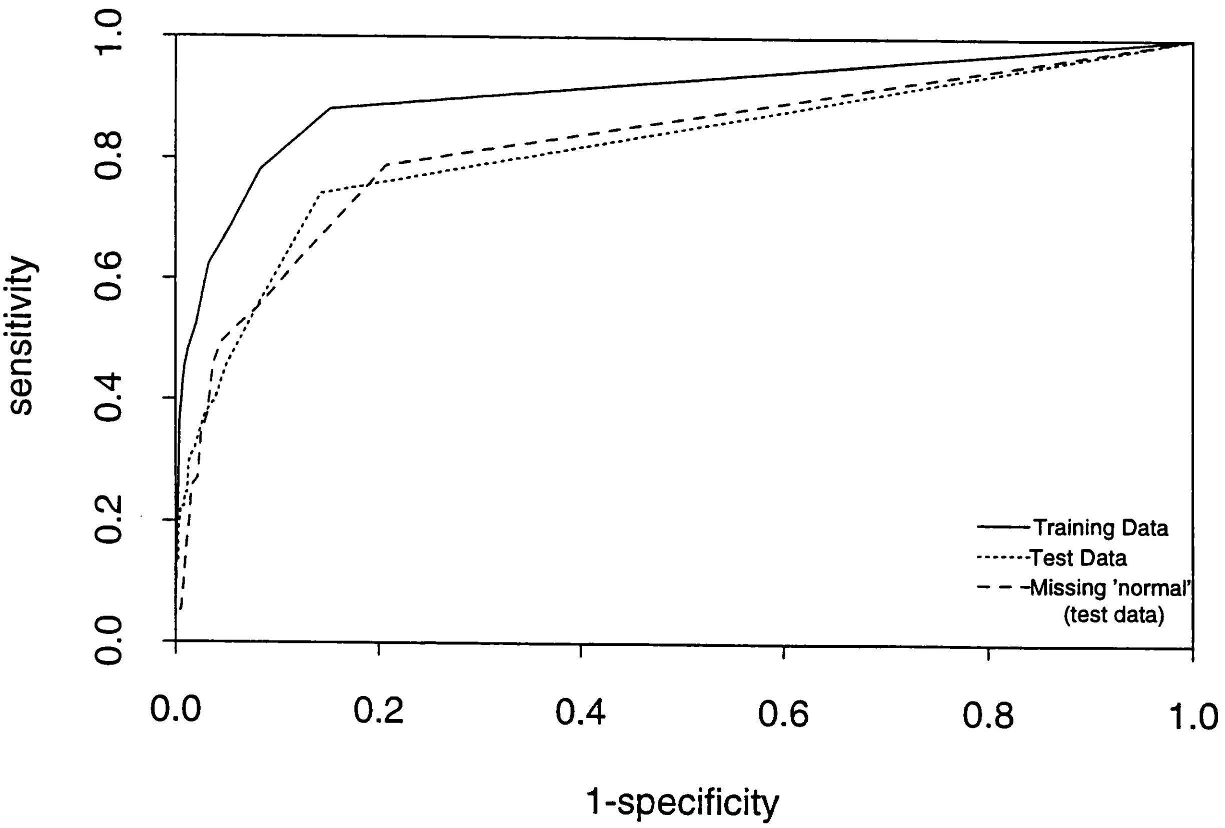


Figure 48: ROC Curves for Models using Variables where "Missing" is a Separate Category

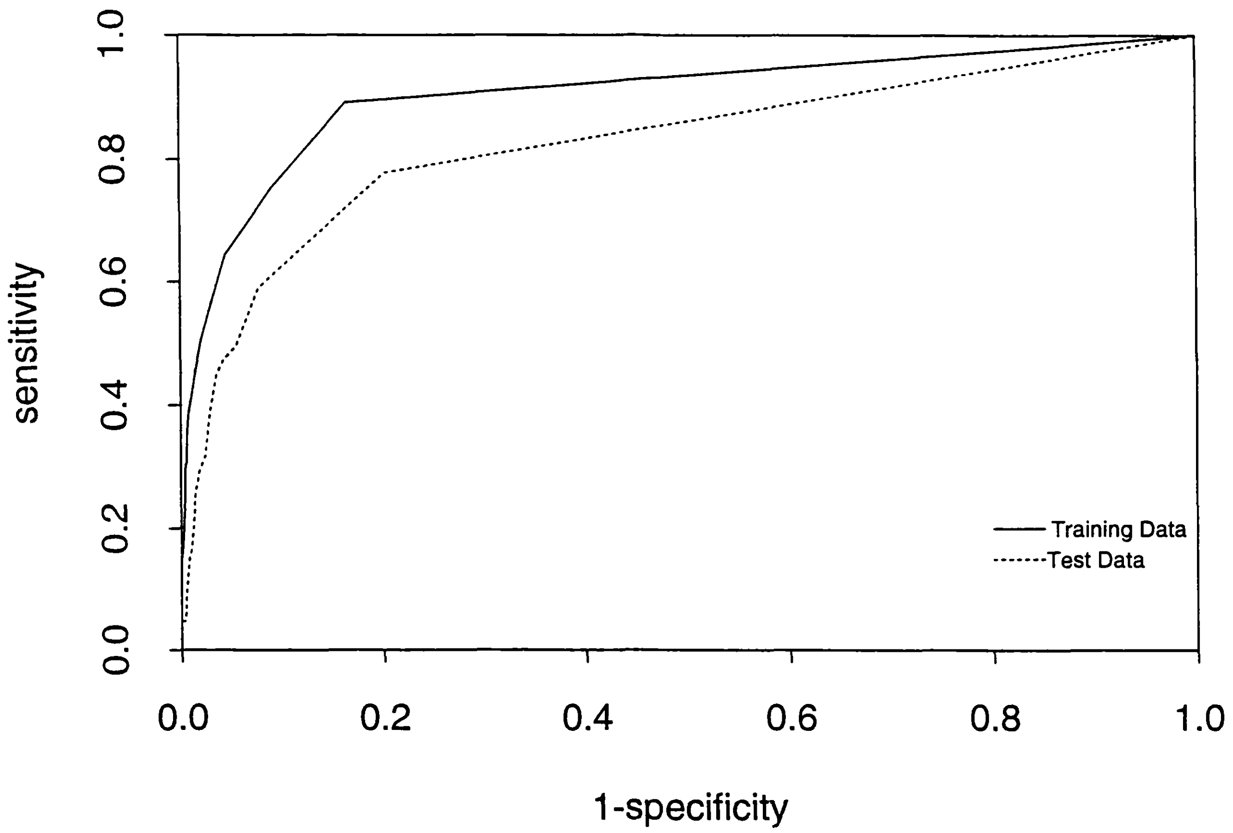


Figure 49(a): ROC Curves for Model containing 11 Variables plus a Missing Value Indicator Variable

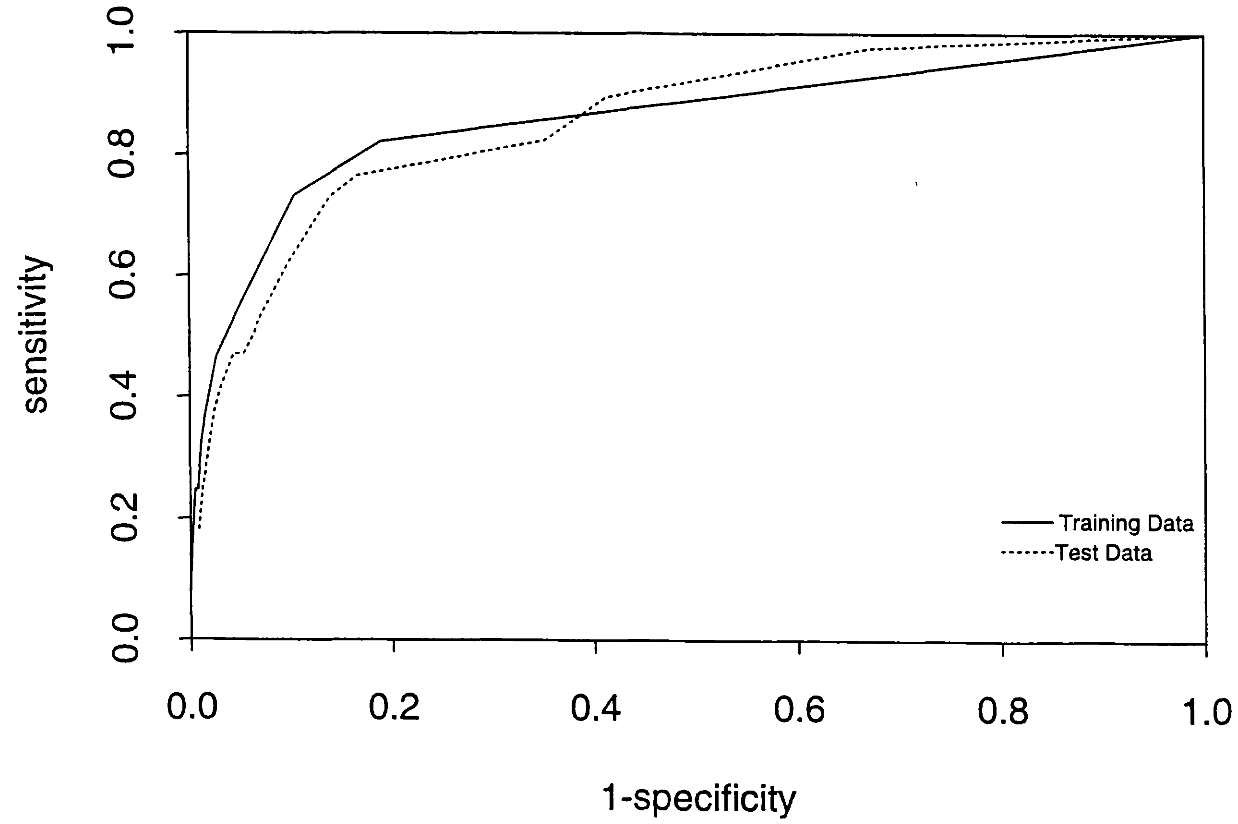


Figure 49(b): ROC Curves for Model containing 4 Experimental Scores plus a Missing Value Indicator Variable

8.5.1.2 INCLUDE A VARIABLE ON COMPLETENESS OF THE DATA

An alternative approach involves assuming all the missing values are "normal" as described previously, but including an indicator variable showing whether data were missing. A stepwise regression was carried out using an indicator variable which took the value 1 if any of the variables considered for inclusion in the model were missing (except for chest X-rays and ECG). There were 1012 patients with this variable equal to 1, who had a mortality rate of 4%, compared with 3% for the 2369 patients with complete data. This yielded a similar model to before, with 12 significant variables including the missing value indicator. By far the largest contributions were again made by age, mode of surgery, emergency admission and operative severity. The indicator variable was the eighth most important, with a p-value of 0.004 in the analysis of deviance. The coefficient was 0.78, implying that those with any data missing have approximately double the mortality rate of those who do not, when the other variables are accounted for. If the variable is then modified to include only the patients who have a missing value for those variables included in the model, it is slightly less important with a p-value of 0.01. This time there are 962 patients with missing data, having a mortality rate of 3%. However, with a coefficient of 0.71, we can still see that risk is increased on average for those with missing data when the other factors are included. Considering simply the 4 most important variables, with another indicator variable for whether any of these 4 are missing, we get this indicator to be more significant than if we consider the ones including other physiological measurements. There are only 136 patients with missing data for these four variables, just one of whom has information for mode of surgery. Of these, 9 (7%) died, explaining the high significance of the variable even if it is included in the fuller model. In order to assess the performance of this model we can look at Q values, proportions of variation explained by the model and compare actual and predicted values as before. We find a value of SSTa of 28.6 for the training data set which is 28.6% of SSTy. This is the second highest value we have achieved, after the above model with missing as a separate category for each variable. This corresponds to a very low Q value of 0.0415. Thus again, by taking missing data into account we are improving the predictive ability of the model. However, looking at the test data set, we see that the model does not transfer so well, with a Q value of 0.0517. Also, we can see that the ROC curves in figure 49(a) are rather far apart. However, this could be caused by the fact that we are modelling variables rather than scores. The predictive accuracy can be seen in table 62. We can see that the model is only a slight improvement on the model with 11 variables without the missing indicator.

	Training Data			Test Data		
Predicted probability of death	Number of patients	Number of deaths	Mortality Rate	Number of patients	Number of deaths	Mortality Rate
<0.25%	1237	1	0.08%	1156	2	0.17%
0.25%-0.5%	598	1	0.17%	520	3	0.58%
0.5%-1%	471	4	0.85%	449	5	1.11%
1%-2%	363	2	0.55%	378	7	1.85%
2%-3%	149	4	2.68%	190	7	3.68%
3%-5%	191	13	6.81%	185	5	2.70%
5%-10%	161	11	6.83%	156	6	3.85%
10%-20%	94	14	14.89%	126	10	7.94%
20%-50%	73	21	28.77%	118	23	18.49%
≥50%	44	30	68.18%	55	17	30.91%

Table 62: Predicted and actual mortality rates for model with twelve variables including Indicator Variable for Missing Data

Since the poor predictions made above could be due to the fact that we are considering variables, we should check it using scores. If we fit the indicator variable for missing along with our "favourite" model of the modified POSSUM scores for operative severity and age, mode of surgery and emergency, the variable turns out to be highly significant. The residual deviance of this model is 17.1 less than the one without the indicator variable, and we account for 21.9% of the total variation in the training data set compared to 20.6% before. The value of Q for this model is 0.0452, which is also an improvement. However, if we apply this model to the test data set, we get a Q value of 0.0836, the highest for any model we have seen. However, the ROC curves (Figure 49(b)) for the training and test data sets are closer together than for the variables model, and the area under them is larger than for the model without the missing value indicator (Figure 45(d)). The minimum predicted probability in the test data set with this model is 1.25%, whereas without the indicator variable it is 0.21%. Thus we are gaining sensitivity with more patients predicted to die, but losing specificity. The model assumes too high a mortality rate for missing data, based on the peculiarities of the training data set. The mortality rate for patients with one of these 4 variables missing is 6.6% in the training data set and only 3.8% in the test set. Thus, the generalisation made for missing data using this method does not even transfer to the other half of the same data set, so it would not be reasonable to try to apply it to a different hospital's data.

Similarly, a continuous variable of the number of missing values from all variables which were included as candidates for a model, excluding the chest X-rays and ECG, was included. The number missing ranged from 0 to 17. The numbers having each value missing can be seen in Table 59. This turned out to be highly significant, with the third most significant coefficient after age and mode of surgery, and a p-value of 0.0001 after all the other variables had been accounted for. When a variable ranging from 0 to 10 was used for the 11 variables remaining in the model, it was still highly significant (p=0.001). The model achieved from logistic regression had a residual Deviance of 506.5. This is not significantly better than the model achieved using an indicator variable. In fact the models are practically the same, with the same factors being entered and a difference in residual deviance between them of only 0.25. Thus we are not gaining any information by using the count of number missing rather than the indicator variable.

Thus patients with missing data tend to be different from those without, and in this data set they have a higher risk of postoperative mortality in general than those with full data. The effect obviously depends on what variables are included in the model, and what variables are included as missing. For example, due to the high mortality rate of patients whose status with respect to malignancy is unknown, the inclusion of this variable makes the missing indicator more significant (p = 0.001 *c.f.* p = 0.01). Including missing variables separately will not give a model that transfers well to different situations.

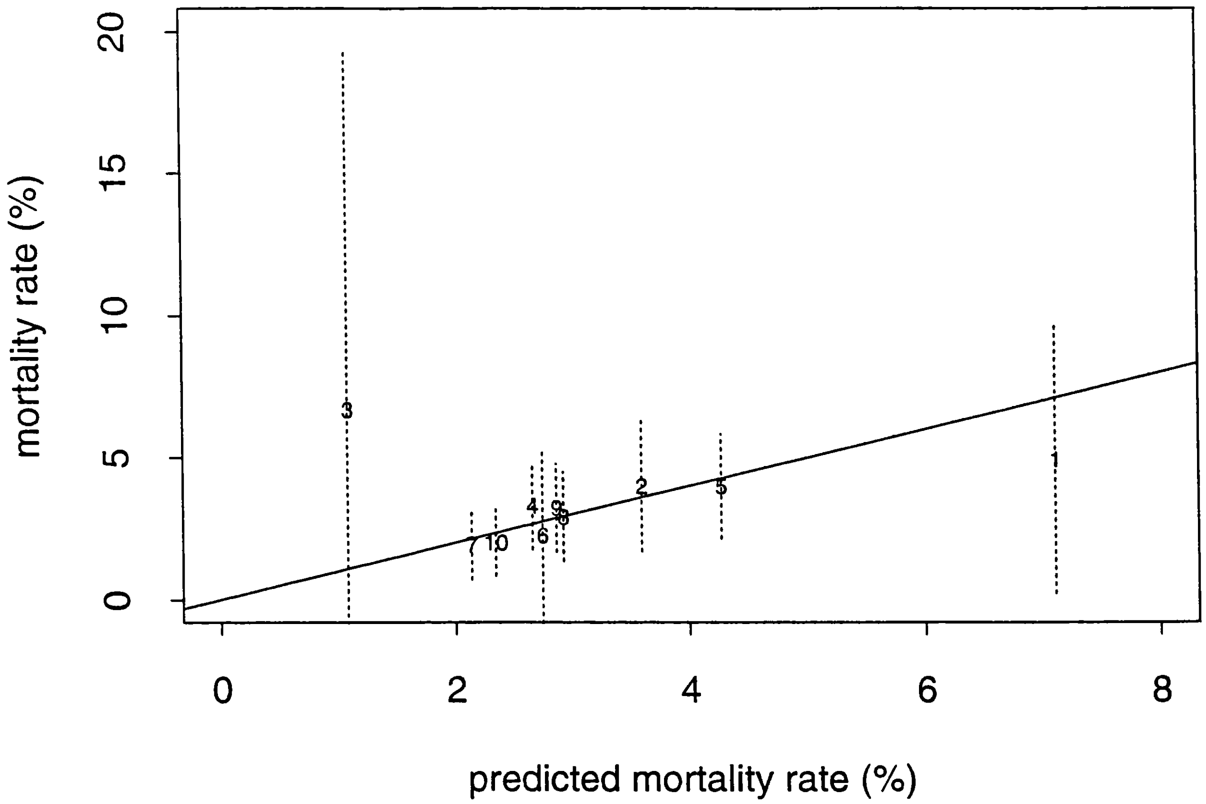
The question is, should a model for audit include a term for missing data? It could be argued that missing data is one symptom of poor care, and as such should not be accounted for. Because it is an unknown, it may be very foolish to try and generalise about the behaviour of missing values from one person to the next. But, since these missing terms are significant, and so missing patients are not the same as "normal" ones when it comes to outcome (except perhaps in the biochemical measurements), any model including these patients as "normal" will have higher expected mortality rates than should be the case. Perhaps the only answer is to work to encourage medical workers to collect fully comprehensive data. This would be facilitated by a reduction in the amount of data required, for example by using the "best" model here which contains only four, very easily elicited scores. It is worth noticing that using this model greatly reduces the problem in itself. For the four variables in the model there are 136 patients with data missing, whereas for the POSSUM model there are 1822.

