Prediction of perioperative mortality after oesophagectomy using the Northern Oesophagogastric Cancer Unit clinical database

Ian Warnell

Thesis submitted for the degree of Doctor of Medicine

Institute of Health and Society

Newcastle University

June 2012

Dedication

To Frances, Harry and Paddy

Declaration

I declare that this thesis was written by me and has not been submitted or accepted for any other degree. Except where specifically acknowledged, I have collected, analysed and interpreted all the data, and have acknowledged any work of others, in accordance with University and School guidance on good academic conduct.

20th June 2012

Ian Warnell

Abbreviations

Abbreviation	Full name
NOGCU	Northern Oesophagogastric Cancer Unit
NYCRIS	Northern and Yorkshire Cancer Registry and Information Service
ASA	American Society of Anesthesiologists
FEV1	Forced expiratory volume in 1 second
FVC	Forced vital capacity
FEV1/FVC	Ratio of FEV1 to FVC
RCRI	Revised cardiac risk index
MNAR	Missing not at random
MCAR	Missing completely at random
MAR	Missing at random
WBC	White blood cell count
pO2	Partial pressure of oxygen in arterial blood
pCO2	Partial pressure of carbon dioxide in arterial blood
Hb	Haemoglobin concentration
K	Serum potassium concentration
SEER	Surveillance Epidemiology and End Results
AUC	Area under curve
ROC	Receiver operating characteristic
POSSUM	Physiological and Operative Severity Score for the enUmeration of Mortality and Morbidity
ASCOT	Assessment of Stomach and Oesophageal Cancer Outcomes from Treatment (database)
RCRI	Revised Cardiac Risk Index
CPX	Cardiopulmonary exercise testing
HL	Hosmer Lemeshow
MET	resting metabolic rate equivalent

Abstract

Perioperative mortality after thoracoabdominal oesophagectomy for cancer is about 4%. Stratifying this risk may assist patients to make treatment choices, facilitate comparative audit, and enhance research. I aimed to explore prediction modelling of this risk, using the Northern Oesophagogastric Cancer Unit (NOGCU) database.

The first section is a systematic review of prediction models and candidate predictors from ‘high surgical volume’ centres. Three models were externally validated but overestimated higher risk mortality; discrimination was moderate. Two groups used prediction models to reduce mortality in practise but there were no clinical impact studies. Candidate predictor definitions and associations with mortality were varied. Age predicts mortality and should be included as a continuous predictor in any model. Risk of bias in primary studies was poorly reported.

In section two, I explored the risk of perioperative mortality using logistic regression on 1575 records from the NOGCU database, from 1991 to 2009. Comorbidity fields required extensive cleaning and recoding, and there were variable amounts of missing data, which caused spurious associations. I compared a prespecified model containing age, operation, albumen and cardiorespiratory comorbidity with a statistical stepwise elimination model and used split-sample validation. Age, gender, operation, white cell count, cardiac risk index, operation and weight loss were associated with mortality but only age, gender, operation and weight loss were significant in multivariate analysis. Discrimination was moderate, at best, for all models and the prediction range was only to a maximum 20%. The best calibrated models contained age, operation and gender, and originated from the most complete datasets.

These models are not suitable for individual risk prediction but could be developed as risk adjusters for provider profiling and research. The sample sizes and high quality data required for further development are most likely to be achieved in larger scale studies, data syntheses or clinical databases.

Table of Contents

List of Tables

List of figures

Preface

The work for this thesis was carried out during my long term involvement as a consultant in anaesthesia and intensive care working with the Northern Oesophago-Gastric Cancer Unit (NOGCU). My interest in perioperative risk arose from observing the serious complication rate of oesophagectomy. Despite our advances in medicine it is impossible to not be struck by the enormous impact this operation has on patient's lives. If there was an alternative viable treatment, patients would surely take it.

I became interested in clinical prediction models as a way of perhaps identifying patients whose risk of surgery was so great that, given reliable information on their likely outcome of surgery, they might wish to choose an alternative treatment. This coincided fortuitously with access to our clinical database, set up in 1990 by Professor Griffin when he first established the Northern Oesophago-Gastric Cancer unit. I decided to try and find out whether nearly two decades of information might offer some answers. I carried out this project under the auspices of the Institute of Health and Society at Newcastle University, because of their expertise in a range of methods used in evidence based medicine. This thesis is the end result.

Acknowledgements

I would like to acknowledge and thank the following colleagues, friends and family for their help:

- Professor Michael Griffin, Mr Nick Hayes, Mr Daya Karat, and Mr Jon Shenfine for allowing me to use the data from the NOGCU clinical database.
- Nick Hayes for providing background information and sources about the management of oesophageal cancer.
- Dr Ian Shaw for his extensive collection of references, advice and for proof reading parts of the thesis.
- Sharon Johnson and Helen Jaretzke, database managers for the Northern Oesophagogastric Cancer Unit clinical database, for their time and help in retrieving and organising data.
- Dr Mahindra Chincholkar, my co-systematic reviewer.
- Erika Gwynett (librarian for the Walton library) for help structuring the search strategy for the systematic review.
- Professor Ewout Steyerberg for help in applying his ‘custom’ ‘code in R for model validation.
- I would like to thank my supervisors Professor Martin Eccles and Nick Steen from the Institute of Health and Society for their invaluable help and advice with all aspects of this thesis. I particularly appreciate their flexibility in providing supervision to fit around illness and my full time work.
- Finally, thank you to Frances, my wife, for proof reading, unstinting support and putting up with all this.

Chapter 1: Introduction

In this thesis I have set out to investigate whether it is possible to use the Northern Oesophago-Gastric Cancer Unit (NOGCU) clinical database to develop an effective clinical prediction model for perioperative mortality after thoracoabdominal oesophagectomy. The ability to reliably predict serious complications of high risk surgery, which may have uncertain success, may help patients and clinicians to weigh the risks and benefits and make informed choices about treatment. It can also enhance comparison of outcome performance between different centres by adjusting for important risk factors. Risk stratification may also enhance experimental and diagnostic research by selecting patients who are likely to benefit most from the intervention (Hernández *et al.*, 2004; Steyerberg, 2009a). For example, 'goal directed fluid therapies' (Abbas and Hill, 2008) to improve outcome and 'cardiopulmonary exercise testing' (Older *et al.*, 1999) for risk stratification, are currently enthusiastically supported by some groups, but their role remains controversial (Moonesinghe *et al.*, 2011). In prospective studies of these interventions, risk stratification may improve study design and clarify their role (Hernández *et al.*, 2004).