9 COMPARISONS BETWEEN SURGEONS IN THE POSSUM STUDY

In chapter 8, we explored a large surgical data set, and arrived at a "favourite" model. To investigate the usefulness of this model, and compare its effectiveness with the other best candidates, we will compare predicted and observed mortality rates for consultants. We will then explore the effect of adjusting individual consultants' relative risks. The models chosen to do this are:

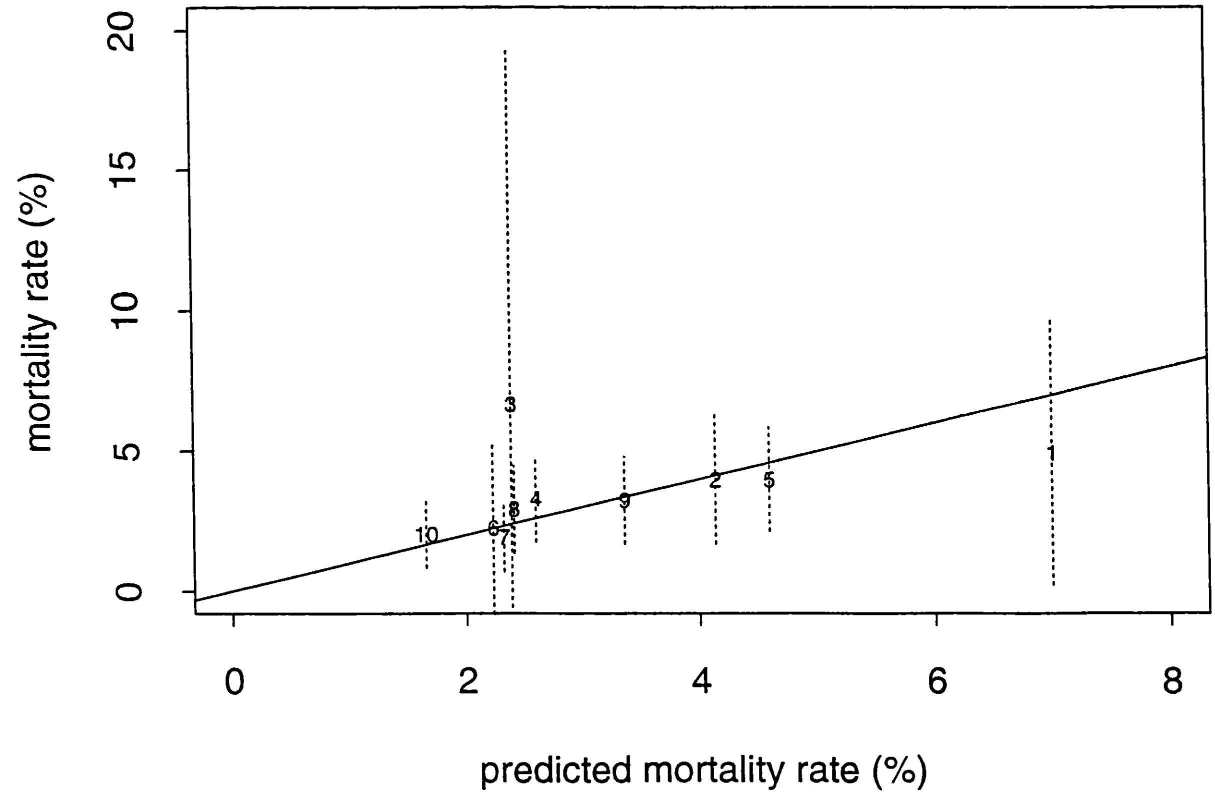
(1) Derived from the POSSUM Physiological and Operative Severity Scores
(2) The "favourite" model, as described in the previous chapter
(3) The model containing 11 experimental weights, which was discarded in favour of (2)

Firstly the predicted mortality rates have been plotted against the observed ones for each consultant. These can be seen in figures 50 and 51 for the training and test data sets respectively. Note that the outlier, consultant 3, had only 14 patients, so any estimates are not likely to be accurate, and the small number of patients with only 1 death gives a misleadingly high observed rate. The other consultants tend to lie close to the line of equality for all three models. This is, as would be expected, more evident in the training data set. In this data set, consultant 1 is consistently predicted to have more deaths than he actually did. This could mean that he is performing better, or that the models are missing something. Note that the rank order of the consultants' observed mortality rates is not the same in each of the data sets, although 7 and 10 have the lowest rates each time, and 1 and 3 the highest. The change in positions of the other six consultants suggests that there is no real difference between them, and that any differences seen are caused by random variation. In the test data set, we can see that the predicted mortality rate for consultant 3 is close to the observed only with the POSSUM scoring model. The two models containing weights make almost identical predictions, despite one of them having seven more included.



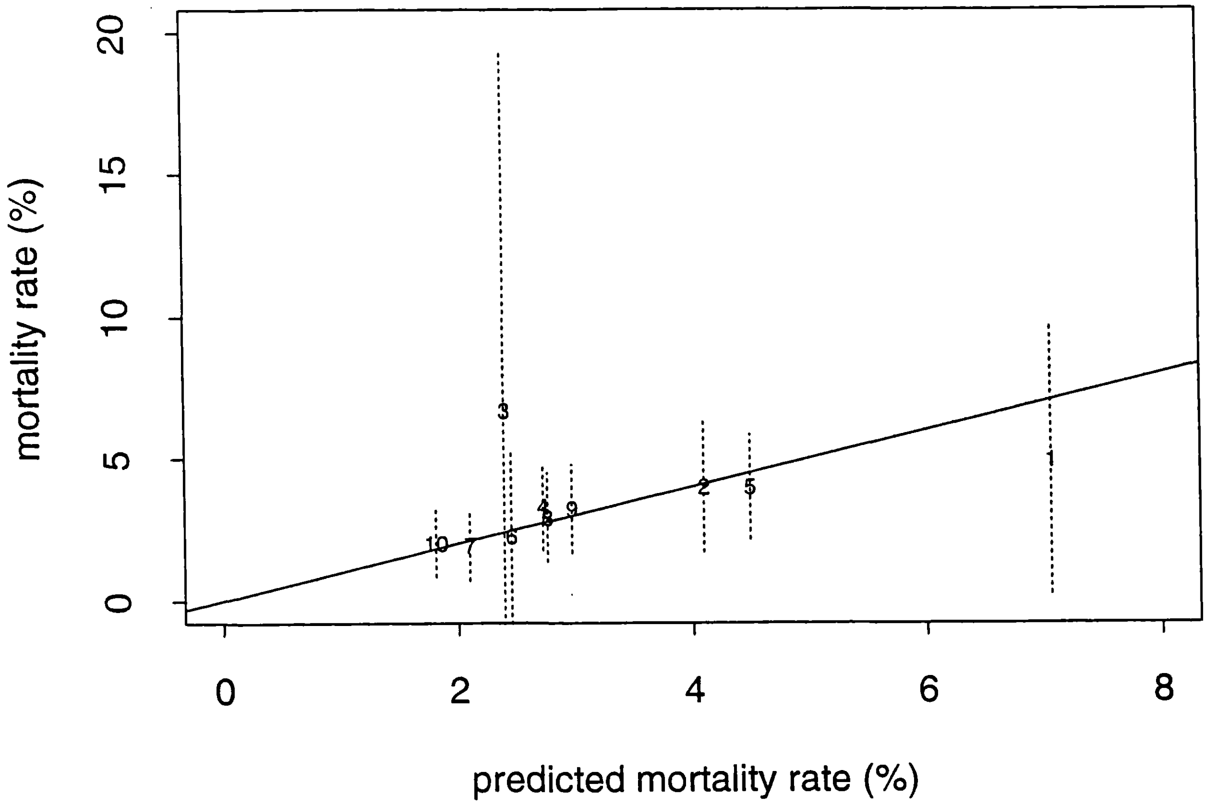
Dotted lines show 95% Confidence Intervals for Mortality Rates

Figure 50(a): Consultants' Actual Mortality Rates *versus* expected for Model calculated from POSSUM Scores (Training Data)



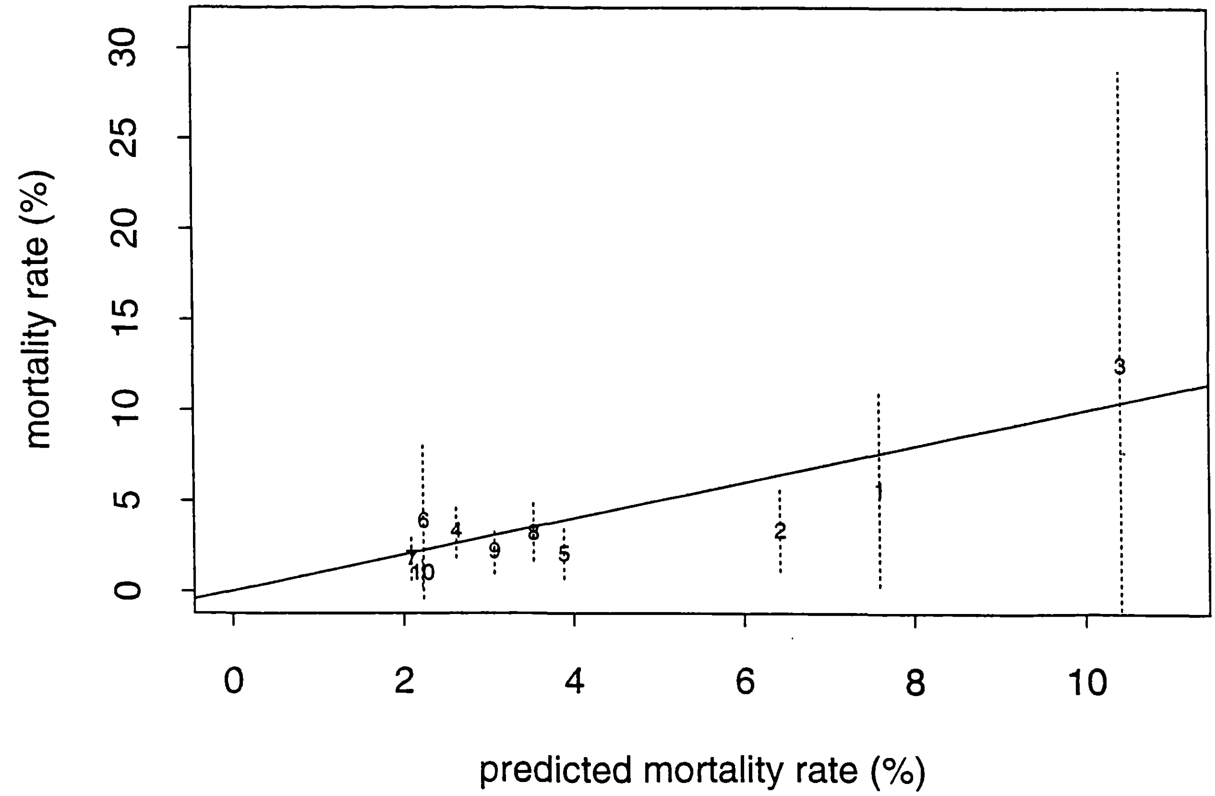
Dotted lines show 95% Confidence Intervals for Mortality Rates

Figure 50(b): Consultants' Actual Mortality Rates *versus* expected for Model calculated from 11 Experimental Weights (Training Data)



Dotted lines show 95% Confidence Intervals for Mortality Rates

Figure 50(c): Consultants' Actual Mortality Rates *versus* expected for "Best" Model (Training Data)



Dotted lines show 95% Confidence Intervals for Mortality Rates

Figure 51(a): Consultants' Actual Mortality Rates *versus* expected for Model calculated from POSSUM Scores (Test Data)

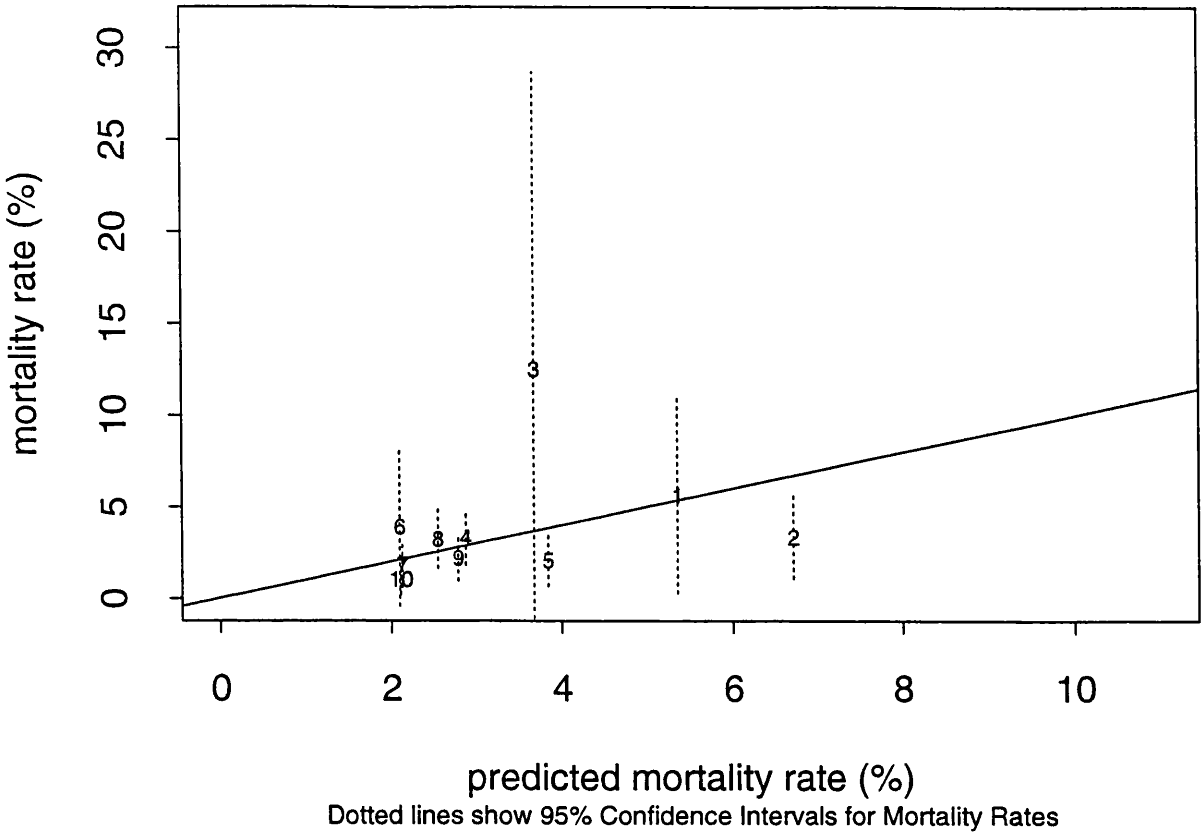


Figure 51(b): Consultants' Actual Mortality Rates *versus* expected for Model with 11 Experimental Weights (Test Data)

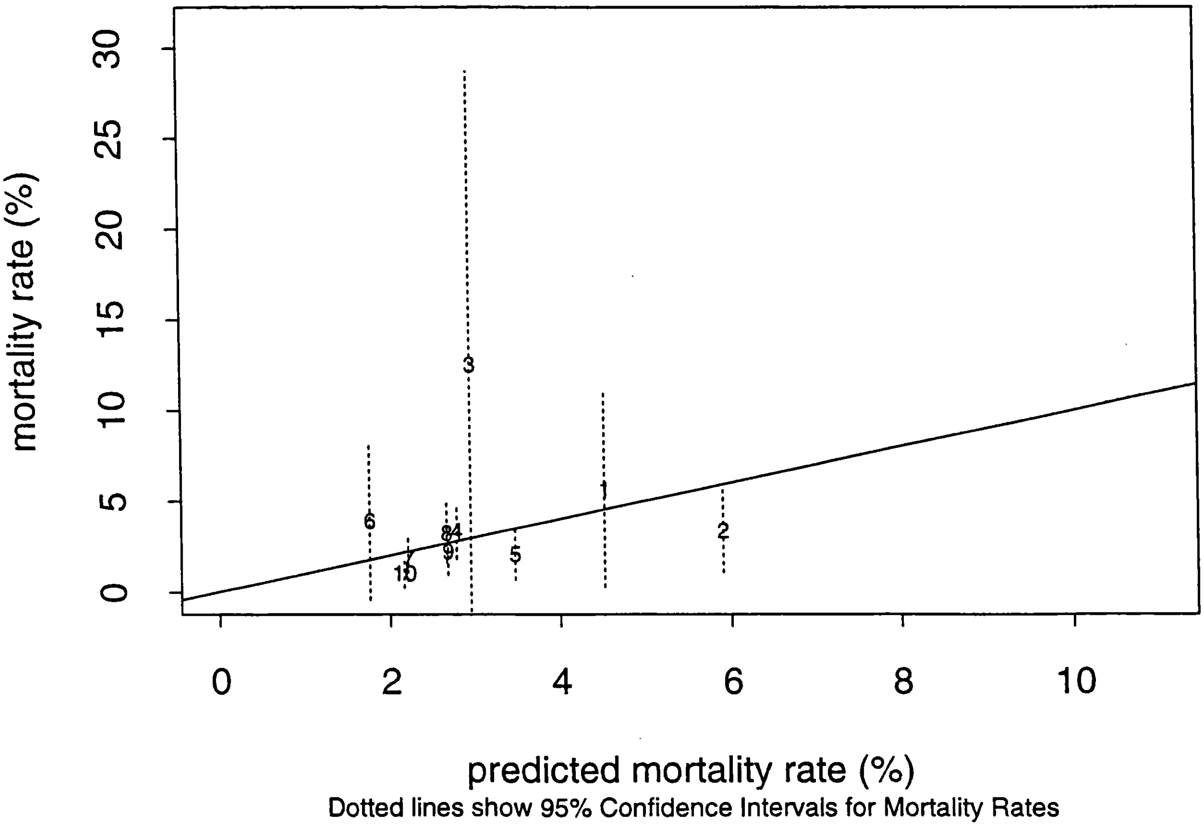


Figure 51(c): Consultants' Actual Mortality Rates *versus* expected for "Best" Model (Test Data)

The next stage is to calculate relative risk confidence intervals as has been done in previous chapters. The easiest way to do this using patient data is to fit a logistic regression model with an indicator variable separately for each consultant. The unadjusted risks are the coefficients obtained by fitting the indicator variable alone, and the adjusted ones are the coefficients of the variable when the relevant other variables are included. The models were fitted in Splus, and the coefficients extracted. Since the coefficients are determined when fitting the model, we cannot use this method for the test data set, as this will mean fitting a different model. The ordered unadjusted intervals can be seen in Figure 52. It can be seen that they all include unity, so there is no evidence of a significant difference between the mortality rates. Recall that these are individual interval estimates, so simultaneous ones would be even wider. The adjusted confidence intervals using the models as described above can be seen in figures 53(a) to (c). The different models tend to have very similar effects on the intervals, and none of them move so that they do not straddle the line. We require data on more consultants in order to fully see the effect of adjustments. The ranks on adjustment for each model are given in table 63.

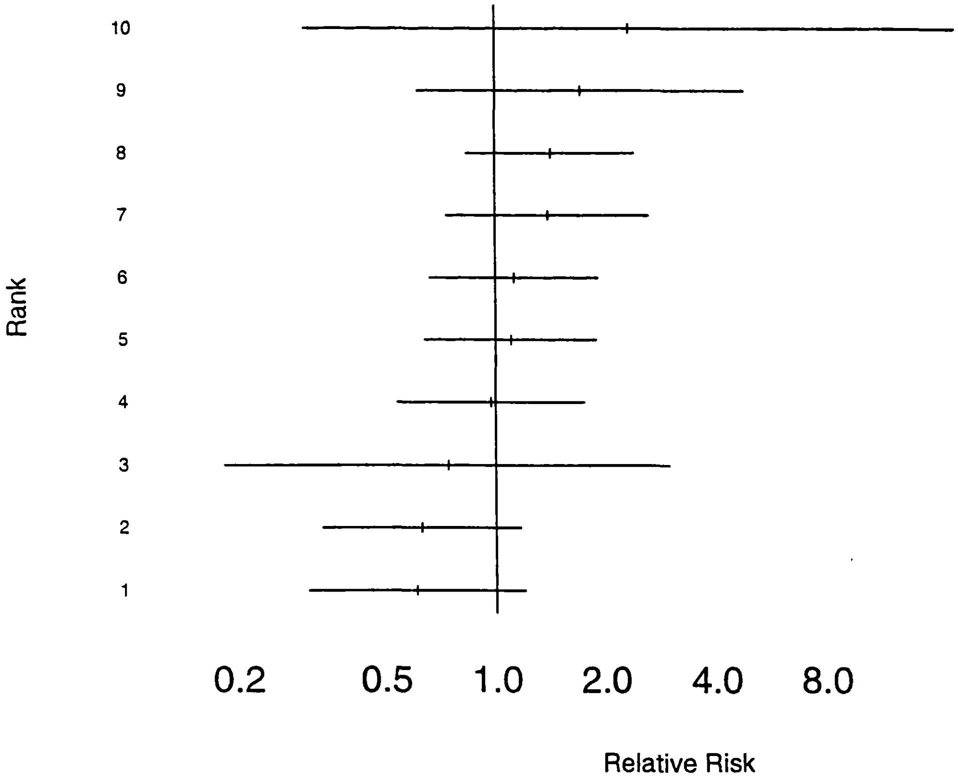


Figure 52: Unadjusted Relative Risk Confidence Intervals for Consultants in POSSUM Study (Training Data)

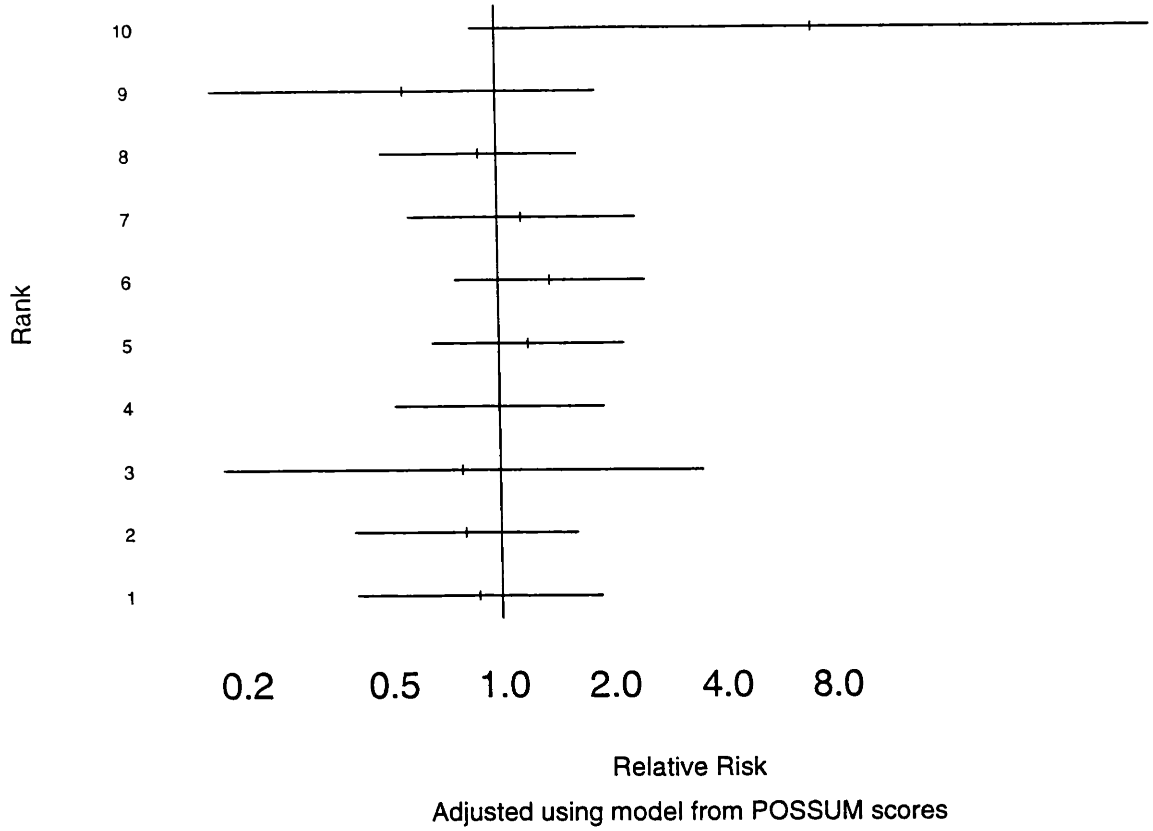


Figure 53(a): Adjusted Relative Risk Confidence Intervals for Consultants in POSSUM Study (Training Data)

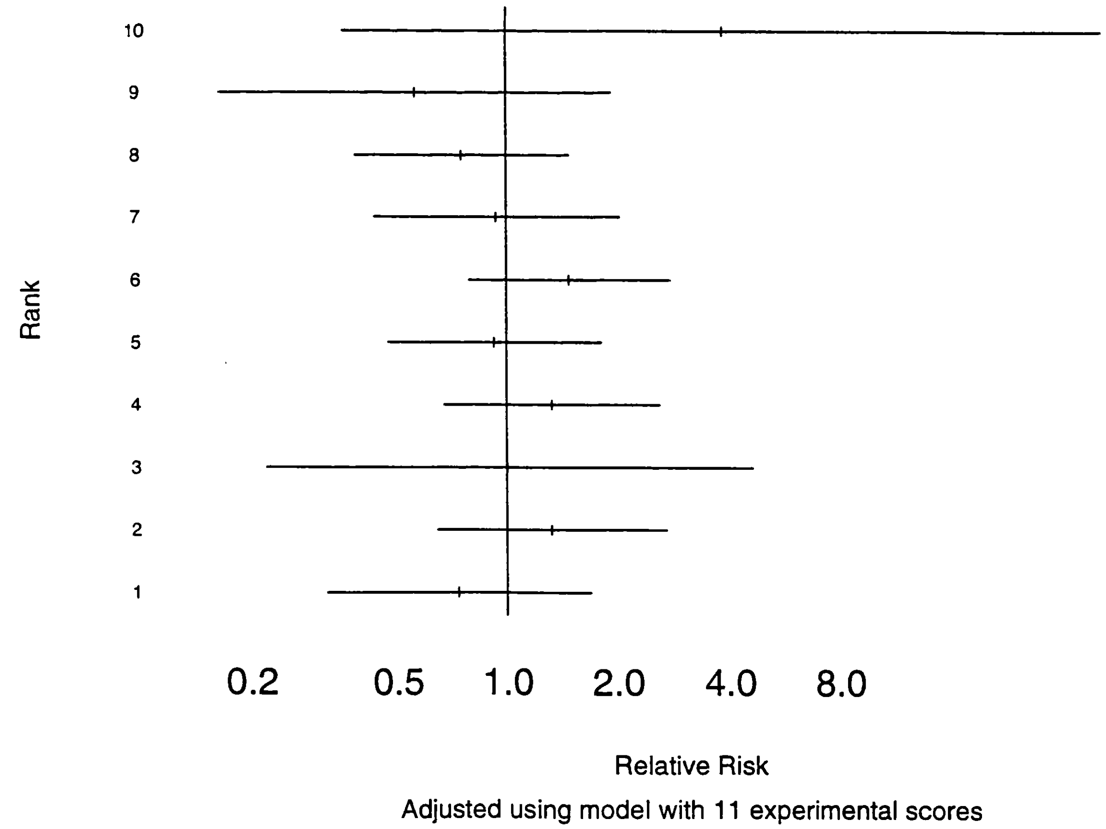


Figure 53(b): Adjusted Relative Risk Confidence Intervals for Consultants in POSSUM Study (Training Data)

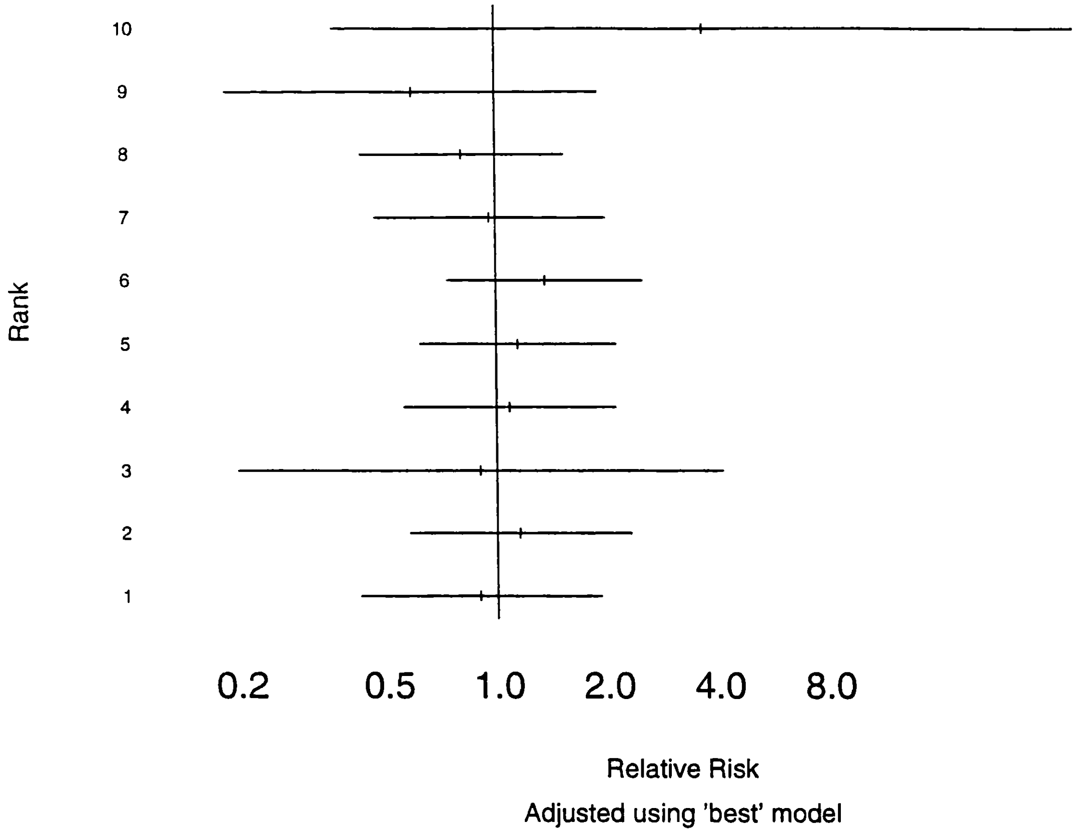


Figure 53(c): Adjusted Relative Risk Confidence Intervals for Consultants in POSSUM Study (Training Data)

	Rank after Adjustment			
Unadjusted Rank	**Adjusted by Model 1**	**Adjusted by Model 2**	**Adjusted by Model 3**	**Aggregate Data**
1	4	2	3	3
2	3	8	8	4
3	2	6	4	2
4	6	7	6	5
5	8	4	7	8
6	9	9	9	7
7	7	5	5	6
8	5	3	2	9
9	1	1	1	1
10	10	10	10	10

Table 63: Consultants' Rankings

Note that, although there are appear to be no significant differences, all the models move the consultant ranked ninth (consultant 1 in the previous plots) to first place due to his high predicted mortality. It is not surprising that there is variation among the rank orders of the consultants in the middle, as there is no evidence of any real difference between them.

9.1 Aggregate Data

The data were aggregated for each consultant and the proportion having each condition were calculated and put up for selection in a stepwise regression. If the operative severity categories were included separately, the only factor to enter the model was the proportion of intermediate procedures, which had a highly negative coefficient. If the operative severity categories were then grouped into "low risk" (minor and intermediate), "medium risk" (major) and "high risk" (major plus and complex major), the only factor to enter is emergency surgery. The model using this was used to adjust. With so few consultants this method does not work too well, but it is interesting to note that the adjusted relative risk of consultant 1 still came out first, although the model was rather weak. We can see from Figure 54 that many of the adjustments using aggregate data do not have much effect.

It would be of interest to adjust the aggregate data using a model similar to the one calculated from patient data. However, if we tried to fit four variables with only 10 consultants, the model is over-specified and the confidence intervals would be ridiculously wide.

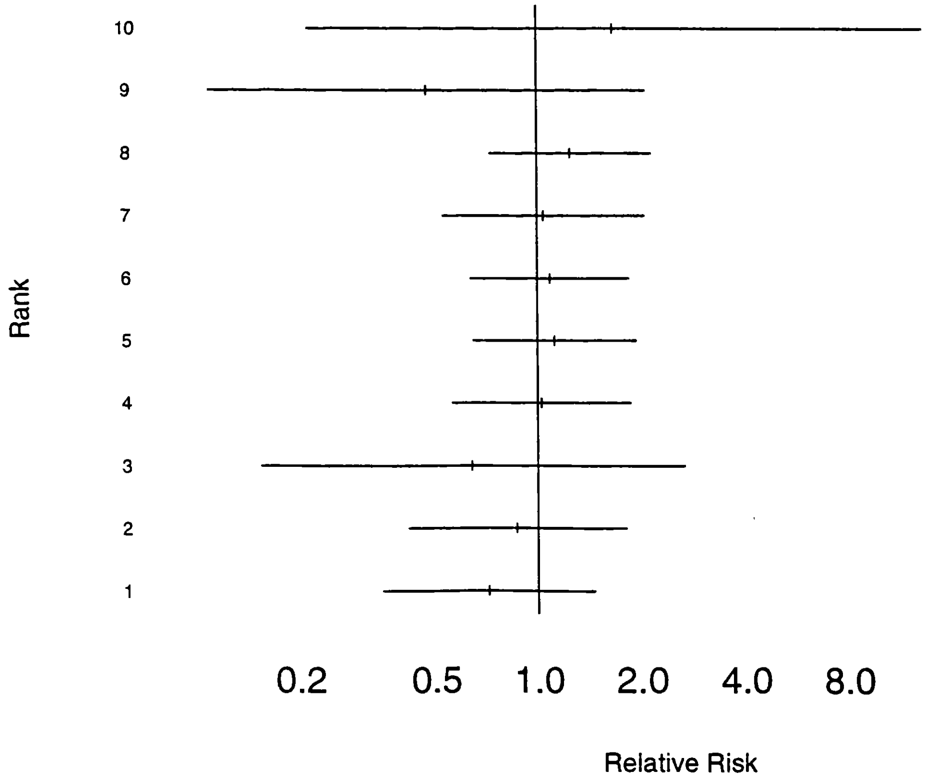


Figure 54: Adjusted Relative Risk Confidence Intervals for Consultants in POSSUM Study (Aggregate Data)

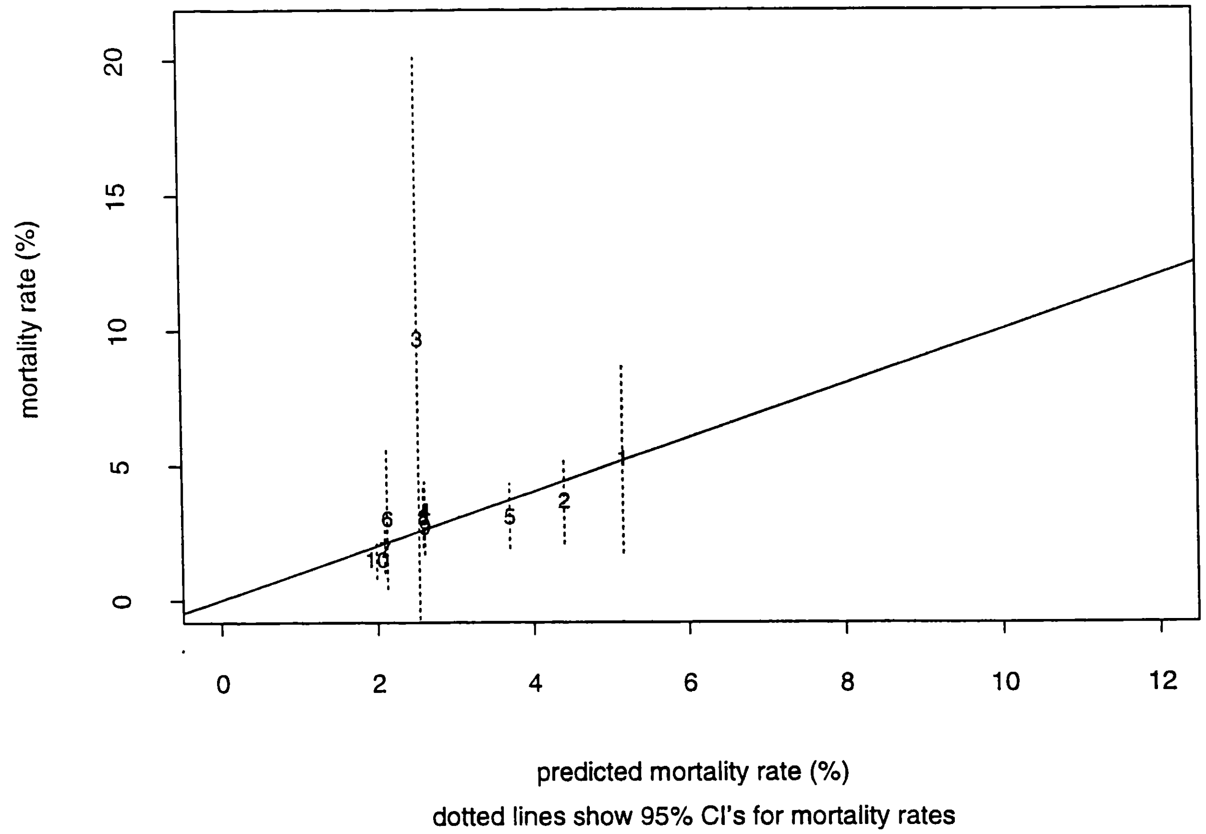


Figure 55: Consultants' Observed Mortality Rates *versus* Rates Predicted using "Best" Model (Whole Data Set)

9.2 Using all the data

In order to gain a definitive model, we put together the training and test data sets and carried out a logistic regression including the scores in the best model with all the data. The model is as follows:

$$\ln\{R/(1-R)\} = 0.25 \times \text{Operative Severity Score} + 0.76 \times \text{Age Score} + 0.26 \times \text{Mode Surgery Score} + 0.19 \times \text{Admission Score} - 7.46.$$

This model is very similar to the one achieved from the training data set alone, and so gives us more confidence in the model. The predicted rates using this model are plotted against the actual rates in figure 55. Again we see that consultant 3 has his estimate too low. But the 95% confidence interval for his mortality rate is very wide due to the fact that he only treated 31 patients, and it does include the predicted mortality rate. This model accounts for 16.6% of the total variation between patients, and has an average quadratic score of 0.0453. We can also calculate the amount of variation between consultants which is accounted for by the model from

SSBa = SSBy - SSBd, where SSBy = $\sum_i n_i(\bar{y}_i - \bar{y})^2$, and SSBd = $\sum_i n_i(\bar{y}_i - \bar{p}_i)^2$.