The Northern Oesophago-Gastric Cancer Unit is a regional centre for the treatment of stomach and oesophageal cancer. At its inception in 1990 an integrated clinical database was set up, which has evolved and now contains extensive clinical information about NOGCU patients, who have received treatment. For each patient it contains demographic information, tumour details, comorbidities, treatment and complications. The NOGCU database is used for clinical audit and has been used to report the outcomes and univariate associations of comorbidity with mortality in 228 patients (Griffin *et al.*, 2002), but has not hitherto been used to study clinical prediction models of perioperative mortality or morbidity.

1.1 Epidemiology of oesophageal cancer

The pattern of oesophageal cancer in the United Kingdom (UK) has changed over recent decades. The incidence has increased from 6.5 to 9.8 per 100,000 of the population, between 1975 and 2008 (Cancer Research UK), and is the ninth most common cancer accounting for 3% of all UK cancer. In 2008 there were 8,173 new cases in the UK (Cancer Research UK). Most striking has been the increase amongst UK males, where the incidence has increased from 8.8 to 14.5 per 100,000. The male to female ratio has increased to between 5 and 10 to 1, and in 2003 UK men were reported to have the highest incidence in Europe (Wild and Hardie, 2003).

The pattern of tumour characteristics in the UK has also changed. Until the early 1990's, squamous cell cancer was most common (Powell *et al.*, 2002). This is typically situated in the middle and upper oesophagus, and is associated with low socioeconomic status, poor nutrition, alcohol intake and smoking. It is still the commonest histological type in many developing areas, such as Asia and Southern Africa. However, in the last twenty years oesophageal adenocarcinoma has become predominant in the UK (Vizcaino *et al.*, 2002). Adenocarcinoma occurs in the lower oesophagus or oesophagogastric junction and is not associated with patient socioeconomic status.

1.2 Treatment options

The options for treatment are primarily defined by tumour stage (Griffin, 2009), with curative surgical resection considered the definitive treatment in early tumours (Wu and Posner, 2003). Chemo- or chemo-radiotherapy, with or without surgery, is more common in more advanced tumours and palliative treatments are used in the very advanced stages. Sadly most tumours are not amenable to surgical resection when they are detected, and currently only about 25% of patients undergo attempted curative surgery (Rouvelas *et al.*, 2005; Al-Sarira *et al.*, 2007). Within these broad categories of treatments there is considerable debate as to which methods produce the best outcomes. The introduction of newer and less invasive techniques may be an attractive option if outcomes match those of current surgical resection methods. For instance, endoscopic tumour resection or photodynamic therapy may be used for both early tumour treatment and advanced palliation, and minimally invasive surgical techniques are increasingly under investigation, as are chemo- and radiotherapy regimes (Allum *et al.*, 2011). With no apparent reduction in the incidence of oesophageal cancer and newer treatments coming on line, outcome information is likely to play an important role in clinical decision making, audit and research.

1.3 Which outcome to investigate?

Treatments for oesophageal cancer incur considerable morbidity, mortality and uncertain success rates and therefore a range of potential categories of outcome are important. Clearly the likelihood of complete cure and long term survival will be a prime concern. If major surgery can reasonably guarantee prolonged survival at little risk, for instance for a very early cancer, this is likely to be the treatment of choice for many people. However, in more advanced cancers when long term survival cannot be guaranteed, other considerations may be important. For example, oesophagectomy seems to diminish quality of life in patients who do not survive more than two years (Blazeby *et al.*, 2000), and to a greater degree than chemo-radiotherapy, which has similar survival rates for locally advanced cancer (Avery *et al.*, 2007). Quality of life measures such as the validated EORTC QLQ-C30 and QLQ-OES18 (Blazeby *et al.*, 2003) were not routinely recorded in the NOGCU database, and therefore this was not an option in this work.

Perioperative morbidity (defined as any clinically significant nonfatal complication (Moonesinghe *et al.*, 2011)) has a much higher incidence than mortality. It is associated with prolonged length of hospital and critical care stay, increased use of resources and is a predictor of mortality (Moonesinghe *et al.*, 2011). It is also increasingly associated with reduced medium term survival after major abdominal surgery (Khuri *et al.*, 2005; Schiesser *et al.*, 2008). A standardised morbidity score, the Postoperative Morbidity Survey (POMS) has been developed and validated in major abdominal surgery (Bennett-Guerrero *et al.*, 1999; Grocott *et al.*, 2007), but most studies report morbidity in a wide variety of ways. This inconsistency, and the potential need to recode and interpret the data retrospectively, gives considerable scope for misinformation bias in the outcome. I therefore chose not to study this outcome.

Perioperative mortality after oesophagectomy is considerable compared with many other procedures, and is therefore likely to remain important when considering treatment options or comparing providers. It is a much 'cleaner' outcome, and 'all cause' mortality should be free of interpretational bias, as no cause needs to be attributed to it, removing the need to 'blind' the data reporter. However, it is statistically uncommon and therefore makes identifying reliable predictors more difficult than, for instance, for the more frequent morbidity. It is important to specify the time period during which mortality is reported. For example, 30 day mortality is frequently used to enable standardised comparisons to be made. However, advances in perioperative care mean that deaths associated with the procedure frequently occur after this period (Griffin *et al.*, 2002; Cromwell *et al.*, 2010), and therefore do not necessarily convey the full nature of the procedure. Longer defined periods of follow up (e.g. 90 days) may include out of hospital deaths, which will inevitably be more difficult to account for (Moonesinghe *et al.*, 2011). Confining the outcome to deaths recorded in hospital is more likely to be reliable as follow up should be easier. Because of these factors I have chosen to study 'all cause', in hospital perioperative mortality associated with the primary tumour resection.