(Recall y_{ij} is the observed mortality for patient j of consultant i, and p_{ij} is the predicted value for that patient from the model). Then SSBa accounts for 43.7% of SSBy. If, however, we omit consultant 3 from this calculation, the amount of variation between consultants which is accounted for by the model increases to 60.7%. This suggests that this method should not be used for consultants with very low numbers of patients, as the results are unreliable. Our model predicts that consultant 3 should have 0.8 deaths when in fact he has 3 out of 31. We cannot know whether he is actually performing worse than expected, or if the model is not sensitive enough to predict such a rate from a small number of patients.

We can now use the above model to find adjusted confidence intervals for the consultants using all 6714 patients. These can be seen in figure 56(a). The unadjusted intervals are plotted in figure 56(b). With the whole data set, the consultant ranked 1 (consultant 10)'s unadjusted relative risk is significantly less than the others', but on adjustment moves to straddle the line of unity.

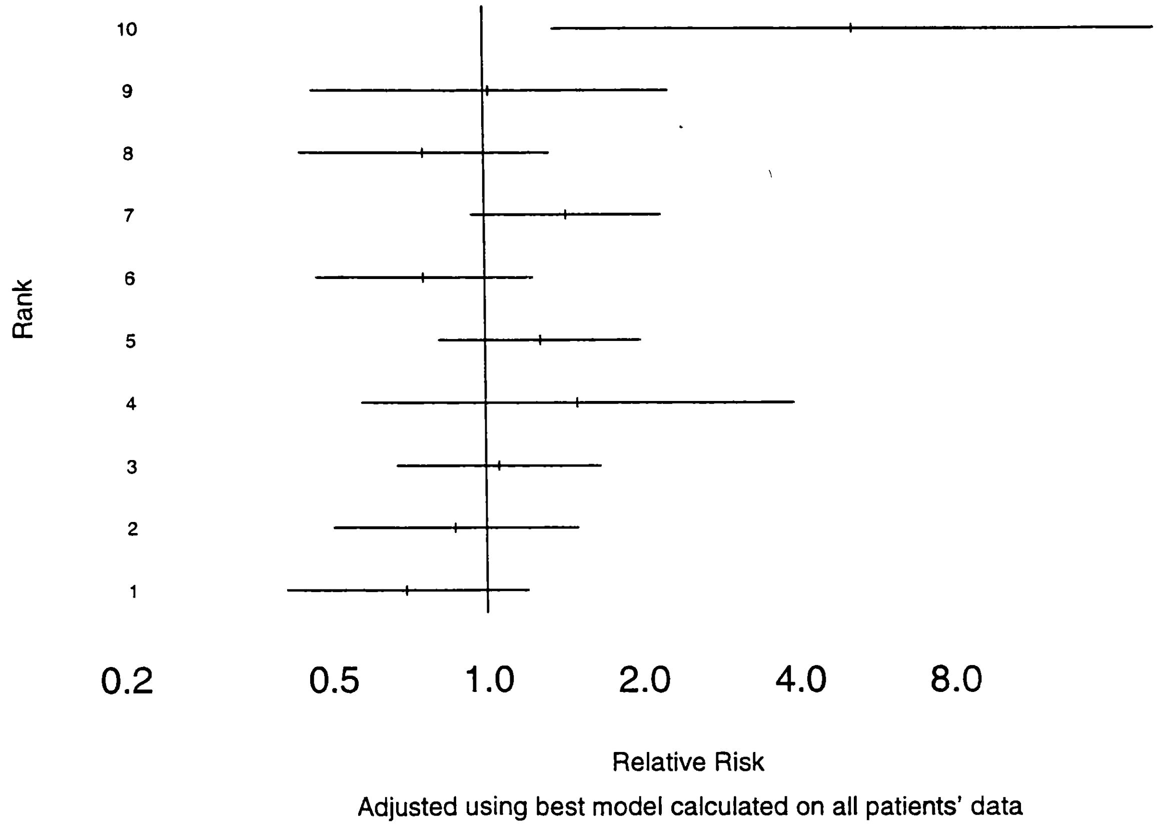


Figure 56(a): Adjusted Relative Risk Confidence Intervals for Consultants in POSSUM Study (Whole Data Set)

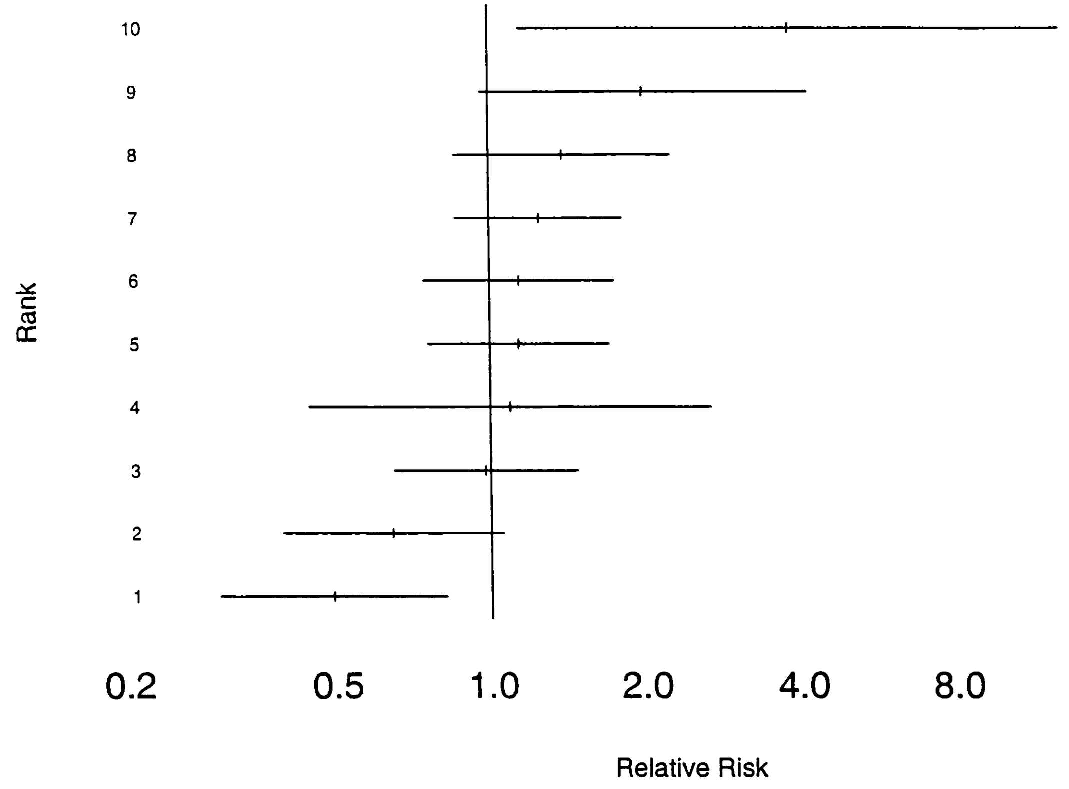


Figure 56(b): Unadjusted Relative Risk Confidence Intervals for Consultants in POSSUM Study (Whole Data Set)

If we fit a categorical variable "consultant" along with the variables in the model, we find that it is not significant (p=0.6). Thus, considering the multiple comparisons aspect, there is no evidence of a difference between the surgeons in this study, when other factors are taken into account. In fact, if we fit this variable alone we still have no evidence of a difference between consultants. Thus we must conclude that the failure of the confidence interval for consultant 3 to include 1 is due to chance.

	Adjusted Rank	
Unadjusted Rank	Best Model	POSSUM Model
1	1	1
2	4	5
3	6	6
4	9	8
5	7	7
6	2	3
7	8	9
8	3	2
9	5	4
10	10	10

Table 64: Consultant Rankings from all patient data

If we compare the results of this model calculated from all the data, from one calculated from all the POSSUM scores (equivalent to model 1 above), they are favourable. The model we achieve for the POSSUM scores is

$$ln\{r/(1-r)\} = 0.15\times\text{Physiological Score} + 0.13\times\text{Operative Severity Score} - 8.34$$

The Q value from this model when used to predict outcomes from all the data is 0.0477, and it accounts for 11.4% of the variation between patients. From these statistics, it would seem that our model is superior. However, if we look at the variability between consultants, this score model accounts for 58.1%. This is because it predicts a higher mortality rate for consultant 3 than our model. If consultant 3 is ignored, this model explains 58.6% of the variation between the other consultants compared to 60.7% by the 'best' model. For consultant 3, the POSSUM model predicted 1.6 deaths. This is because it consistently estimates higher probabilities of death.

In table 60, we can see that the models do not make a substantial difference to the rankings. Looking at the Confidence Interval plots in figure 56 however, it can be seen that the interval for the consultant ranked 10 (consultant 3) moves to include 1 when adjusted using the POSSUM model, but not with the new model. Similarly, the consultant ranked 7 (consultant 4) moves entirely above 1 with the POSSUM model, but not with the new one. Thus we would come to differing conclusions depending on the model used. However, looking more closely we can see that consultant 4 has a predicted mortality rate of 0.026 from the new model and of 0.024 from the POSSUM model. This small difference with a large number of patients makes a substantial difference to the position of the confidence interval.

9.3 Conclusions

While none of the models we have found are accurate at patient level, overall they tend to perform satisfactorily. A small set of variables is better for prediction on a new data set than many variables, so it is hoped that the need for less data will facilitate more accurate collection, thus removing the missing values problem. Missing values are a great problem, in that their presence is a significant predictor of mortality, but accounting for them in a model reduces the transferability of the model due to differences in collection mechanisms. It was found that scores provide more robust models than using actual variable measurements. The POSSUM Operative Severity and Physiological Scores perform well, but involve a large amount of often unavailable information. Changing some of the scores so that they tie in with the observed mortality rates improves the model. The model with the altered scores for age and operative severity, the POSSUM score for mode of surgery, and a score for emergency admission performed almost as well as one with many more variables included, and was chosen as the 'best model'. This has been used to compare consultants. Using the POSSUM score model on all the patients gives slightly different conclusions. Our model does not perform well with few patients. Overall, however, it performs similarly to the POSSUM model for consultants, although not for patients. This is illustrated in figures 57(a) and 57(b), where the patient predictions do not correspond at all for the two models, but on average for the consultants they correspond very well. As Smith (1994) said, if a model existed which could accurately predict outcome for a single patient, the way medicine is carried out would be completely different, as prediction of death in an individual is very difficult, even with highly detailed information.

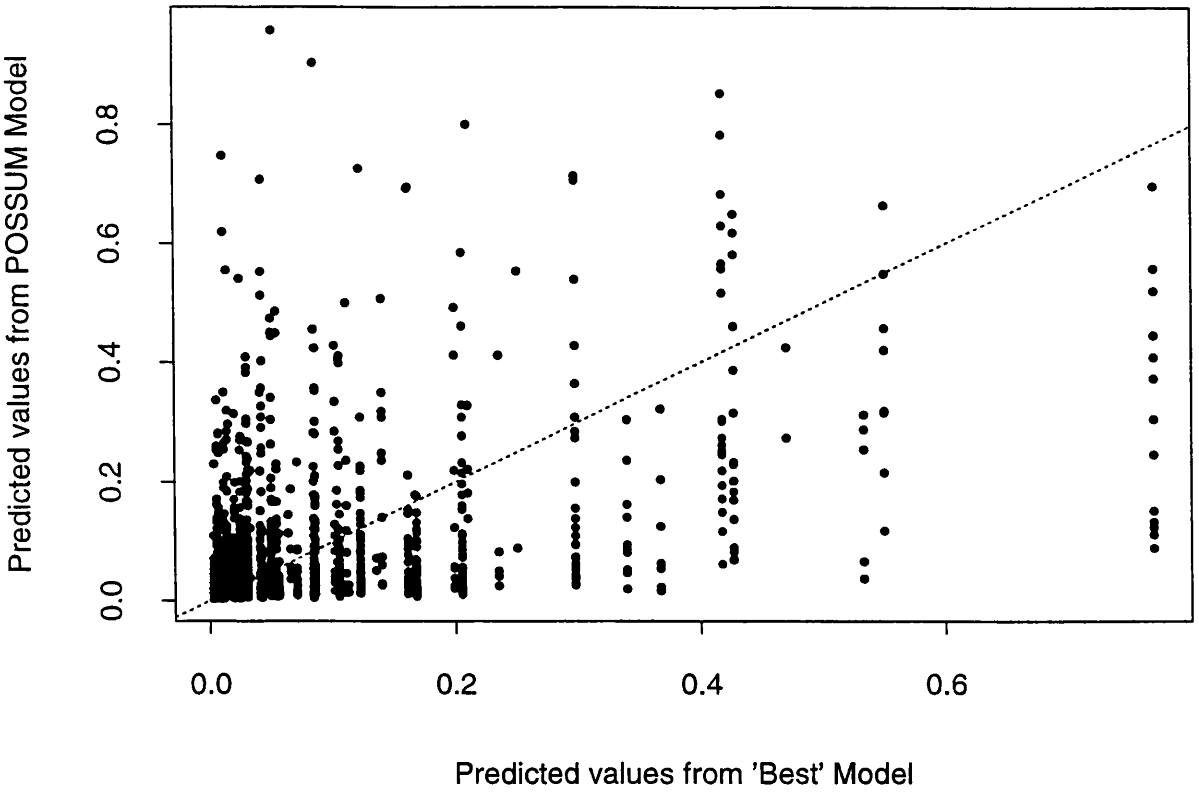


Figure 57: Comparison of Patients' Predicted Probabilities of Mortality from Two Models (All Data)

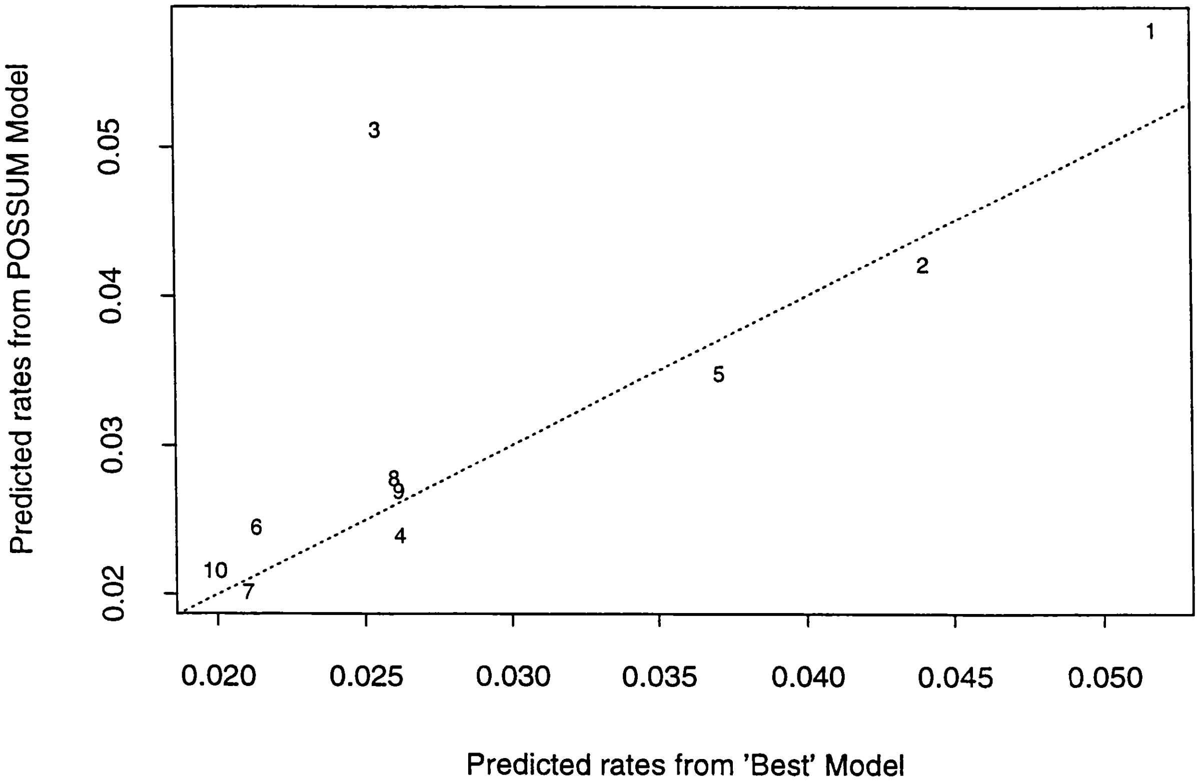


Figure 58: Comparison of Consultants' Predicted Mortality Rates from Two Models (All Data)

10 COMPARISONS USING AUDIT DATA ON A PARTICULAR DIAGNOSIS

10.1 Introduction

Up until now, we have been concentrating on modelling General Surgery. This is an area which has received much attention, and where work needs to be done. However, while we have obtained some reasonable results, it may be that we will never achieve very powerful predictive systems due to the vary wide range of patients being considered. Overall at a consultant level the models are good, but at a patient level, predictions are very inaccurate. We have used the very rare outcome measure of postoperative mortality as there are no other reliable measurements in data sets which we have been able to acquire. It may be that the way forward for surgical audit is to consider one diagnosis or procedure at a time, and compare consultants' performance in these. The advantages of doing this are that there will be more specific prognostic variables, and more relevant outcome measures. In this chapter, then, we consider a large audit of colorectal cancer patients, and compare the consultants involved in this study.

10.2 The CRAG Study

The CRAG Study was set up in Lothian and the West of Scotland, as a large scale continuation of the Colorectal Cancer Study described in chapter 6, after large differences were found between participating consultants' survival. We obtained data from the Study, which covered admissions for colorectal cancer from 1991 until 1994. There were 1622 patients from 5 Lothian hospitals and 1807 patients from 8 hospitals in the West of Scotland. These included data on 76 consultants.

The variables from these two area studies which were considered to be important were put together in one data file. These included the variable "class" which we derived. This was an indicator of the type of operation combined with severity of the cancer, derived from available data. It had six categories. These are shown in table 61, with the numbers of patients in each. Dukes' stage was either given, or based upon pathological variables such as the presence of metastatic spread or fixity. Any resected patient with Dukes' D was considered to have received a palliative resection.

		Numbers in Class	
Class	Description	Lothian	West
0	Curative - Dukes' Stage Unknown	10	106
1	Curative - Dukes' A	159	58
2	Curative - Dukes' B	577	620
3	Curative - Dukes' C	356	358
4	Palliative Resection	489	387
5	Palliative- other treatment	29	74
6	No Operation	2	204
		1622	1807

Table 65: Numbers of patients in each "class"

For our analyses, we have used only patients with resections and whose pathology is known, i.e. classes 1 to 4. This gave a total of 3004 patients, 1581 from Lothian and 1423 from West Scotland.

The other variables considered were age, sex, presentation (elective or emergency), site, surgeon's status, city, hospital and consultant. Unfortunately, there were no available data on social class. The outcome measures were mortality within 30 days of operation and the presence of any leak or abscess.

10.3 Analysis

10.3.1Modelling

Logistic Regression was carried out for 30 day mortality, and for any poor outcome (leak, abscess or death) to investigate which variables were significant. In the original study of 645 patients by McArdle and Hole (1991), the overall postoperative mortality rate was 16%, and varied among surgeons from 8% to 30%. This, however, included those patients with no resection. Including only those with resections, the rates varied from 0% to 20%. In this data set, the postoperative mortality rate varied among consultants from 0% to 24%. We have modelled postoperative mortality for the McArdle and Hole data in chapter 6 when looking at aggregate data. In this study, the variables Dukes' Stage, differentiation and local spread have been combined in the variable "class" along with type of resection.

10.3.1.1 POST-OPERATIVE MORTALITY

Stepwise Logistic Regression for postoperative mortality arrived at a model containing presentation (emergency versus elective), class, age and sex (males being higher risk), although sex only entered with a p value of 0.08. It was found that a categorical variable of whether the patient was over 75 gave a better fit than age as a continuous variable. Presentation was the most important factor. It was also found that, once the other factors are accounted for, there is no significant difference between classes 1, 2 and 3. Thus this factor was reduced to two classes, curative and palliative resection. Of the 2990 patients with data on these factors, there were 164 deaths, a rate of 5%.

10.3.1.2 ALL POOR OUTCOMES

For any adverse outcome, which includes postoperative death, and any leak or abscess, there were more significant variables. There were 270 poor outcomes (9%). The most important predictor of this was again presentation, followed by age over 75. The next variable to be entered was whether or not the site of the tumour was the colon. This reflects the fact that the colon is lower risk than the rectum for anastomotic leakage. Sex and palliative resection were also included in the model, with all factors being highly significant. The terms in the model are given in the Table 66.

Term	Coefficient	Standard Error
Presentation	0.738	0.13
Sex (female = 1)	-0.413	0.13
Over 75	0.549	0.13
Colon	-0.477	0.14
Palliative	0.325	0.14
Constant	-2.543	0.13

Table 66: Coefficients in the Logistic Regression Model for all Poor Outcomes

In our comparisons, we shall consider this variable, as we have obtained a more satisfactory model, and poor outcome is more appropriate as an outcome measure for comparing surgical success.

10.3.2 Comparisons of hospitals

In order to compare the thirteen hospitals in the study, relative risks of poor outcome for each hospital have been calculated. Hospital 5 (Lothian) is excluded as it had only 8 patients, with no poor outcomes. These were calculated by fitting binary variables for each hospital in separate logistic regressions, on their own to obtain the unadjusted values and with the above explanatory variables to obtain the adjusted ones. The confidence intervals can be seen plotted in Figures 59 and 60. These show that the adjustments have made very little difference to the rank order of the hospitals, and seem to show that the hospital ranked twelfth is performing worse than the rest. We must, however take the fact that these are individual interval estimates into account. If we fit "hospital" as a categorical variable in a logistic regression along with the other variables, it is not significant in the analysis of deviance. It is therefore possible that the difference is due to chance. Otherwise there could either have been some unaccounted for factor causing this hospital to have almost double the average failure rate of the other hospitals, or it could actually be administering a poorer quality of care.

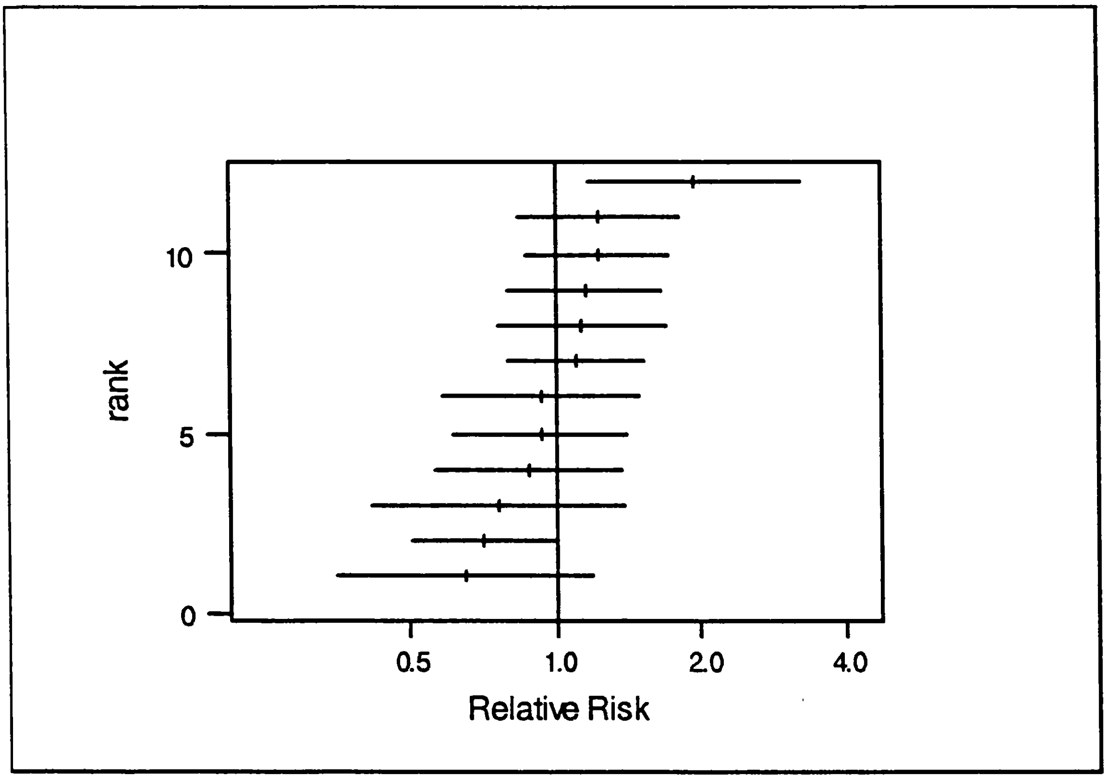


Figure 59: Unadjusted Relative Risk Confidence Intervals for Poor Outcome for Hospitals (CRAG Study Resection Data)

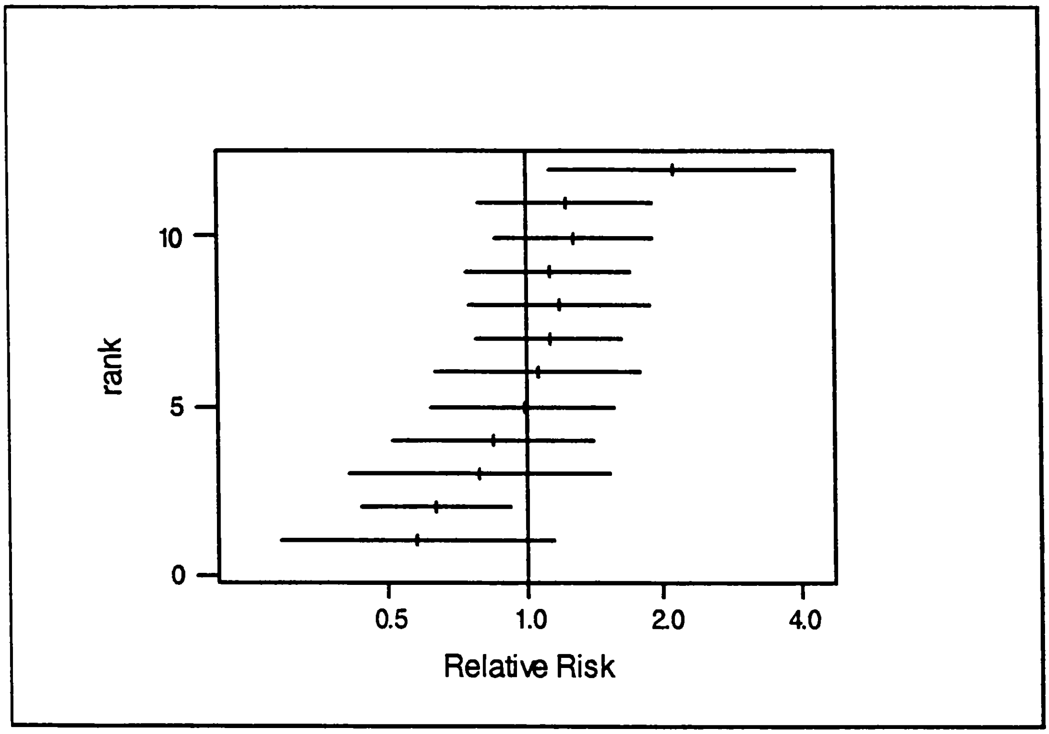


Figure 60: Adjusted Relative Risk Confidence Intervals for Poor Outcome for Hospitals (CRAG Study Resection Data)

10.3.3 Comparisons of consultants

Unadjusted and adjusted relative risks for consultants were calculated in the same way as for hospitals. The confidence intervals have been plotted in Figures 61 and 62. All the consultants who treated less than 10 patients have been included together as one "consultant", giving a total of 67. Of these, 10 had no poor outcomes so have not been included in the plots. The "consultant" representing those with less than 10 patients was ranked 35 (25 on plots) for its unadjusted value, having 6 poor outcomes out of a total of 70 patients. On adjustment, its rank moved up to 27 (17). This is, perhaps, a surprising result as one might expect those consultants treating very few patients to do worse. The relationship between volume and outcome will be explored in more detail in the next section.

The adjustments of consultants' relative risks had more effect than the hospitals'. Most of the unadjusted intervals (88%) contained 1. This figure is the same as for the adjusted intervals, although the same consultants do not have significant intervals each time. The movements are very slight compared with the RCS results. This could be due to the relatively small number of patients per consultant. To check whether there are actual differences between the consultants, we can fit it as a categorical variable in logistic regression as we did with the hospital data. This time, the variable is highly significant in the analysis of deviance, giving strong evidence that a difference exists even after adjustment. We may conclude that one consultant is performing significantly better in terms of numbers of poor outcomes than any of the others. The reasons for this should be investigated, and any beneficial aspects of practice noticed. Six consultants appear to be performing significantly worse than the others given their case mix, and should also investigate the reasons for this.

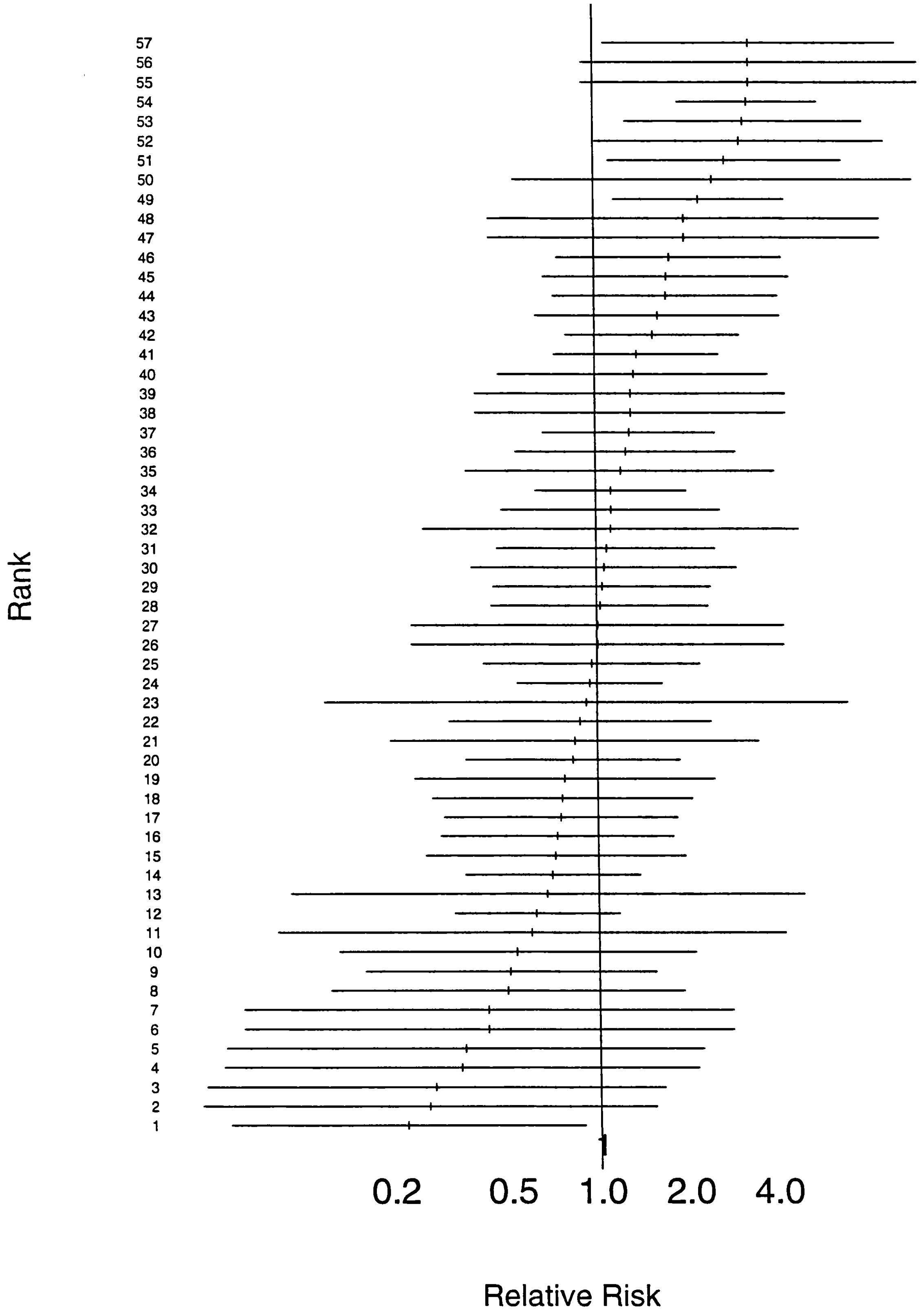


Figure 61: Unadjusted Confidence Intervals for Relative Risk of Poor Outcome for Consultants (CRAG Study, All Resections)

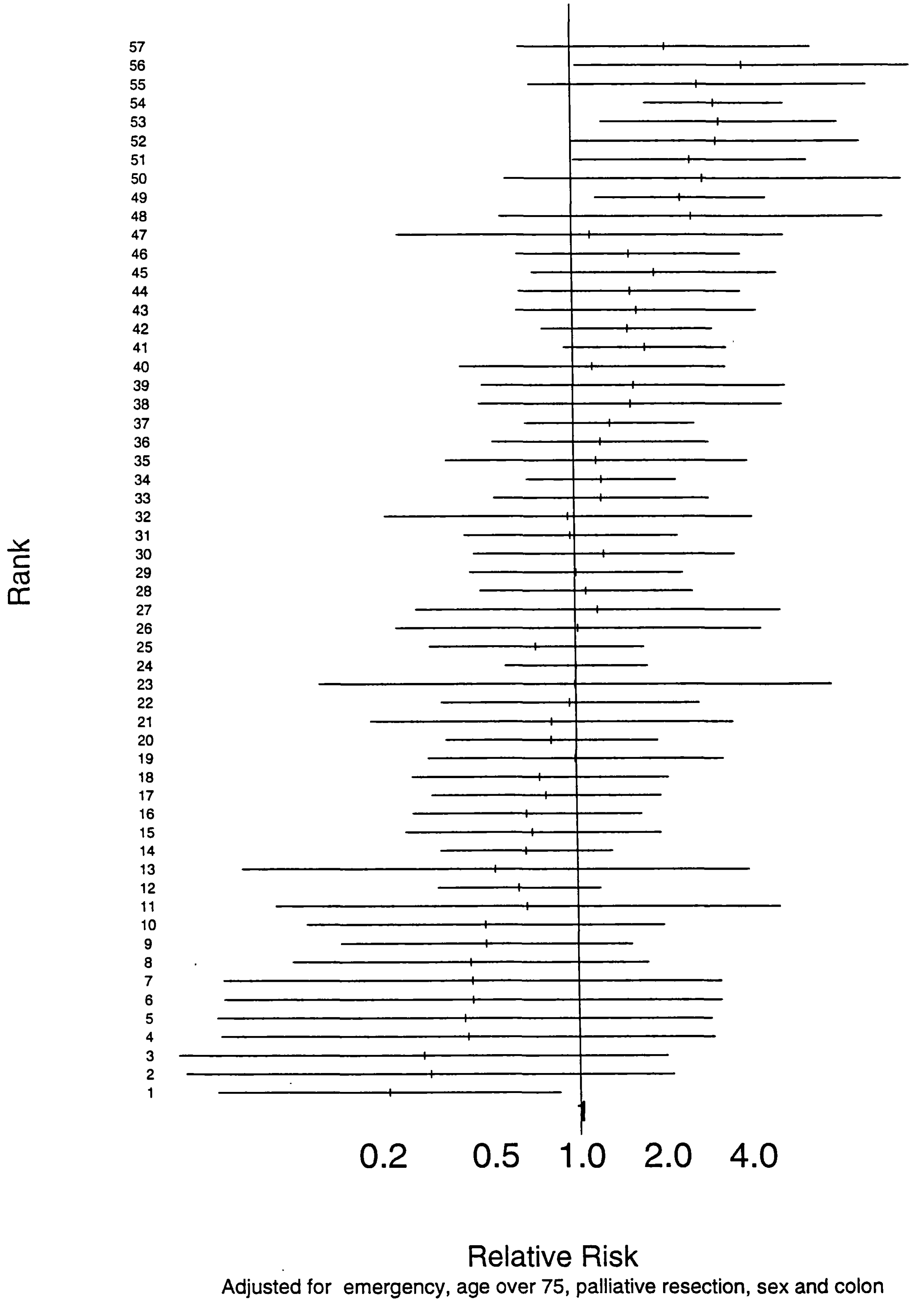
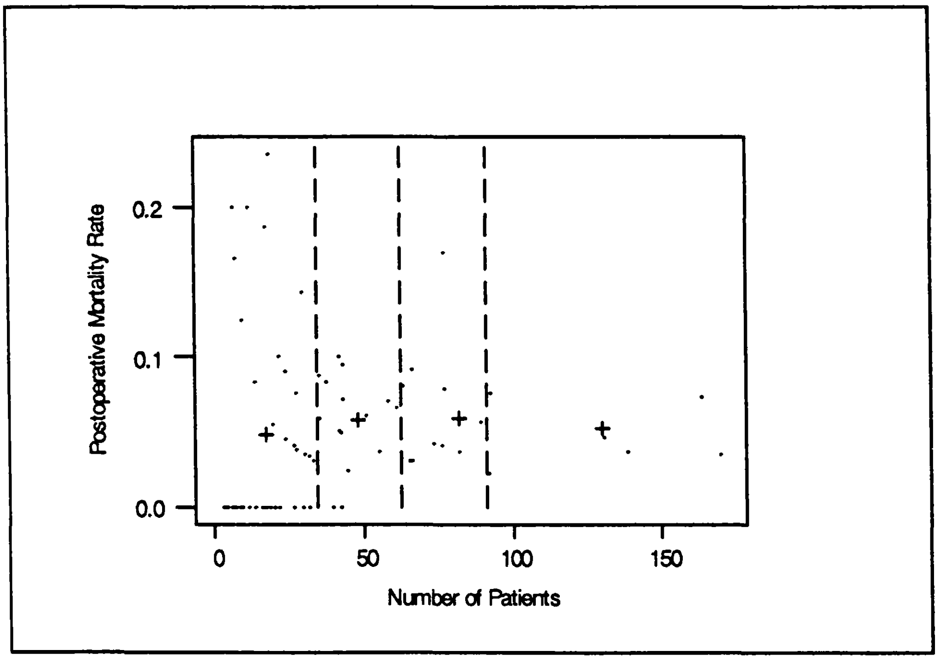


Figure 62: Adjusted Confidence Intervals for Relative Risk of Poor Outcome for Consultants (CRAG Study, All Resections)