1.4 Perioperative mortality after oesophagectomy

Perioperative mortality from oesophagectomy has steadily decreased from 72% in 1941(Ochsner and DeBakey, 1941), to 29% between 1960 and 1979 (Earlam and Cunha-Melo, 1980), 13% between 1980 and 1988 (Muller *et al.*, 1990), and 6.7% between 1990 and 2000 (Jamieson *et al.*, 2004) This trend has been observed in the United Kingdom (Al-Sarira *et al.*, 2007), the United States (Hofstetter *et al.*, 2002; Dimick *et al.*, 2005b), Sweden (Rouvelas *et al.*, 2005), France (Sauvanet *et al.*, 2005) and the Far East (Jamieson *et al.*, 2004). Developments in the surgical and perioperative management of oesophageal cancer are likely to have contributed to these improvements in the last four decades. These have included improved patient selection and tumour staging, perioperative nutritional support, improvements in anaesthetic techniques and the concentration of skills in "large volume" centres. Despite this progress, the perioperative mortality rate is still daunting with most centres currently reporting in-hospital mortality rates of about 5% (McCulloch *et al.*, 2003; Steyerberg *et al.*, 2006; Cromwell *et al.*, 2010). Overall complication rates for oesophagectomy can reach 60% and in-hospital mortality up to 14 % (McCulloch *et al.*, 2003).

The causes of perioperative mortality are typically associated with anastomotic breakdown, necrosis of the gastric remnant, or respiratory and cardiovascular complications, which may be primary or secondary to surgical complications (Law *et al.*, 1994; Griffin *et al.*, 2002; Law *et al.*, 2004). Surgical mechanical failure is a major cause of death, which it would seem, is unlikely to be predicted by preoperative comorbidity. However, nutritional state and general health may cause impaired healing (Law *et al.*, 1973; Fekete and Belghiti, 1988) as may the tissue hypoxia that can result from cardiorespiratory impairment. Perhaps a proportion of anastomotic or gastric breakdown may be predictable.

Transthoracic oesophagectomy involves chest wall surgery, prolonged operating time, one lung ventilation, mechanical retraction of lung tissue, thoracic lymphadenectomy and potentially large and complex body fluid shifts. With these intraoperative physical insults to the chest wall and lung, it is unsurprising that respiratory complications are frequent and serious, with reported rates up to 32% (Law *et al.*, 1994; Whooley *et al.*, 2001; Law *et al.*, 2004). Respiratory complications are also a major cause of mortality, for example contributing up to 55% of all perioperative fatalities (Whooley *et al.*, 2001; Law *et al.*, 2004). The incidence of cardiorespiratory complications would suggest that comorbidity of these systems may predict outcome.

1.5 Studying prediction of perioperative mortality

There are three main steps to predicting prognosis. Firstly the development of a suitable prediction model, secondly its validation and finally assessing its clinical impact (Moons *et al.*, 2009b). In this project I will be using data collected prospectively from a cohort of patients who have undergone oesophagectomy, but the data will necessarily be analysed retrospectively. This gives scope for model development and some degree of validation, but external validation and clinical impact studies are beyond the scope of this project.

Patient variability and heterogeneous causes of mortality mean that single predictors are unlikely to effectively predict outcome. Multivariable models are likely to be more effective and therefore selection of candidate predictors will be an important initial step in prediction (Moons *et al.*, 2009b). Perioperative mortality after oesophagectomy has been associated with a variety of predictors including age, tumour stage, pulmonary dysfunction, impaired general health, smoking, diabetes, cardiac dysfunction and hepatic dysfunction (Pennefather, 2007). All these candidate predictors are represented in some form in the NOGCU database and will be considered. The overall dataset is about 1576 cases of gastric and oesophageal surgery with about 87 deaths. This is a relatively small dataset so in an ideal situation we would be able to use existing information to select predictors with known 'weights' and apply these directly to our dataset. We could then update and recalibrate the model to suit our population (Steyerberg, 2009i). This would require prior knowledge of which predictors to include, together with their form and magnitude. In the absence of this information, predictor exploration and selection would become important.

A range of other multivariable models to predict mortality have already been described (Shende *et al.*, 2007), for instance the 'Physiological and Operative Severity Score for the enumeration of Mortality and Morbidity', known as POSSUM (Copeland *et al.*, 1991; Prytherch *et al.*, 1998; Tekkis *et al.*, 2004) and the 'Rotterdam' model (Steyerberg *et al.*, 2006). It is possible that any of these may be applicable and perform acceptably for our data. However, there are no systematic reviews, which address the applicability of these prediction models, or supply the prerequisite knowledge to enable selection of candidate predictor for model development; therefore it will be one of my objectives to use systematic review methods to supply this information.

An important step before modelling the outcome is to prepare a suitable dataset and select appropriate predictors. This includes evaluating data quality, the extent of missing values and determining how important predictors are to be handled (Royston *et al.*, 2009). After selecting a set of candidate predictors I will use standard recognised methods to select and investigate the performance of potential clinical prediction models. These methods have been described by Steyerberg (Steyerberg, 2009e) and by Moons, Royston, Altman and colleagues (Altman *et al.*, 2009; Moons *et al.*, 2009a; Moons *et al.*, 2009b; Royston *et al.*, 2009).

For a prognostic prediction model to be accepted in clinical practise it should be reliable and "transportable" to new patient groups. Reliability is assessed by validation procedures, which include calibration (a measure of accuracy) and discrimination (a measure of whether the model can allocate correct outcome between different risk groups). A patient, who is deciding whether to undergo a major procedure, requires accurate estimates for outcomes; this is calibration. If the problem is deciding whether to allocate further diagnostic stratification tests, identifying high or low risk groups may be important; this is discrimination (Steyerberg, 2009b). Another important aspect of prediction models is whether their predictors are easily "transportable" to new patient groups.

1.6 Summary of the aims and objectives of the thesis

1.6.1 Aim of thesis

To study a clinical prediction model of perioperative mortality after oesophagectomy based on data from the NOGCU clinical database, and to consider its potential for application and further development.

1.6.2 Objectives of the thesis

1. To carry out an original systematic review to:

i. Assess existing clinical prediction models of perioperative mortality after oesophagectomy, which could be used for risk stratification in patients of a 'high volume' unit in the United Kingdom.

ii. Identify candidate pre-operative predictors, which should be considered for inclusion in any such prediction model, and if possible to estimate their effects on perioperative mortality.

2. To develop and internally validate a clinical prediction model for perioperative mortality after oesophagectomy, using data from the NOGCU clinical database and to consider further development.

1.7 Outline structure of thesis

In Chapter 2 I use systematic review methodology to identify and assess the performance and applicability of existing clinical prediction models. I also use these methods to identify and, where possible, quantify the effects of candidate predictors, which may be considered for inclusion in the prediction model, which I intend to study.

In Chapter 3 I report an investigation of the data contained in the NOGCU database and attempt to prepare a suitable dataset, with which to explore a clinical prediction model. In particular I investigate data quality, missing data, and the structure of relevant predictors.

In Chapter 4 I use logistic regression to explore a clinical prediction model of perioperative mortality using a subset of data from the NOGCU clinical database. The model will be developed and validated using split sample design.

Finally, in Chapter 5, the general discussion, I report a narrative summary of the work carried out in this thesis and discuss the potential applications of suitable prediction models. I will go on to discuss potential areas for further development of such models for oesophageal surgery.

The appendices will contain supplementary information and are listed in the Table of Contents.

Chapter 2: A systematic review of clinical prediction models and their candidate predictors for perioperative mortality after oesophagectomy

2.1 Introduction

The first step in this project is to identify existing clinical prediction models of perioperative mortality, and to determine whether they can be validated on the NOGCU database and subsequently applied to prospective patients, who present to this unit. The second step is to identify from published studies, which preoperative predictors should be considered for inclusion in any future clinical prediction model, and whether their effects can be quantitatively estimated from data synthesis. There were no published systematic reviews to answer these questions and therefore I decided this would be an appropriate starting point.

Unlike studies of therapeutic interventions, the methods for systematic reviewing for clinical prediction are less well developed. Consequently, I constructed the review methods from several sources. These included generic methodological recommendations for systematic reviews (Centre for Reviews and Dissemination, 2009) and recommendations of the Ottawa Methods Centre for reporting them ('Preferred reporting items for systematic reviews and metanalyses: the PRISMA statement' (Moher *et al.*, 2009)). I also drew on recommendations for primary prognostic study methods (Altman and Lyman, 1998), the reporting of systematic reviews of prognostic studies (Altman and Riley, 2005), metanalyses of observational studies (Stroup *et al.*, 2000), and recommendations for the assessment of the potential risk of bias in systematic reviews of prognostic studies (Hayden *et al.*, 2006). There are no specific Cochrane guidelines for prognostic reviews, but there is a Cochrane group, the Cochrane Prognostic Review Network (Cochrane Prognosis Methods Group, 2011), which is developing the methodology, and with whom I consulted.

I intended to apply lessons from this review to patients from our United Kingdom regional centre and therefore the inclusion criteria for primary studies reflected two important characteristics of our centre. Firstly, the

NOGCU carries out at least fifty subtotal oesophagectomies annually, and can reasonably be classified as a ‘high volume’ centre, so studies were only included if the reporting centre performed at least 10 procedures per year (Killeen *et al.*, 2005). Studies using population or multicentre databases were also included as overall effects might be more generally applicable. Secondly, the data collection for our clinical database started in 1991; I therefore only included primary studies reported after 1990. This matches our data collection period and allows comparison of effects from periods of similar perioperative mortality rates.

2.1.1 Aims and objectives of the systematic review

Aims

1. The first aim is to identify clinical prediction models of perioperative mortality after oesophagectomy, which we could potentially validate on the NOGCU database and use for patients in a ‘high volume’ oesophagogastric cancer unit in the United Kingdom.

2. The second aim is to clarify which individual predictors that were routinely collected pre-operatively, should be considered for inclusion in a prediction model, and whether their estimated effects are known well enough to incorporate from the outset.

Objectives

1. To identify studies of prediction models and of individual predictors for perioperative mortality after oesophagectomy for cancer, which were carried out in ‘high volume’ surgical centres, or reported from multicentre studies or population databases after 1990.

2. To report clinical prediction model reliability and ‘transportability’ to other populations.

3. To report which individual predictors have been studied, their definitions and descriptions, and their effects on mortality. Consideration will also be given to including the summary effects in a quantitative data synthesis.

4. To report potential for risk of bias within primary studies.

2.2 Systematic Review Methods

I used the checklist of items from the PRISMA checklist (Moher *et al.*, 2009) as a template for the reporting of this systematic review. Typically systematic reviews structure research questions by defining 'concepts' such as the population of interest (P), the intervention (I), the outcome comparison (C) and the study design (S), frequently abbreviated to 'PICOS'. Studies of clinical prediction models do not neatly fit this structure, so I have defined four 'concepts' to define the inclusion criteria for this review.

2.2.1 Inclusion criteria

Study Population

1. The population of interest is adults undergoing elective oesophagectomy for oesophageal cancer. The primary studies should focus on, or contain an easily identifiable subgroup of patients, who underwent oesophageal cancer surgery. Oesophageal cancer resections in our centre were almost exclusively thoracoabdominal procedures and therefore only studies focussing on thoracoabdominal procedures were considered.
2. I aimed to use the results of this review to inform a study on the NOGCU clinical database, which started to collect data in 1990, therefore I selected articles published or carried out, in or after 1990.
3. Outcomes from complex major surgery, such as oesophagectomy, may be better in hospitals where larger volumes are carried out. This is because patient assessment, surgical skills and supporting services (radiology, anaesthesia, critical care, and nursing) are concentrated in fewer hands enabling them to improve through experience and clinical audit. There is evidence to support this both generally (Killeen *et al.*, 2005), and for specific geographical regions, for example the UK (Bachmann *et al.*, 2002; Al-Sarira *et al.*, 2007), the USA (Birkmeyer *et al.*, 2002; Allareddy *et al.*, 2007), and Sweden (Rouvelas *et al.*, 2005). This effect is complex because of potential confounding by the caseload of individual surgeons (Birkmeyer *et al.*, 2003; Dimick *et al.*, 2005a; Migliore *et al.*, 2007),

teaching hospital status (Dimick *et al.*, 2004; Verhoef *et al.*, 2007) and the use of small samples to compare hospitals (Dimick *et al.*, 2004). Despite this controversy, the weight of opinion seems to favour this view and therefore I only selected studies from 'high volume centres'. Defining a 'high volume' is difficult. Killeen (Killeen *et al.*, 2005) reviewed studies addressing this issue and the primary studies variously described 'low volume' as 2 to 13 and 'high volume' as 6 to 83. The investigators calculated the number of operations required by a 'high volume' centre needed to reduce perioperative mortality by 1 instance per year. In the case of oesophagectomy this appeared to be about 8 or 9 operated cases per year. I arbitrarily defined 'high volume' as enough cases to produce this annual reduction and included only reports from centres, which carried out at least 10 procedures per year.

Perioperative clinical outcomes

Articles were considered if they specified 'all cause' mortality associated with the hospital admission for the main surgical procedure, and a specified time period, e.g. 'in-hospital' or '30-day' mortality.