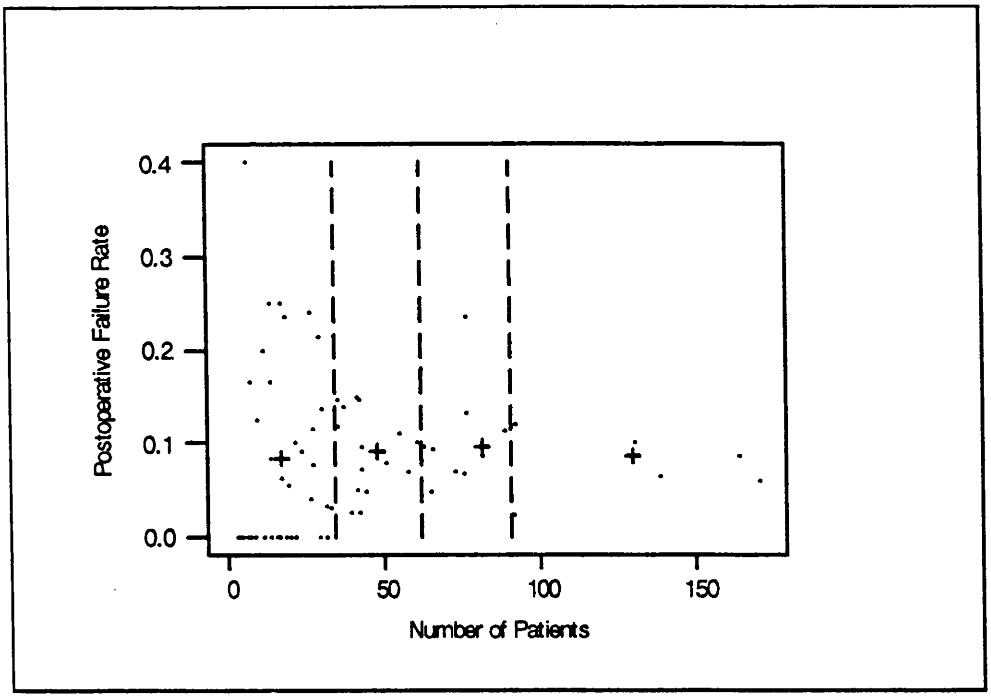
10.4 Relationship of number of patients treated with outcome

This area is important, as current thinking suggests that there is a case for having increased specialisation in treatment of cancers, as a larger volume of patients is seen as being related to higher success rates. McKee and Hunter (1995) produced a plot of mortality rate versus number of episodes per year for carcinoma of the colon, and suggested a threshold of about 400 episodes above which survival is higher. This was a study of 22 hospitals, and data were at a hospital level. Looking at their plot, it can be seen that only 3 hospitals had a volume of episodes greater than 400, and the spread of mortality rates among the other hospitals is very wide. If we plot our data for consultants (Figure 63), a similar pattern can be seen, where there appears to be a downward trend in mortality rate with increasing patient numbers. The plot has, however, been divided into quartiles of number of patients, so that approximately a quarter of patients treated lie in each section. The crosses show the mean mortality rate for each quartile, which are all around 0.05. This seems to show that the pattern in the plot is an illusion caused by larger variance with smaller numbers. Figure 64 shows a similar plot, but for all poor outcomes. The uniformity of patient quartile failure rates is even more striking than for mortality.



+ shows mean mortality rate for quartile

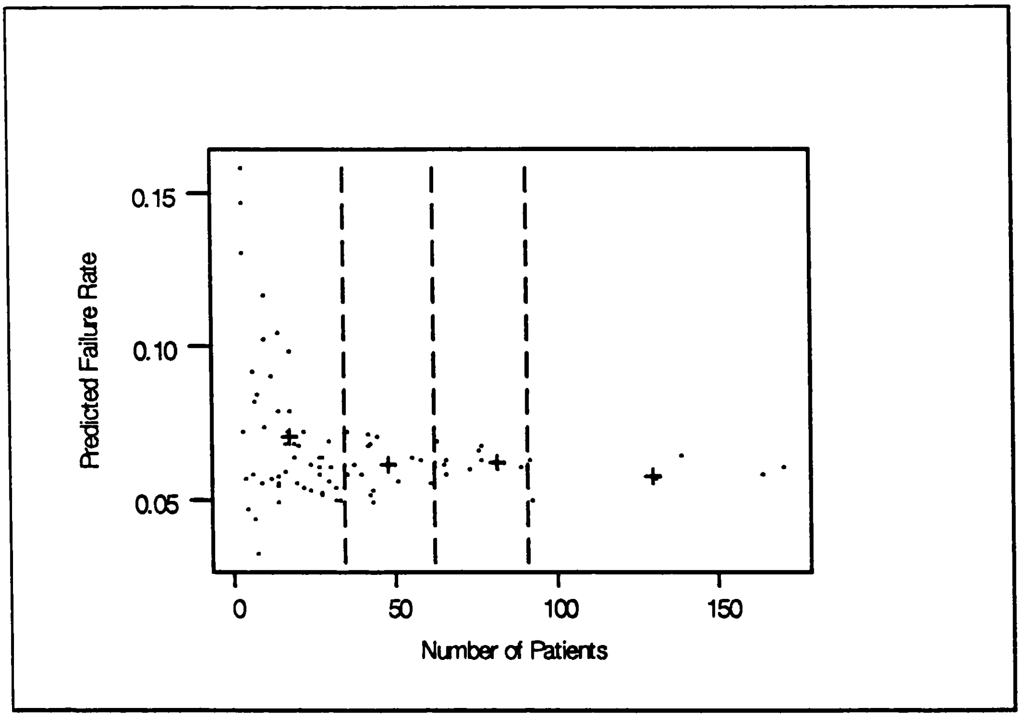
Figure 63: Postoperative Mortality Rate *versus* number of Patients Treated (CRAG Study)



+ shows mean failure rate for quartile

Figure 64: Postoperative Failure Rate *versus* Number of Patients Treated (CRAG Study)

In order to see whether the failure rates may be attributable to case mix, we could look at the predicted failure rates rather than the observed ones. These can be seen plotted against the number of patients in Figure 65. They were calculated using the logistic regression equation above to get predicted probabilities of adverse outcome for each patient, which were averaged to get the consultant's value. If it was the case that the consultants with the most patients also received the most difficult patients, their predicted failure rate would be far higher than the others. This is not the case.



+ shows mean failure rate for quartile

Figure 65: Predicted Rates of Poor Outcomes *versus* Numbers of Patients Treated (CRAG Study)

As well as these plots, investigation of the effect of patient volume was carried out by including number of patients as a variable in stepwise logistic regression. This was not entered into the model. If number of patients is fitted to postoperative poor outcome alone, it has a coefficient of -0.001 (s.e. = 0.001), so has a negligible effect. Categorical variables of whether the patient was treated by a consultant with under 20 patients or under 10 patients were also not significant. The quartile of number of patients into which the patients' consultant fell was also tried, but did not produce a significant result.

A factor which has not been accounted for here, but could affect the conclusions we have reached about the relationship between volume and outcome, would be the late entry to or early withdrawal from the study by any surgeons. We did not have the relevant data to assess whether this is an issue.

10.5 Discussion

In these analyses we have concentrated mainly on the outcome measure of postoperative failure, which includes mortality or any leak or abscess. We have shown up differences between consultants for this measure, but no significant difference between hospitals in the study. This outcome measure is, however, not the most important end result of resection for colorectal cancer. The ultimate survival of the patient, i.e. that they are cured by the treatment, is the key indicator of success of the surgery. It could be that a poor short term outcome rate is the result of a tendency to more aggressive surgery, which may achieve more favourable results in the long term. Audit of a specific procedure rather than all of general surgery would allow specific, relevant outcome measures such as longer term survival to be measured. Although consideration of long term outcome measures does render the analyses somewhat out of date, they are still very important for comparing surgical performance, as was shown by McArdle and Hole (1991).

As well as relevant outcome measures, studies of specific procedures allow more appropriate prognostic factors to be included. For example, Dukes' grade or site of the tumour are important variables in an audit of colorectal cancer, but it would not make sense to collect them on a general surgical data base. Any procedure one could choose would have specific prognostic factors which would allow simpler, more powerful statistical models to be found, and thus more accurate adjustments to outcome measures than are possible for general surgery. It is encouraging to note, however, that here presentation is, as in our modelling of general surgery, the most significant factor.

11 DISCUSSION

We set out to discuss statistical aspects of Surgical Audit. The majority of work has been concerned with what has become known as "Comparative Audit", where surgeons or units are compared with each other. It was stressed that the reasons for these comparisons should be to improve the quality of treatment rather than to reward and punish. Demand for comparative audit data is increasing, and large tables of performance figures are now published for hospitals, the latest being for England and Wales in June 1995. These statistics covered numbers of patients and lengths of waiting lists, but there is demand for more comparisons of outcome measures rather than process. Publication of mortality rates is a very sensitive area, which demands much care. If this type of outcome measure is to be released, it should be adjusted for patient severity. The data should also always be presented as confidence intervals, so as significant differences can be separated from those which could simply be due to random variation. In order to adjust outcome rates, we require predictive models. These should be well calibrated and have high sensitivity and specificity. We need well defined and relevant outcome measures with easily measured, objective prognostic factors that depend on the patient and not the clinician.

We have modelled mortality, which is not a particularly relevant outcome measure, for general surgery in two large studies. There are several reasons for this. Firstly, there is a need to look at general surgery and attempt to compare consultants. While mortality is not the best indicator of surgical skill, it is still sensitive enough to show up large differences between surgeons, and it is a measure that people understand and are interested in. Secondly, data on other outcome measures, such as complications, are poorly defined and notoriously poorly recorded, and as such are not yet reliable enough to use. It is difficult to obtain good quality data on individual procedures, such as we have for the colorectal cancer study, unless they have been collected for this sort of specific study. Large amounts of data are constantly being gathered on audit systems, which it is rather a shame to waste. The types of analyses we have considered are suitable for use on even the most general of surgical databases. We have highlighted some of the drawbacks of comparing surgeons by looking at general surgery mortality rates. The problems include the difficulty of modelling and the lack of sensitivity with such a rare outcome, as well as the irrelevance of the measure to the majority of patients. However, overall it is possible, and can discriminate between consultants given large amounts of accurate data. In many surgical departments, the data collected are very poor, often with wrongly coded or missing diagnoses. A recent study of hospital databases found that the records were unacceptably incomplete and inaccurate (Cleary *et al.*, 1994). This situation is improving, but until there are reliable audit data the models will not have much credibility. We first met this problem in our analyses of the data from the Royal College of Surgeons of England. These annual totals are submitted by some of the keenest advocates of audit, and yet many of the data are incomplete. It was hoped that this would improve as computer systems became established, but from the evidence of the 1994 data, this is not yet the case. The difficulties of obtaining the necessary data were only fully realised when we came to deal with a real hospital data base. The problems encountered with the Glasgow Royal Infirmary data could well be replicated around the country. The data from the Portsmouth hospitals, which were gathered for an investigation of the POSSUM system were another illustration of the difficulties of data collection. Many of the variables were missing, so several assumptions had to be made about the patients. This highlights the major drawback of the POSSUM system, which is currently the most popular predictive model for use in general surgery. That is that it relies on a large amount of data which is not routinely collected. The work in this thesis should be regarded as an exploration of the methodology, and its value in practice, rather than a source of definitive models for hospital use.

Logistic regression is a standard method of modelling binary outcomes, such as mortality, and is the "obvious" choice of method for the work done here. Presentation of mortality data as relative risks rather than simple mortality rates is a new way of approaching data presentation in this area. This allows one to immediately see whether a particular consultant's rate is different from average, where mortality rates do not. The "logarithm method" proposed by Katz *et al.* of calculating the unadjusted risks is as good as any other proposed method for the numbers involved in the RCS audit. Adjusting these relative risks is also an innovative step in this field. Previously, predicted mortality rates have been quoted (McKee and Hunter, 1995), and ratios of observed to expected mortality rates (Copeland *et al.*, 1995), but without suitable confidence intervals. While the plots are perhaps a little difficult to understand, most surgeons who attended the RCS comparative audit meeting found the confidence intervals given on their individual printouts very useful. If the data were improved and the methods refined to include an adjustment for operative severity rather than the inaccurate diagnostic risk groupings which we have used here, these could be of some use for future audits of this type.

We saw that with large enough numbers of consultants and patients, aggregate data can give almost as reliable adjustments as patient data, but superior analyses can always be done using patient level data. With the increase in computer power, it should be possible for data to be collected without the use of extensive paper forms, either on disk or via the internet, and so collection of patient details on a national basis should be possible. The prospect of collecting patient data has been suggested by consultants at recent RCS comparative audit meetings, and the comparative audit organisers are considering the possibilities.

If it is preferable to collect aggregate data, it could be that it would be more useful to collect scores rather than numbers in categories. For example, a consultant could score each patient on age, and then calculate the average age score for all his patients. This could be done for the POSSUM variables, or a subset of them. We saw in chapter 8 that scores gave more reliable models than actual measurements when considering patient data. We have not explored whether average scores would be any more powerful for modelling aggregate data than proportions in groups. It would, however, have the advantage that there was only one term in the model for each case mix variable.

We have shown that it is not necessary to collect all the data that are required by POSSUM to obtain a reasonable predictive system. In fact, our model containing only four scores was just as effective for adjusting consultants' risks. Thus, perhaps, the way forward is to gather large volumes of data, accurately on a limited number of variables, in order to make comparisons between consultants. A few variables could easily be collected as routine. The models could even be incorporated into a computer audit system, although it should be stressed that individual patient predictions will not be accurate.

Having described these analyses of general surgery, it should be acknowledged that this is perhaps not an ideal way to compare surgeons. A major problem with modelling general surgery is that there is such a wide range of operations, with so much difference in risk between them. The relationship between severity of the procedure or illness and risk of complication or death could be so strong that it dominates the model, so while there is overall accuracy, underlying details are hidden. As already stated, the outcome measure has to be relevant, and for general surgery the mortality rate is often very low. Different outcomes may be relevant for different procedures. If, instead of considering general surgery, we were to consider one particular procedure, the range of values of the risk would be greatly reduced and we would effectively 'stretch' a small section of the overall risk scale. If we considered only cholecystectomies for example we would be looking at a different scale of outcomes than for major laparotomies. The predictions should thus depend totally on characteristics of the patients and we could expect to spot more subtle differences in outcome due to management. Another advantage of considering only one operation in a predictive model is that the predictive variables can be more specific to the particular surgery, for example Dukes' grade for Colorectal Cancer. Also a few of them can be more powerful for prognosis. This makes measurement and calculation easier, as well as the prediction more precise.

Obviously, every type of operation carried out by general surgeons could not be covered, but an idea of overall performance could be gained by comparing observed and expected outcomes in a few indicator areas. Different types of procedure could be studied in a succession of audits. The 'indicator procedures' used would have to be fairly common so as enough were performed to make regular comparisons and so that a reasonable amount of data was available from which to draw conclusions. They would also have to be taxing enough to reflect surgical skill. A possible drawback of observing particular procedures could be that surgeons may, consciously or unconsciously, make more effort with those operations that 'count', at the expense of other operations not included in the exercise. Also, we can never obtain the large volumes of data which are available for general surgery. This leads to wide confidence intervals and less discrimination between consultants.

Mortality data on individual diagnoses were published for hospitals in Scotland in December 1994. They were part of a larger publication of 17 clinical outcome indicators. Mortality rates were published for heart attacks, strokes and fractured femurs, and large differences were shown between different hospitals. It was stressed that these differences could have been due to differences in patient case mix and not differences in quality of care. The only surgical outcome data produced were on reoperation for prostate surgery. The publication of outcome indicators is an improvement on the previous figures, which dealt with managerial and not medical aspects of care. If hospitals were found to have poorer outcome rates, they conducted investigations into why this was happening. It is important that the published figures are meaningful, and that case mix adjustments are used in future, as this type of information will be used by fundholders in future to decide where to send their patients. Data on social class and smoking, as well as improved measures of disease severity will be essential if realistic adjustments are to be made.

Having compared hospitals, it is only a matter of time before there will be demands to publicly compare individual consultants. While most consultants perform well, there are some who are recognised by their peers as substandard. Reliable data are required to objectively show up these consultants so that steps can be taken to improve their performance.

Comparing surgeons would be pointless if we did not act on the information gained. The problem is, what should be done, and how? There are many ethical problems involved. A consultant may be shown to be under performing compared with his peers. Although, by adjusting, we have tried to account for the major factors, we can never be sure that there is not some other reason, apart from surgical skill, for any differences. However, the data will in future be used to judge surgeons' ability. It is possible that a licensing system will be introduced, whereby only those who perform adequately will be permitted to perform specific operations.

The idea of licensing provokes a problem with emergency surgery, which is often carried out by the only available surgeon at the time. This is probably in the middle of the night when the surgeon is tired. An unlicensed surgeon may through necessity have to undertake the procedure, but it would be likely to have a poor outcome, due to lack of practice in an elective setting. This system would breed bad feeling among surgeons put in this position. They could be justified in refusing to carry out operations for which they did not have a licence.

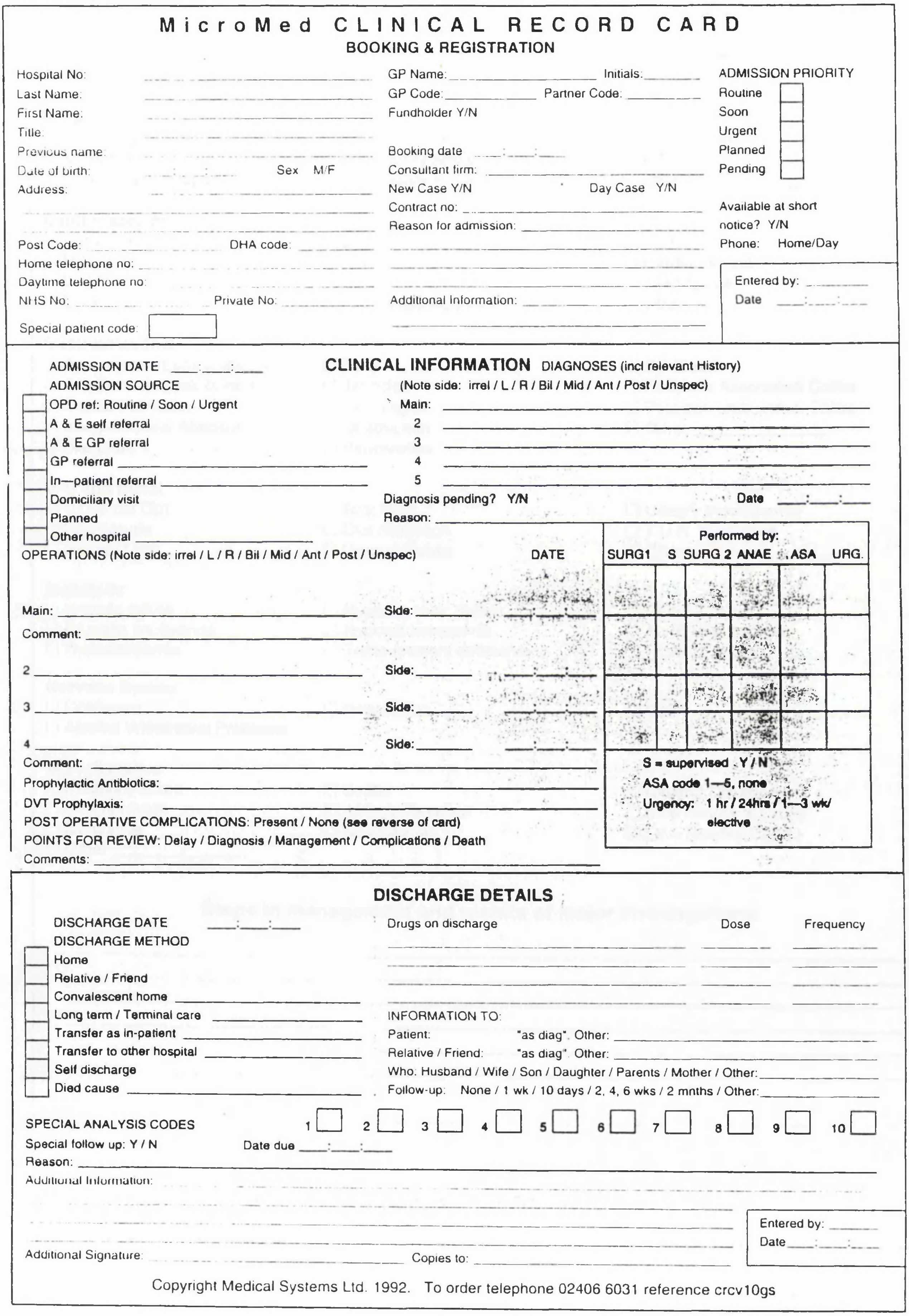
Making comparative audit results public brings difficult ethical issues from the patient point of view. They would then be entitled to ask why they had been operated on by a substandard surgeon, and why the initial training had not been good enough. While these questions must be answered for quality to improve, there is a real danger that more litigation will result. This could lead to an atmosphere of fear, and a reluctance for surgeons to take risks. Those performing best would see their waiting lists grow, and the less successful surgeons would find themselves with only the difficult emergency patients to treat, thus keeping their ratings at the bottom of the table. Another problem is the sensitivity of using figures such as mortality as measures to compare surgeons. This is often seen as distasteful by the general public, and figures are often blown out of proportion by the media.