Study design

Observational or randomised studies (including cohort, clinical database, prospective or retrospective studies), which attempted to develop clinical prediction models, or estimate the effects of preoperative predictors on perioperative mortality were considered.

Prognostic predictors

For the purposes of the searches, 'prognostic predictor' included any individual preoperative predictor of perioperative mortality, and any clinical prediction model (combining more than one predictor). These general concepts were mapped to search terms, which included general terms (risk assessment, risk assessment tools) as well as more specific ones (e.g. cardiovascular comorbidity and its methods of assessment). Only articles which considered methods likely to be routinely available preoperatively, were considered.

2.2.2 Exclusion criteria

1. Studies of laparoscopic, thoracoscopic, minimally invasive and endoscopic procedures.

2. Studies mainly carried out, or published, before 1990 (see main introduction for discussion of mortality rates over time).

2. Studies carried out in centres where less than 10 cases per year on average were estimated to have been performed over the study period.

3. Studies were confined to English language reports because this was likely to constitute the largest reading and reporting audience.

2.2.3 Search strategy

Electronic databases

The search strategy was developed and carried out using Ovid Technologies, initially through the British Medical Association 'Medline Plus', but subsequently using Newcastle University Library Ovid Technology. Both accessed the Medline and Embase databases produced by the National Library of Medicine in the USA.

Search terms

The general concepts, which described the inclusion criteria, were specified and further mapped to specific terms using Collins Thesaurus, relevant journal articles, OVID Medline Medical Subject Heading (MeSH) mapping and its' permuted index function (Appendix A.). Searching the literature for studies on prognosis is more complex than for therapeutic interventions because of the wide range of study designs, the variety of synonyms for prognostic and observational studies, and a lack of standardised methodology. Therefore, I also incorporated other validated 'filters' (Appendix B.), which have been used in this type of study, and I also consulted with Erika Gwynnett (Newcastle University Walton Librarian), to facilitate search strategy development. The original 'concepts', which defined the inclusion criteria, were combined to produce search output, which was focussed on the defined population, but which had high sensitivity to include as many types of prediction and prognostic study as possible. The concepts were combined using logical operators as follows:

'Population' **AND** 'study type' **AND** ('prognostic predictor' **OR** 'clinical outcome').The search was run initially in April 2009, and updated on 18/09/2010 using the Ovid 'Autoalert' for 'selective dissemination of information'. The full search strategy is listed in Appendix C.

Alterations to search strategy during or after the searches

After initial searches, it was clear that several studies (known to myself) were missing and so the search strategy was modified as below:

1. The search term "Ivor adj Lewis" was added.

2. The "P" in "Possum" was capitalised

3. The text word oesophagectomy was shortened to 'oesoph' to retrieve any term with this root.

Selection of articles from electronic databases

I screened article titles to identify potentially relevant articles. Two reviewers, IW & Mahindra Chincholkar (anaesthesia specialist registrar; MC) screened titles and abstracts from this subset and examined full text versions of selected articles for inclusion criteria.

Other search sources

Hand searches were made of reference lists from primary research studies, review articles (Pennefather, 2007; Shende *et al.*, 2007), standard texts (Shaw, 2008), and personal collections of articles (IW and Dr I Shaw, consultant anaesthetists in the NOGCU, Royal Victoria Infirmary, Newcastle upon Tyne). The article selection process is summarised in Figure 1.

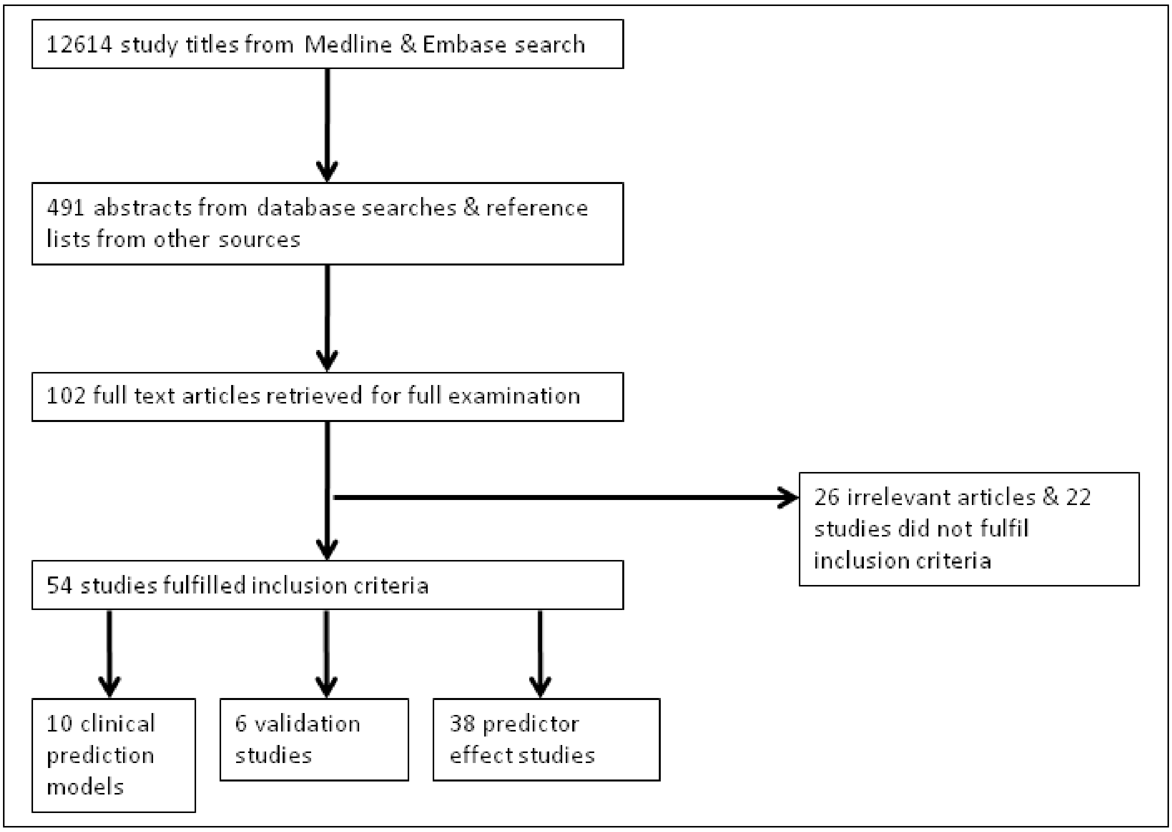

12614 study titles from Medline & Embase search
491 abstracts from database searches & reference lists from other sources
102 full text articles retrieved for full examination
26 irrelevant articles & 22 studies did not fulfil inclusion criteria
54 studies fulfilled inclusion criteria
10 clinical prediction models
6 validation studies
38 predictor effect studies

Figure 1 Selection process for included articles

2.2.4 Data extraction

The data items, which I intended to extract from each study are summarised in (Table 1).

Table 1 Data items for extraction from primary studies

Main data category	Data item
Study description	Author, publication date, period of data collection, study design, geographical location, number/type of centre or database,
Characteristics of study sample	Sample size, mean annual operative volume for study period, male/female ratio, tumour histology incidence, use of neoadjuvant therapy, surgical procedure, definition of perioperative mortality, 'hospital mortality' rate for study sample
Predictors investigated	Description & definition, of predictors and their effect on perioperative mortality
Performance of clinical prediction models	Modelling method, model fitting, calibration, discrimination, observed effect of clinical application in practise

I based assessment for potential bias on recommendations for systematic reviews of prognostic studies (Hayden *et al.*, 2006) and primary prognostic study design (Altman and Lyman, 1998; Altman and Riley, 2005). I adapted these for the prediction of perioperative mortality in oesophageal cancer resection and they are listed in Table 2 on the following page. The scoring criteria for individual items are explained in the table and are: M, fully met; P, partially met; N, not met; U, unclear; na, not applicable.

Table 2 Items to evaluate risk of bias in primary studies adapted from Hayden (Hayden et al., 2006)

Main category of potential bias	Items to consider in assessing potential for bias	Scoring method
The sample adequately represents the population of interest	Patients who were eligible for surgery but excluded are described and reasons given (e.g. surgical reasons or unfitness)	Reported, described and quantified, M; incomplete report e.g. surgical exclusions only, not quantified, P; not reported, N; unclear, U
	The sample includes all patients undergoing oesophagectomy during the stated period	Evidence that all oesophagectomies were included in the sample (e.g. statement that consecutive cases were included), M; excluded oesophagectomies from sample described, P; otherwise U #
	Sample key characteristics are described adequately including gender distribution, tumour histology, surgical procedure, neoadjuvant therapy, surgical operative volume, geographical location, period of study, study type, overall study mortality rate	All characteristics described, M; partially described, P; not described, N; otherwise unclear, U
The data represents the sample	Follow up rate is reported and acceptable (Kristman *et al.*, 2004)	Number of survivors and fatalities stated, with no losses to follow up, or evidence that losses are MAR or MCAR§, or less than 5% of sample, M; follow up rate is deducible from article, P; unreported or unclear or unknown, U
	Patients lost to follow up differ in characteristics from the sample	Characteristics of patients lost to follow up reported, M; not reported or missing not stated, unclear, or unknown U.
	Prospective or retrospective data collection	Prospective, P; retrospective, R; unclear or unknown, U

Main category of potential bias	Items to consider in assessing potential for bias	Scoring method
The data represents the sample	Evidence of data validation	Data audit or double entry described, M; partial validation e.g. data cross checked with more than one database, P; not stated or not done, N; unclear, U
	Missing values reported	All missing values reported, M; missing value quantities possibly deducible from tables or partially stated, P; no report or unclear, U;
	Description of missing value procedures (Vach, 1997)	No missing values, values MAR or MCAR§, or acceptable missing value procedure reported and described for all relevant missing data, M; partial missing data handling procedure, e.g. some information given, P; no report or unclear, U
	Records with missing values differ from the rest of the sample in other characteristics	Characteristics and outcome of records with missing values compared with rest of sample, M; no report or unclear, U
Important prognostic factors adequately measured (age, gender, cardiovascular, respiratory, nutritional and immune status; activity capacity; tumour stage, histology, surgical procedure, neoadjuvant therapy)	Adequate description or definition of prognostic factor (e.g. “transportable” to another population)	All prognostic factors in prediction model clearly defined and/or described, M; some prognostic factors described or not fully “transportable”, P; prognostic factors not defined, N; unclear, U
	Valid measurement of prognostic factor.	All main prognostic factors measured appropriately, M; some factors measured appropriately, P; not measured appropriately, N; unclear U

Main category of potential bias	**Items to consider in assessing potential for bias**	**Scoring method**
Prognostic factors	Continuous variables used or otherwise handled appropriately	Continuous variables used, M; predefined cut points with rational basis, P; 'data-driven' or unhelpful cut points, N; unclear, U
Outcome is adequately measured	Clear definition of outcome (follow up 30 to 90 days or "in hospital" mortality).	Period of follow up to perioperative mortality clearly defined, M; deducible from text, P; not stated, N; unclear U
Potential known confounders of prognostic variables are accounted for (includes all important prognostic variables if single variable is investigated)	Important potential cofounders, if not investigated as prognostic variables are defined, measured and recorded	All important confounders defined, measured and recorded, M; some defined, measured and recorded, P; none recorded, N; unclear, U
	Important potential cofounders are included in study design or accounted for in the data analysis	Important confounders included in study design prospectively, or included in prognostic model, M; confounders tabulated to allow statistical analysis, P; not recorded, N; unclear, U
Appropriate data analysis	Description of appropriate statistical model	Selection of statistical model and variables is appropriate and based on conceptual model, M; inappropriate model, N; unclear, U
	Sufficient information given to assess adequacy of analysis	Adequate model description and presentation of appropriate results, including regression coefficients or equivalent & statistical significance, variable collinearity and interaction, & model testing, M; statistical model described but incomplete details, P; inadequate information or unclear , U

Main category of potential bias	Items to consider in assessing potential for bias	Scoring method
Appropriate data analysis	Adequate sample size; two sample size calculated with on line calculator (Type I error 5%, Type II error 20%)(Pezzullo, Updated May 2009); at least 10 events per predictor variable in linear regression	Sample size is large enough to detect statistically significant differences for clinically significant outcomes, M; sample too small, N; unclear, U

# If the description was, "we included 100 oesophagectomies in the study sample", this could have been a selected sub-group and therefore did not fully meet the criteria for no selection bias. Studies based on large population databases or where data was submitted from several centres were classified as not satisfying our criteria for clearly representing a defined population, because the process of case selection from multiple centres is unlikely to be reliably known.