Obviously, by the nature of lists, someone must be top and someone must be bottom. Positions in a list will not stay constant over time. It is essential that confidence intervals are used to show comparative audit data, so that differences due to chance are separated from differences which are highly unlikely to be merely random variation.

One could question whether it is legitimate at all to compare surgeons, who on the whole increase patients' survival prospects. Comparisons could lead to increased competition, and thus less communication of new ideas. In every field, from artists to engineers, there are people who are more skilled than others. They are all permitted to practise and improve. It is because of the life and death nature of surgeons' work that their performance is seen as critical. But with long hours and limited resources, it may be unreasonable to expect that each patient is treated perfectly.

APPENDICES

APPENDIX 1: Micromed Clinical Record Card



MicroMed CLINICAL RECORD CARD

BOOKING & REGISTRATION

Hospital No: ____

Last Name: ____

First Name: ____

Title: ____

Previous name: ____

Date of birth: __:__:__ Sex M/F

Address: ____

Post Code: ____ DHA code: ____

Home telephone no: ____

Daytime telephone no: ____

NHS No: ____ Private No: ____

Special patient code: □

GP Name: ____ Initials: ____

GP Code: ____ Partner Code: ____

Fundholder Y/N

Booking date __:__:__

Consultant firm: ____

New Case Y/N Day Case Y/N

Contract no: ____

Reason for admission: ____

Additional Information: ____

ADMISSION PRIORITY

Routine □

Soon □

Urgent □

Planned □

Pending □

Available at short notice? Y/N

Phone: Home/Day

Entered by: ____

Date __:__:__

CLINICAL INFORMATION

ADMISSION DATE __:__:__

ADMISSION SOURCE

□ OPD ref: Routine / Soon / Urgent

□ A & E self referral

□ A & E GP referral

□ GP referral ____

□ In—patient referral ____

□ Domiciliary visit

□ Planned

□ Other hospital ____

DIAGNOSES (incl relevant History)

(Note side: irrel / L / R / Bil / Mid / Ant / Post / Unspec)

Main: ____

2 ____

3 ____

4 ____

5 ____

Diagnosis pending? Y/N Date __:__:__

Reason: ____

OPERATIONS (Note side: irrel / L / R / Bil / Mid / Ant / Post / Unspec) DATE

Main: ____ Side: ____ __:__:__

Comment: ____

2 ____ Side: ____ __:__:__

3 ____ Side: ____ __:__:__

4 ____ Side: ____ __:__:__

Comment: ____

Prophylactic Antibiotics: ____

DVT Prophylaxis: ____

POST OPERATIVE COMPLICATIONS: Present / None (see reverse of card)

POINTS FOR REVIEW: Delay / Diagnosis / Management / Complications / Death

Comments: ____

Performed by:

SURG1	S	SURG 2	ANAE	ASA	URG.

S = supervised Y / N

ASA code 1—5, none

Urgency: 1 hr / 24hrs / 1—3 wk/ elective

DISCHARGE DETAILS

DISCHARGE DATE __:__:__

DISCHARGE METHOD

□ Home

□ Relative / Friend

□ Convalescent home ____

□ Long term / Terminal care ____

□ Transfer as in-patient ____

□ Transfer to other hospital ____

□ Self discharge

□ Died cause ____

Drugs on discharge	Dose	Frequency

INFORMATION TO:

Patient: "as diag". Other: ____

Relative / Friend: "as diag". Other: ____

Who: Husband / Wife / Son / Daughter / Parents / Mother / Other: ____

Follow-up: None / 1 wk / 10 days / 2, 4, 6 wks / 2 mnths / Other: ____

SPECIAL ANALYSIS CODES 1 □ 2 □ 3 □ 4 □ 5 □ 6 □ 7 □ 8 □ 9 □ 10 □

Special follow up: Y / N Date due __:__:__

Reason: ____

Additional Information: ____

Entered by: ____

Date __:__:__

Additional Signature: ____ Copies to: ____

Copyright Medical Systems Ltd. 1992. To order telephone 02406 6031 reference crcv10gs

Complications of Operation or Treatment

Wound

□ Deep infection
□ Haematoma
□ Cellulitis
□ Other ______________
□ Superficial Infection
□ Serious Collection
□ Wound Sinus
□ Incisional hernia
□ Ischaemic wound failure
□ Dehiscence

Respiratory

□ Sputum Retention
□ Aspiration
□ Pneumonia/Collapse
□ Pneumothorax
□ Respiratory Arrest
□ Respiratory Failure
□ Shock Lung
□ A.R.D.S.
□ Other ______________

Cardiovascular

□ Arrhythmia
□ M.I.
□ C.C.F.
□ L.V.F.
□ D.V.T.
□ C.V.A.
□ P.E.
□ Hypovolaemic shock
□ T.I.A.
□ Cardiac Arrest
□ Graft failure
□ Other ______________

Gastrointestinal

□ Anastomotic Leak (clinical)
□ Anastomotic Leak (x-ray)
□ Fistula Formation
□ Intraperitoneal Abscess
□ Bile Leak
□ Ileus
□ Jaundice
□ Cholangitis
□ Obstruction
□ Pancreatitis
□ G-I Bleed
□ Antibiotic Associated Colitis
□ Pseudomembranous Colitis
□ Other ______________

G-U and Renal

□ UTI (Post Op)
□ Haematuria
□ Urinary Leak
□ Retention
□ Clot Retention
□ Urinary Fistula
□ Urinary incontinence
□ T.U.R. Syndrome
□ Other ______________

Metabolic

□ Hepatic failure
□ Glucose intolerance
□ Hyponatraemia
□ Hyperosmolar States
□ Hypoalbuminaemia
□ Trace element deficiency
□ Renal Failure
□ Acid/base disorder
□ Other ______________

Nervous System

□ Confusion
□ Alcohol Withdrawal Problems
□ Depression
□ Other ______________

Miscellaneous

□ Pressure Sores
□ Haemorrhage
□ Ischaemia
□ Other ______________
□ Septic
□ Allergic Reaction
□ Septicaemia
□ Clotting Disorder
□ Side effects of Drugs
□ I.V.I. Related Sepsis

Steps in management and results of major investigations

____

____

____

____

____

____

____

____

____

____

Reorder from Medical Systems Ltd. (02406) 6031 Ref. crcv10gs

APPENDIX 2

Royal College of Surgeons of England 14th June 1994. Questionnaire to assess response to new Comparative Audit presentation methods.

Please fill in your personal identification number ____________

What is your position? Consultant Other (please state)________________

How did you collate the data? 1. Commercial audit package (please specify)

____

2. Hospital Activity Analysis (HAA) system
3. Hospital Audit Department
4. Other ____________________

Did you receive a new style personal printout? Yes No

Presentation Methods

There were three new methods presented: triangle plots showing case mix, confidence interval plots of unadjusted and adjusted relative mortalities and a personal sheet giving case mix and confidence interval figures.

1. Clarity

Please circle a score to indicate how clear and easy to understand each of the methods of presentation was to you.

Completely clear........................Incomprehensible

Triangle Plots	5	4	3	2	1
Confidence Intervals	5	4	3	2	1
Personal Information	5	4	3	2	1

2. Usefulness

Please circle a score on the scale to indicate how useful each of the methods of presentation was to you.

Highly informative........................Useless

Triangle Plots	5	4	3	2	1
Confidence Intervals	5	4	3	2	1
Personal Information	5	4	3	2	1

PTO

Will you carry out any investigations in the light of your results? (please circle)

Yes **No**

Would you like to see further development of these methods for future use?

Yes **No**

Are there any other pieces of information you would like to see presented in future?

____

____

____

How did these methods compare with those used in previous years? (please circle)

more informative **equally informative** **less informative**

Any other comments/ suggestions

____

____

____

____

____

____

____

Please leave your completed questionnaire in one of the boxes provided.

Thank you very much for your help.

APPENDIX 3: Example Consultant Printout

COMPARATIVE AUDIT 1994 **Consultant XXX**

Total deaths: 19 **Total admissions:** 1343

Mortality rate: 1.4% **Overall mean:** 1.7% (range 0.0% to 3.95%)

Relative mortality: 80% (95% CI 51% to 126%) **Rank:** 34 out of 89

Number of admissions with no operation: 437

Number of non-operative deaths: 11

<u>Case Mix</u> (percent of total with information supplied)

1. **Age**

	10 and under	11 to 60	over 60
your data	**8%**	**63%**	**28%**
overall mean	5%	55%	40%

Relative mortality adjusted for age: 95% (95% CI 60% to 150%)
Rank: 42 out of 88

2. **Admission category**

	Daycase	Elective	Emergency
your data	**15%**	**31%**	**53%**
overall mean	37%	32%	31%

Relative mortality adjusted for admission:59% (95% CI 37% to 92%)
Rank: 15 out of 89

3. **Diagnostic risk category**

	Low	Medium	High
your data	**37%**	**31%**	**33%**
overall mean	20%	40%	40%

Relative mortality adjusted for diagnosis: 87% (95% CI 56% to 139%)
Rank: 43 out of 83

Relative mortality adjusted for all of the above: 72% (95% CI 43% to 121%)
Rank after adjustment: 22 out of 82

APPENDIX 4

Quotes from Questionnaire to Royal College of Surgeons on New Presentation Methods

Other information requested

* "Relate to complexity of operation - better reflection of case mix"
* "Include ASA data for a <u>much</u> better handle on case mix"
* "Categorisation of expected vs. unexpected deaths"
* "Mortality should indicate deaths from terminal disease"
* "Post op/ non operative deaths separated"
* "Analysis by operator as well as consultant"
* "How many hospitals responded?"
* "Relative mortalities within main specialist groups (e.g. vascular)"

Comments and Suggestions

* "Non operative mortality should not be included because it may just reflect lack of local hospice provision"
* "We have a commitment to palliative care and terminal care"
* "Collecting mortality data without any other items is meaningless UNLESS one can categorise whether the death was 'inevitable' or 'preventable'. Our audit system does not provide such a split in the data and many of our deaths were 'inevitable' and occurred because of terminal disease. This must apply to other units as well!"
* "Leave out the BUPA classification of operations and ASA grades as almost impossible to collect"
* "Data collection package"
* "Without in-house audit, I find collection of data extremely difficult"
* "Identification of patients remains a problem. How many of us identify patients for audit ourselves and how many rely on the hospital information system?"
* "Evaluation of methodology needs to be discussed. It might help next year to ask, say, 5-6 surgeons to present their methods of data collection."

* “Is the data validated by another method? One should be sceptical of e.g. Micromed data collected for a fee (£250). Valuable data will only come from interested surgeons, who are interested in accurate data retrieval.”
* “Unfortunately in our case the case mix data collected does not provide any information on the illness or severity of the patients treated, nor does it collect all the operations formed since the list of codes used was incomplete”
* “Separate categories for day case surgery, outpatient and inpatient cases”
* “Diagnostic groups could be further split to separate high/ low risk groups”
  “Include ‘Abdominal Pain’ (7890) as a separate diagnostic category”
* “Can ASA grading be included in mortality computations? (VERY important)”
* “I am unconvinced about the use of diagnostic risk category as helpful - it might be better to use high risk procedures”
* “The adjustments for age, mode of admission and diagnostic category are an improvement over the raw data, but still have great potential for abuse. The more refined methods of risk adjustment (i.e. POSSUM) would be difficult, however, at present throughout the country, but may be more applicable”
* “Would like to review against POSSUM”
* “Need to encourage more responders - perhaps newsletter type information to all hospitals / consultants to emphasise the positive information that has come from Comparative Audit studies”
* “The comments from presenters and the audience suggest extremely over-confident interpretation. This probably means ‘that’ when the quality of ‘this’ is completely unknown”
* “Can you clarify the figures given - there seem to be two percentages - which is the most important? E.g. a% (b% CI d% to e%). Which of a% or b% is the figure to give more weight to and what are the two different figures representing?”
* “I find it difficult to comprehend the ‘relative mortality’ concept and cannot see how it has been related to age or case mix!”

“Delighted with results - keep it up!”

APPENDIX 5

Equations used for log-linear model Simulations

The data for two variables are of the form

		age			
		1	2	3	total
	1	n_{11}	n_{12}	n_{13}	n_{s1}
status	2	n_{21}	n_{22}	n_{23}	n_{s2}
	3	n_{31}	n_{32}	n_{33}	n_{s3}
	total	n_{a1}	n_{a2}	n_{a3}	N

We have the model

$$\log(E(n_{ij})) = \theta + \lambda_{s_i} + \lambda_{a_j} + \beta(i-2)(j-2)$$

We know the marginal totals $n_{s_i} = \sum_{j=1}^{3} n_{ij}$ i=1,2,3

and $n_{a_j} = \sum_{i=1}^{3} n_{ij}$ j=1,2,3

Since $E(n_{s_i}) = n_{s_i}$ and $E(n_{a_j}) = n_{a_j}$

$$E(\sum_j n_{ij}) = \sum_j E(n_{ij}) = n_{s_i} \quad \Rightarrow \quad \log(\sum_j E(n_{ij})) = \log(n_{s_i})$$

From the model, $E(n_{ij}) = e^{\theta} e^{\lambda s_i} e^{\lambda a_j} e^{\beta(i-2)(j-2)}$ so

$$n_{s_i} = \sum_j E(n_{ij}) = e^{\theta} e^{\lambda s_i} e^{\lambda a_1} e^{-\beta(i-2)} + e^{\theta} e^{\lambda s_i} e^{\lambda a_2} + e^{\theta} e^{\lambda s_i} e^{\lambda a_3} e^{\beta(i-2)}$$

$$n_{a_j} = \sum_i E(n_{ij}) = e^{\theta} e^{\lambda s_1} e^{\lambda aj} e^{-\beta(j-2)} + e^{\theta} e^{\lambda s_2} e^{\lambda a_j} + e^{\theta} e^{\lambda s_3} e^{\lambda a_j} e^{\beta(j-2)}$$

which give the six equations which must be solved.

APPENDIX 6: Calculation of POSSUM Scores

The score consists of two parts: Physiological and Operative Severity. These consist of separate factors which are given weights of 1, 2, 4 or 8 depending on their severity. The weights are then summed to obtain the two scores, which are incorporated in a logistic regression equation. The scores are assigned as follows.

Physiological Score

	Score			
	1	2	4	8
Age (years)	≤ 60	61-70	≥ 71	
Cardiac Signs	No failure	Diuretic, digoxin, antianginal or hypertensive therapy	Peripheral oedema; warfarin therapy	Raised jugular venous pressure
Chest radiograph			Borderline cardiomegaly	Cardiomegaly
Respiratory History	No dyspnoea	Dyspnoea on exertion	Limiting dyspnoea (one flight)	Dyspnoea at rest (r ≥ 30/min)
Chest radiograph		Mild COAD	Moderate COAD	Fibrosis or consolidation
Blood Pressure (systolic) (mmHg)	110-130	131-170	≥ 171	≤ 89
		100-109	90-99	
Pulse (beats/min)	50-80			≥ 121
				≤ 39
Glasgow coma score	15	12-14	9-11	≤ 8
Haemoglobin (g/100ml)	13-16	11.5-12.9	10.0-11.4	≤ 9.9
		16.1-17.0	17.1-18.0	≥ 18.1
White cell count ($\times 10^{12}$/l)	4-10	10.1-20.0	≥20.1	
		31.-4.0	≤ 3.0	
Urea (mmol/l)	≤ 7.5	7.6-10.0	10.1-15.0	≥ 15.1
Sodium (mmol/l)	≥ 136	131-135	126-130	≤ 125
Potassium (mmol/l)	3.5-5.0	3.2-3.4	2.9-3.1	≥ 2.8
		5.1-5.3	5.4-5.9	≥ 6.0
Electrocardiogram	Normal		Atrial fibrillation (rate 60-90)	Any other abnon rhythm or ≥ 5 ectopics/min Q waves or ST/T wave changes

COAD=Chronic Obstructive Airways Disease

Operative Severity Score

	Score			
	1	2	4	8
Operative severity	Minor	Moderate	Major	Major+
Multiple procedures	1		2	>2
Total blood loss (ml)	≤ 100	101-500	501-999	≥ 1000
Peritoneal Soiling	None	Minor (serious fluid)	Local pus	Free bowel cont pus or blood
Presence of malignancy	None	Primary only	Nodal metastases	Distant metastases
Mode of surgery	Elective		Urgent (<24 hours)	Emergency(<2 ho

Any variable with missing data is given a weight of 1.

REFERENCES

American Society of Anesthesiologists, 1963. New classification for physical status. *Anesthesiology*;**24**:111

Ausobsky JR *et al.,* 1982. Delayed hypersensitivity testing for the prediction of postoperative complications. *Br J Surg*;**69**:346-8

Bailey BJR,1987. Confidence Limits to the Risk Ratio. *Biometrics*;**43**:201-5

Baker JP, Detsky AS, Wesson DE, Wolman SL, Stewart S, Whitewell J, Langer B, JeejeebhoyKN, 1982. Nutritional assessment - a comparison of clinical judgement and objective measurements. *N Engl J Med.*;**306**:969-72

Baron JH, 1988. Quality control and audit - Annual General Meeting of the Royal College of Surgeons of England, 9 December 1987, London. *Theor Surg*;3:27-8

Baum M, 1991. New approach for recruitment into randomised trials. *Lancet* 1993;**341**:812-3 Boeke S *et al.* Psychological variables as predictors of the length of postoperative hospitalization. *J Psychosom Res*;**35**:281-8

Beecham L, 1993. Consultants outraged by league tables. *BMJ*;**307**:699

Boeke S, Stronks D, Verhage F, Zwaveling A, 1991. Psychological variables as predictors of the length of postoperative hospitalization. *J Psychosom Res*;**35**:281-8

Boyd O, Grounds RM, 1993. Physiological Scoring Systems and Audit. *Lancet*;**341**:1573-4

Brenner U, Walters U, Muller JM, 1989. A simple point system for preoperative assessment of operative risk. *Theor Surg*;**4**:17-21

Brindle B, 1994. NHS to run death rate leagues. *Guardian* 23 Nov;1

British United Provident Association 1989. BUPA Schedule of Procedures. London: BUPA.

Brook RH, Appel FA, 1973.Quality of care assessment:choosing a method for peer review. *New Engl J Med*;**288**:1323-9

Buchman TG, Kubos KL, Seidler AJ, Siegforth MJ, 1994. A comparison of statistical and connectionist models for the prediction of chronicity in an intensive care unit. *Crit Care Med.*;**22**:750-62

Buck N, Devlin HB, Lunn JN, 1987. The Report of a Confidential Enquiry into Perioperative Deaths. *Nuffield Provincial Hospitals Trust and the King's Fund,* London

Buzby GP, Mullen JL, Matthews DC, Hobbs CL, Rosato EF, 1980. Prognostic Nutritional Index in Gastrointestinal Surgery. *Am J Surg.*;**139**:160-7

Byar DP, 1980. Why Data Bases Should Not Replace Randomized Clinical Trials. *Biometrics*;**36**:337-42

Cale ARJ, King PM, Macleod DAD, 1991. Practical surgical audit: a morbidity profile. *J R Coll Surg Edin*;**36**:41-4

Chapuis PH *et al.*, 1985. A multivariate analysis of clinical and pathalogical variables in prognosis after resection of large bowel cancer. *Br J Surg*.;72:698-702

Christou NV *et al.*, 1981. The predictive role of DH in preoperative patients. *Surg Gynecol Obstet.*;152:297-301

Cleary R, Beard R, Coles J, Devlin B, Hopkins A, Schumacher D, Wickings I., 1994. Comparative hospital databases: value for management and quality. *Quality in Health Care*;3:3-10

Consultant Surgeons and Pathologists of the Lothian and Borders Health Boards, 1995. Lothian and Borders large bowel cancer project: immediate outcome after surgery. *Br J Surg*.;82:888-90

Copeland GP, Jones D, Walters M, 1991. POSSUM: a scoring system for surgical audit *Br J Surg*.;**78**:356-60

Copeland GP, Jones D, Harris PL, Wilcox A, 1993. Comparative Vascular Audit using the POSSUM scoring system. *Anns R Coll Surg Engl*;**75**:175-7

Copeland GP, 1993. Comparative Audit: fact *versus* fantasy. *Br J Surg*.;**80**:1424-5

Copeland GP, Sagar P, Brennan J, Roberts G, Ward J, Cornford P, Millar A, Harris C, 1995. Risk-adjusted analysis of surgeon performance: a 1-year study. *Br J Surg*.;82:408-411

Crombie IK, Davies HTO, 1993.Missing link in the audit cycle *Quality in Health Care*;**2**:47-8

Daley J, Jencks S, Draper D, Lenhart G, Thomas N, Walker J, 1988. Predicting hospital associated mortality for Medicare patients, *JAMA*;**260**:3617-24

Deans GT, Odling-Smee W, McKelvey STD, Parks GT, Roy DA, 1987. Auditing perioperative mortality *Anns R Coll Surg Engl*;**69**:183-7

Deans GT, Heatley M, Patterson CC, Moorehead RJ, Parks TG, Rowlands BJ, Spence RAJ, 1994. Colorectal carcinoma: importance of clinical and pathological factors in survival. *Anns R Coll Surg Engl*;**76**:59-64

Dempsey DT, Mullen JL, Buzby GP, 1988. The link between nutritional status and clinical outcome: Can nutritional intervention modify it? *Am J Clin Nutr*.;**47**:352-6

Devlin B, 1990. Audit and the quality of clinical care. *Anns R Coll Surg Engl*.;72:supp3-14

Doyle HR, Dvorchik I, Mitchell S, Marino IR, Ebert FH, McMichael J, Fung JJ, 1994. Predicting outcomes after liver transplantation - a connectionist approach. *Anns Surg*.;**219**:408-15

Dudley HAF, 1974. Necessity for Surgical Audit *BMJ* (i):275-7

Dunn DC, Dale RF, 1986. Combined computer generated discharge documents and surgical audit *BMJ*;**292**:816-8

Dunn DC, 1988. Audit of a surgical firm by microcomputer: 5 years experience. *BMJ*;**296**:687-91

Dunn DC, Fowler S, 1990. Comparative audit: an experimental study of 147,882 general surgical admissions during. *Br J Surg* 1992;**79**:1073-6

Dunn DC, Dale RF, Gumpert JRW, Duffy TJ, 1992. Combined surgical audit by microcomputer involving units in four health regions. *Anns R Coll Surg Engl*;**74**:47-53

Editorial 1974: Towards medical audit. *BMJ*:255

Editorial 1976: Separating the sheep from the goats. *BMJ*: 1218

Edwards G, Morton HJV, Pask EA, Wylie WD, 1956. Deaths associated with anaesthesia. *Anaesthesia*;**11**:194-220

Ellis BW, Michie HR, Esufali ST, Pyper RJD, Dudley HAF, 1987. Development of a microcomputer-based system for surgical audit and patient administration: a review. *J Roy Soc Med*;**80**:157-61

Ellis BW, 1989. How to set up an audit BMJ;**298**:1635-7

Emberton M, Rivett R, Ellis BW, 1991. Comparative Audit: A new method of delivering audit *Anns R Coll Surg Engl.*;**73**:suppl. 117-20

Fielding LP, Stewart-Brown S, Blesovsky L, Kearney G, 1980. Anastomotic integrity after operations for large bowel cancer: a multicentre study. *BMJ*;**288**:411-4

Gardner MJ, Altman DG, 1989. *Statistics with Confidence*. BMJ

Gart JJ, Nam J, 1988. Approximate Interval Estimation of the Ratio of Binomial Parameters: A Review and Corrections for Skewness. *Biometrics*;**44**:328-38

Gilmore OJA, Griffiths NJ, Connoly JC, Dunlop AW, Hart S, Thomson JPS, Todd IP, 1980. Surgical audit: comparison of the workload and result of two hospitals in the same district. *BMJ*;**281**:1050-2

Goldman L, Caldera Dlm, Nussbaum SR, Southwick FS, Krogstad D, Murray B, Burke DS, O'Malley TA, Goroll AH, Caplan CH, Nolan J, Carabello B, Slater EE, 1977. Multifactorial index of cardiac risk in noncardiac surgical procedures *N Engl J Med.*;**297**:845-50

Gonnella JS, Hornbrook MC, Louis DZ, 1984. Staging of Disease. A Case-Mix Measurement. *JAMA*;**251**:637-44

Gore SM, 1981. Assessing clinical trials - Why randomise? *BMJ*;**282**:1958-60

Gough MH, Kettlewell MGW, Marks CG, Holmes SJK, Holderness J, 1980. Audit: an annual assessment of the work and performance of a surgical firm in a regional teaching hospital. *BMJ*;**281**:913-918

Green J, Winfield N, Sharkey P, Passman LJ, 1990. The importance of severity of illness in assessing hospital mortality. *JAMA*;**263**:241-6

Gruer R, Gordon DS, Gunn AA, Ruckley CV, 1986. Audit of surgical audit.;*Lancet*(i):23-6

Hartley MN, Sagar PM, 1994. The surgeon's 'gut feeling' as a predictor of post-operative outcome *Anns R Coll Surg Engl.*;**76**:suppl. 277-8

Harvey KB, Moldawer LL, Bistrian BR, Blackburn GL, 1981. Biological measures for the formulation of a hospital prognostic index. *Am J Clin Nutr.*;**34**:2013-22

Hart A, Wyatt J, 1990. Evaluating black-boxes as medical decision aids: Issues arising from the study of neural networks. *Med Inf*;**15**:229-36

Heywood AJ, Wilson IH, Sinclair JR, 1989. Perioperative mortality in Zambia. *Anns R Coll Surg Eng*;**71**:185-7

Hirshberg A, Adar R, 1990. Preoperative prediction of postoperative complications. *Isr J Med Sci*;**26**:123-4

Houghton A, 1994. Variation in outcome of surgical procedures. *Br J Surg*;**81**:653-60

Jones DR, Copeland GP, Decossart L, 1992. Comparison of POSSUM with APACHE II for prediction of outcome from a surgical high dependency unit*Br J Surg*;**79**:1293-6

Jones J, 1993. Doctors condemn NHS league table. *Independent*. September 10: 2

Jencks SF, Williams DK, Kay TL, 1988. Assessing Hospital-Associated Deaths From Discharge Data: The Role of Length of Stay and Comorbidities. *JAMA*;**260**:2240-6

Jencks SF, Daley J, Draper D, Thomas N, Lenhart G, Walker J, 1988. Interpreting hospital mortality data. The role of clinical risk adjustment. *JAMA*:**260**:3611-6

Katz D, Baptista J, Azen SP, Pike MC, 1978. Obtaining Confidence Intervals for the Risk Ratio in Cohort Studies. *Biometrics*;**34**:469-74

Kennedy RH, Al-Mufti RAM, Brewster SF, Sherry EN, Magee TR, Irvin, TT, 1994. The acute surgical admission: is mortality predictable in the elderly? *Anns R Coll Surg Eng*;**76**:342-5

Klidjian AM *et al.*, 1980. Relation of anthropometric and dynamometric variables to serious postoperative complications. *BMJ*;**281**:899-901

Knaus WA, Zimmerman JE, Wagner DP, Draper EA, Lawrence DE, 1981. APACHE - acute physiology and chronic health evaluation: a physiologically based classification system. *Crit Care Med*:**9**:591-7

Knaus WA, Draper EA, Wagner DP, Zimmerman JE, 1985. APACHE II: a severity of disease classification system. *Crit Care Med.*;**13**:818-29

Knaus WA. The science of prediction and its implications for clinicians today. *Theor Surg* 1988;3:93-101

Knaus WA, Wagner DP, Draper EA, Zimmerman JE, *et al.*, 1991. The APACHE III Prognostic System. *Chest*;**100**:1619-36

Kohler L, Viell, B, Bode C, Vestweber K-H, Troidl H, 1988. A prospective trial of the Prognostic Nutritional Index in prediction of postoperative morbidity and mortality. *Theor Surg.*;**3**:3-7

Koopman PAR, 1984. Confidence Intervals for the Ratio of Two Binomial Proportions. *Biometrics*;**40**:513-7

Krenzien J, Roding H, Mummelthey R, 1989. Operative risk in octogenerians - a statistical prognostic index and its prospective validation. *Theor Surg*;**4**:10-16

Le Gall JR, Lemeshow S, Saulnier F, 1993. A new simplified acute physiology score (SAPS - II) based on a European Nort American multicenter study. *JAMA*;**270**:2957-63

Lemeshow S, Le Gall JR, 1994. Modeling the severity of illness of ICU patients - a systems update. *JAMA*;**272**:1049-55

Lemeshow S, Teres D, Klar J, Avrunin JS, Gehlbach SH, Rapoport J, 1993. Mortality Probability Models (MPM-II) based on an international cohort of intensive care patients. *JAMA*;**270**:2478-86

McArdle CS, Hole D, 1991. Impact of variability among surgeons on postoperative morbidity and mortality and ultimate survival. *BMJ*;**302**:1501-5

McKee M, Hunter D, 1995. Mortality league tables: do they inform or mislead? *Quality in Health Care*;**4**:5-12

Meakins JL *et al.*, 1980. Predicting surgical infection after operation. *World J Surg.*;**4**:439-50

Miettinen O, Nurminen M, 1985. Comparative analysis of two rates. *Statistics in Medicine*;**4**:213-26

Miholic J, Haumer M, Moeschl P, Schemper M, 1988. Risk factors for mortality after total gastrectomy, evaluated by logistic regression analysis. *Theor Surg*;**3**:78-82

Mishriki SF, Law DJW, Jeffrey PJ, 1990. Factors affecting the incidence of postoperative wound infection. *J Hosp Inf*;**16**:223-30

Moore EE *et al.*, 1981. Penetrating Abdominal Trauma Index. *J Trauma*;**21**:439-45

Mortensen N, 1989. Wide variations in surgical mortality. *BMJ*;**298**:344-5

Moser K-H, Bouillon B, Troidl H, Koppen L, 1989. Validation of the Continuous APACHE Score (CAPS) for a better prediction of outcome in surgical ICU patients. *Theor Surg*;**3**:192-7

Moses LE, Mosteller F, 1968. Institutional differences in postoperative death rates. *JAMA*;**203**:150-2

Murray GD, 1977. A Cautionary Note on Selection of Variables in Discriminant Analysis. *J R Statist Soc C(Applied Statistics).*;**26**:246-50

Murray GD, 1986. Use of an international data bank to compare outcome following severe head injury in different centres. *Statistics in Medicine*;**5**:103-12

Murray GD, Murray LS, Barlow P, Teasdale GM, Jennett WB, 1986. Assessing the performance and clinical impact of a computerised prognostic system in severe head injury. *Statistics in Medicine*;**5**:403-10

Murray LS, Teasdale GM, Murray GD, Jennett B, Miller JD, Pickard JD, Shaw MDM, Achilles J, Bailey S, Jones P, Kelly D, Lacey J, 1993. Does prediction of outcome alter patient management? *Lancet*;**341**:1487-91

Murray GD, Hayes C, Fowler S, Dunn DC, 1995. Presentation of Comparative Audit Data. *Br J Surg*;**82**:329-32

Neugebauer E, Troidl H, Spangenberger W, Dietrich A, Lefering R and the Cholecystectomy Study Group, 1991. Conventional *versus* laparoscopic cholecystectomy and the randomized controlled trial. *Br J Surg*;**78**:150-4

Nixon, SJ, 1990. Defining essential hospital data. *BMJ*;**300**:380-1

Nixon SJ, 1992. Does audit result in change of practice? The Lothian surgical experience. *Quality in Health Care*;1 supp:S25-S27

Oguz M, Sayar A, Yalin R, 1990. Preoperative prediction of postoperative complications*Isr J Med Sci*;**26**:147-9

Ottow RD *et al.*, 1984. Clinical judgement versus delayed hypersensitivity skin testing for the prediction of postoperative sepsis and mortality. *Surg Gynecol Obstet.*;**159**:475-7

Pedersen T, Eliasen K, Henriksen K, 1990. A prospective study of mortality associated with anaesthesia and surgery: risk indicators of mortality in hospital. *Acta Anaesthesiol Scand*;**34**:176-82

Pettigrew RA, Hill GL, 1986. Indicators of surgical risk and clinical judgement. *Br J Surg.*;**73**:47-51

Pettigrew RA, Burns JG, Carter, 1987. Evaluating sugical risk: the importance of technical factors in determining outcome*Br J Surg.*;**74**:791-4

Pettigrew RA, McDonald JR, van Rij AM, 1991. Developing a system for surgical audit. *Aust NZ J Surg*;**61**:563-9

Playforth MJ, Smith GMR, Evans M, Pollock AV, 1987. Preoperative assessment of Fitness Score. *Br J Surg.*;**74**:890-2

Pollock AV, Evans M 1989. *Surgical Audit.* Butterworths

Pollock AV, 1993. Surgical evaluation at the crossroads. *Br J Surg*; **80**:964-6

Ramsay G, McGregor JR, Murray GD, Neithercut D, Ledingham IMcA, George WD, 1986. Prediction of surgical risk in adults. *Surg Res Comm.*;**3**:95-103

Rees JL, 1982. Accuracy of hospital activity analysis data in estimating the incidence of proximal femoral fracture. *BMJ*;**284**:1856-7

Ripley BD, 1994. Neural Networks and Related Methods for Classification. *J R Statist Soc B.*:**56**;409-437

Ruckley CV, 1984. Mechanisms of audit:discussion paper. *J Roy Soc Med*;**77**:40-44

Russell I, 1987. *Lecture Notes on Methods for Health Care Evaluation.* Health Services Research Unit, University of Aberdeen, Occasional Paper No. 1 (reprint).

Ryan JA, Taft DA, 1980. Preoperative nutritional assessment does not predict morbidity and mortality in abdominal operations. *Surg Forum*;**31**:96-8

Sagar PM, Hartley MN, Mancey-Jones B, Sedman PC, May J, MacFie J, 1994. Comparative Audit of Colorectal Resection with the POSSUM scoring system. *Br J Surg.*;**81**:1492-4

Saklad M, 1941. Grading of patients for surgical procedures. *Anesthesiology*;**2**:281-4

Sarmiento J *et al.*, 1991. Statistical modelling of prognostic indices for evaluation of critically ill patients. *Crit Care Med.*;**19**:867-70

Schackert HK *et al.*, 1986. The predictive role of delayed cutaneous hypersensitivity testing in postoperative complications *Surg Gynecol Obstet.*;**162**:563-8

Schein M, 1988. Acute surgical disease and scoring systems in daily surgical practice. *Br J Surg.*;**75**:731-2

Schwartz D, Flamant R, Lellouch J, 1980. *Clinical Trials*. Academic Press

Secretaries of State for Health, Wales, Northern Ireland and Scotland, 1989. *Working for Patients*. London:HMSO

Smith CW, 1994. Evaluating risk adjustment by partitioning variation in hospital mortality rates. *Statistics in Medicine*;**13**:1001-13

Shaw CD, 1980. Acceptability of Audit *BMJ*;**281**:1443-5

Titterington DM, Murray GD, Murray LS, Spiegelhalter DJ, Skene AM, Habbema JDF, Gelpke GJ, 1981. Comparison of Discrimination Techniques Applied to a Complex Data Set of Head Injured Patients. *J R Statist Soc A*;**144**:145-75

Toynbee P, 1991. Nervous Surgeons' Best Kept Secret: Their rate of success. *Daily Mail*: 9 November

Toynbee P, 1993. Opinion. *Radio Times*: 26 June;22

Urbach P, 1993. The value of randomization and control in clinical trials. *Statistics in Medicine*;**12**:1421-31

Vacanti CJ *et al.*, 1970. A statistical analysis of the relationship of physical status to postoperative mortality in 68,388 cases. *Anesth Analg*;**49**:564-6

Whates PD, Birzgalis AR, Irving M, 1982. Accuracy of hospital activity analysis operation codes. *BMJ*;**284**:1857-8

Whiteley MS, Prytherch D, Higgins B, Weaver PC, Prout WG, 1995. Comparative audit of colorectal resection with the POSSUM scoring system (letter to the editor). *Br J Surg.*;**82**:425

Wilkin A, McColl I, 1987. Surgical audit: the clinician's view. *Theor Surg*;**1**:195-206

Warnold I, Lundholm K, 1984. Clinical significance of preoperative nutritional status in 215 non-cancer patients. *Ann Surg.*;**199**:299-305

BIBLIOGRAPHY

BMDP Statistical Software, Inc. *BMDP Manual*, 1990.

Crombie IK, Davies HTO, Abraham SCS, du V Florey C. *The Audit Handbook.* Wiley, England, 1993.

Minitab Inc. *MINITAB* 9 *Reference Manual* , 1993

Pollock A, Evans M. *Surgical Audit.* Butterworths, London 1989.

Statistical Sciences, Inc. *S-PLUS Reference Manual, Version 3.2*, Seattle: StatSci, a division of MathSoft, Inc., 1993.