§ Abbreviations: MAR, Missing at random; MCAR, Missing completely at random

The data was extracted into a data entry form and transferred into an Excel 2003 spreadsheet. Factual items were extracted by myself and checked independently by MC (second reviewer). Items, which addressed potential for bias in primary studies, were extracted independently by two reviewers (IW/MC) into an Excel 2003 spreadsheet. Discrepancies and disagreements were resolved by 'face to face' discussion. Most disagreements were due to unclear reporting of definitions, and difficulties finding relevant data in the studies. The final results were entered into the spreadsheet by IW.

Data synthesis

The criteria for attempting a quantitative data synthesis of the estimated effects of individual predictors were:

1. Whether definitions of the predictors and outcomes across candidate studies were consistent.

2. Whether the summary effects of predictors were reported in a way to allow a quantitative synthesis.

3. Whether the estimated potential for bias would support combination of summary effects.

2.3 Results

2.3.1 Organisation of results

1. Summary of included and excluded studies.
2. Description of included studies.
3. Geographical location of studies
4. Type of study centre and source of data
5. Size of study samples
6. Description of clinical prediction models.
7. Description of the effects of candidate predictors on mortality.
8. Potential risk of bias in primary studies
9. Table 12 Characteristics of studies fulfilling inclusion criteria (end of chapter).
10. Table 13 Excluded studies and reasons for exclusion (end of chapter).
11. Table 14 Studies of clinical prediction models (end of chapter).

2.3.2 Summary of included and excluded studies

At the time of the initial searches (5/12/2007), no relevant systematic reviews had been retrieved from Medline, Embase or Cochrane databases. Fifty four studies fulfilled the inclusion criteria and these are referenced in Table 12 at the end of the chapter. Excluded studies are listed in Table 13 at the end of the chapter. Reasons for exclusion were: unclear definition or follow up period for perioperative mortality (Bonavina *et al.*, 2003; Di Martino *et al.*, 2005; Alexiou *et al.*, 2006; Morgan *et al.*, 2007; Skipworth *et al.*, 2009), data collection before 1990 (Lund *et al.*, 1990; Charoenpan *et al.*, 1993; Gulliford *et al.*, 1993; Liedman *et al.*, 1995), outcome not clearly defined as perioperative mortality (Ferguson and Durkin, 2002; Mokart *et al.*, 2005; Jiao *et al.*, 2006; Baba *et al.*, 2008; Lagarde *et al.*, 2008; Wright *et al.*, 2009), no relationship or unclear perioperative mortality (Karl *et al.*, 2000; Nozoe *et al.*, 2002; Blazeby *et al.*, 2005b), operative volume less than our predefined inclusion criteria (Cariati *et al.*, 2002; Golubovi and Golubovi, 2002), expanded data from previous study (Bartels *et al.*, 2000), no identifiable group of oesophagectomies (Chamogeorgakis *et al.*, 2007).

2.3.3 Description of included studies

Ten studies developed clinical prediction models (Law *et al.*, 1994; Zhang *et al.*, 1994; Bartels *et al.*, 1998; Liu *et al.*, 2000; Bailey *et al.*, 2003; McCulloch *et al.*, 2003; Tekkis *et al.*, 2004; Sanz *et al.*, 2006; Ra *et al.*, 2008; Steyerberg, 2009a). Three studies evaluated existing prediction models (Zafirellis *et al.*, 2002; Schroder *et al.*, 2006; Lagarde *et al.*, 2007) and three compared and evaluated existing models (Lai *et al.*, 2007; Nagabhushan *et al.*, 2007; Zingg *et al.*, 2009). All studies, except evaluation and validation studies, investigated the effect of prognostic variables on 'in-hospital', '30 day' or time defined 'in hospital' mortality.

2.3.4 Geographical location of studies

Twenty six studies were based in Europe. Thirteen were in the United Kingdom(Adam *et al.*, 1996; Alexiou *et al.*, 1998; Griffin *et al.*, 2002; Zafirellis *et al.*, 2002, {Leigh, 2006 #60; Rahamim *et al.*, 2003; Tekkis *et al.*, 2004; Abunasra *et al.*, 2005; Alexiou *et al.*, 2005; Murray *et al.*, 2007; Nagabhushan *et al.*, 2007; Forshaw *et al.*, 2008)}, 3 in Germany (Bartels *et al.*, 1998; Gockel *et al.*, 2005; Schroder *et al.*, 2006), 3 in France (Thomas *et al.*, 1996; Jougon *et al.*, 1997; Sauvanet *et al.*, 2005), 2 in Italy (Ruol *et al.*, 2007(b)), 2 in the Netherlands (Han-Geurts *et al.*, 2006; Lagarde *et al.*, 2007), and one from each of Spain(Sanz *et al.*, 2006), Sweden(Johansson and Walther, 2000) and the Irish republic(Healy *et al.*, 2008). Nine were based in the USA (Ferguson *et al.*, 1997; Ellis Jr *et al.*, 1998; Sabel *et al.*, 2002; Bailey *et al.*, 2003; Rentz *et al.*, 2003; Atkins *et al.*, 2004; Moskovitz *et al.*, 2006; Finlayson *et al.*, 2007; Ra *et al.*, 2008), 5 in Hong Kong (Law *et al.*, 1994; Poon *et al.*, 1998; Whooley *et al.*, 2001; Law *et al.*, 2004; Lai *et al.*, 2007), 5 in Japan (Saito *et al.*, 1993; Zhang *et al.*, 1994; Kuwano *et al.*, 1998; Fang *et al.*, 2001; Kinugasa *et al.*, 2001), and one each from Australia (Liu *et al.*, 2000), and Taiwan(Tsai *et al.*, 2003). One was based jointly between the Netherlands and the USA (Steyerberg *et al.*, 2006).

2.3.5 Type of study centre and source of data

Most studies were from single centres, and a few were based on two or three centres. Two studies from the USA used data from the Department of Veteran Affairs National Surgical Quality Improvement Program, sampling 109 centres (Bailey *et al.*, 2003; Rentz *et al.*, 2003) and one used the Nationwide Inpatient Sample, a national database containing hospital discharge data on all paying patients (Finlayson *et al.*, 2007). Ra (Ra *et al.*, 2008) used data from the Surveillance, Epidemiology and End Results Program (SEER-Medicare) to identify patients with oesophageal cancer and linked this to the Medicare Provider Analysis and Review file to collect information about patients who had oesophagectomy. Steyerberg (Steyerberg *et al.*, 2006) used the SEER population database, a Netherlands registry and surgical centre in the Netherlands.

Of the European studies, Sauvanet (Sauvanet *et al.*, 2005) collected voluntarily submitted data from members of the French Association of Surgery in 37 centres. In the United Kingdom, one study used the Assessment of Stomach and Oesophageal Cancer Outcomes from Treatment (ASCOT) database, and the Risk Scoring Collaborative to collect data from 36 centres (Tekkis *et al.*, 2004), and the audit report from the database of the Association of Upper Gastrointestinal Surgeons of Great Britain and Ireland included data from 37 UK centres(Griffin *et al.*, 2002).In a study from Hong Kong, data from 14 Hospital Authority hospitals was collected through the Hospital Authorities Integrated Administration System and Central Management System (Lai *et al.*, 2007).

2.3.6 Size of study samples

In single centre studies sample sizes ranged from 32 (Liu *et al.*, 2000) to 785 (Tsai *et al.*, 2003) (median 382.5). In larger multicentre or population database studies, the sample size ranged from 538 (Tekkis *et al.*, 2004) to 27957 (Finlayson *et al.*, 2007) (median 1192). Estimated annual surgical volumes were available or deducible all individual units (sample size or operated cases averaged over the study period) and ranged from 9.17

(Nagabhushan *et al.*, 2007) to 63.16 (Atkins *et al.*, 2004) (median 37.51, excluding multiunit or population database studies). Male to female ratios were deducible or stated in all but 2 (Rentz *et al.*, 2003; Murray *et al.*, 2007) studies, and ranged from 1.9 (Nagabhushan *et al.*, 2007) to 110 (Bailey *et al.*, 2003) (median 3.6).

2.3.7 Clinical prediction models (Table 14)

Bailey (Bailey *et al.*, 2003), Ra (Ra *et al.*, 2008) and Steyerberg (Steyerberg, 2009a) developed prediction models using regression methods on data from USA population databases. Bailey (Bailey *et al.*, 2003) used the Veterans Affairs National Surgical Improvement Program. Ra (Ra *et al.*, 2008) and Steyerberg (Steyerberg *et al.*, 2006) used records from the Surveillance, Epidemiology and End Results (SEER) Medicare database. Tekkis (Tekkis *et al.*, 2004) developed a POSSUM score specifically for oesophagogastric surgery (O-POSSUM) from UK clinical databases and McCulloch developed a clinical prediction model from a subset of the ASCOT database (McCulloch *et al.*, 2003). Bartels (Bartels *et al.*, 1998), Law (Law *et al.*, 1994), Liu (Liu *et al.*, 2000), Sanz (Sanz *et al.*, 2006) and Zhang (Zhang *et al.*, 1994) modelled mortality on data from their own clinical centres.

Steyerberg (Steyerberg *et al.*, 2006) validated the prediction model using bootstrap methods on the modelling sample and applied the model to a SEER cohort from a subsequent period, and also to cohorts from a Netherlands population database and clinical centre. A simple scoring system was developed to predict 30 day mortality, which included age, comorbidity count, type of neoadjuvant therapy, and hospital surgical volume mortality. Discrimination was reported as poor (receiver operator AUC 0.56-0.7) but calibration was described as excellent for SEER patients and pooled data, but reported as “problematic” when applied to cohorts from the Netherlands.

Bartels and Zhang (Zhang *et al.*, 1994; Bartels *et al.*, 1998) validated their models on prospective samples from their own centres. Bartels used multivariate and discriminant analysis to associate degrees of organ dysfunction with mortality, and created a risk score. Similar mortality rates were observed in high risk groups in modelling and validation samples. Zhang used multivariate regression methods to develop a risk score, which had similar specificities in modelling, and validation samples, but whose sensitivity deteriorated considerably.

Tekkis (Tekkis *et al.*, 2004) developed the O-POSSUM on 70% of a randomly split sample, and validated on 30%. The O-POSSUM fitted the data well and discriminated well (C-index was 74.6%). It also compared favourably with the P-POSSUM, which overestimated mortality by about 20%.

Studies which validated models on development samples (apparent internal validation (Steyerberg *et al.*, 2006)) reported that model fit and discrimination was acceptable (Law *et al.*, 1994; Bailey *et al.*, 2003; Ra *et al.*, 2008). Liu and Sanz did not report formal validation procedures (Liu *et al.*, 2000; Sanz *et al.*, 2006).

Four studies compared and externally validated POSSUM models in oesophagogastric surgery (Zafirellis *et al.*, 2002; Lagarde *et al.*, 2007; Lai *et al.*, 2007; Nagabhushan *et al.*, 2007). Discrimination and calibration were poor for the original POSSUM and O-POSSUM (Zafirellis *et al.*, 2002; Lagarde *et al.*, 2007). The P-POSSUM performed reasonably in a comparison with O- and the original POSSUM (Lai *et al.*, 2007), but poorly in other comparisons with the O-POSSUM (Tekkis *et al.*, 2004; Nagabhushan *et al.*, 2007). Overestimation of predicted mortality was common (Lagarde *et al.*, 2007; Lai *et al.*, 2007; Nagabhushan *et al.*, 2007).

Schroder (Schroder *et al.*, 2006) evaluated Bartels' (Bartels *et al.*, 1998) model prospectively on 126 patients. Discrimination and calibration were not formally tested but sensitivity and specificity were deducible from the results. Schroder's predicted 'high' risk group had 16.7% mortality, considerably lower than in Bartels original modelling study, again suggesting a tendency to over estimate mortality predictions.

Zingg (Zingg *et al.*, 2009) compared the performance of the 'Rotterdam' (Steyerberg *et al.*, 2006), 'Munich' (Bartels *et al.*, 1998), and 'Philadelphia' (Ra *et al.*, 2008) models along with the American Society of Anesthesiologists (ASA) physical status score (Saklad, 1941) on cohorts of transthoracic oesophagectomies from Switzerland and Australia. Discrimination and details of calibration were not reported. The Philadelphia and Rotterdam models had some predictive value assessed on Nagelkerke's R squared from logistic regression in pooled data but calibration (Hosmer-Lemeshow) was poor. Only the Philadelphia score had any value in the Swiss cohort. The Munich score was reported to be an ineffective predictor.

2.3.8 Candidate predictors and perioperative mortality

Age: Summary of studies which investigated age

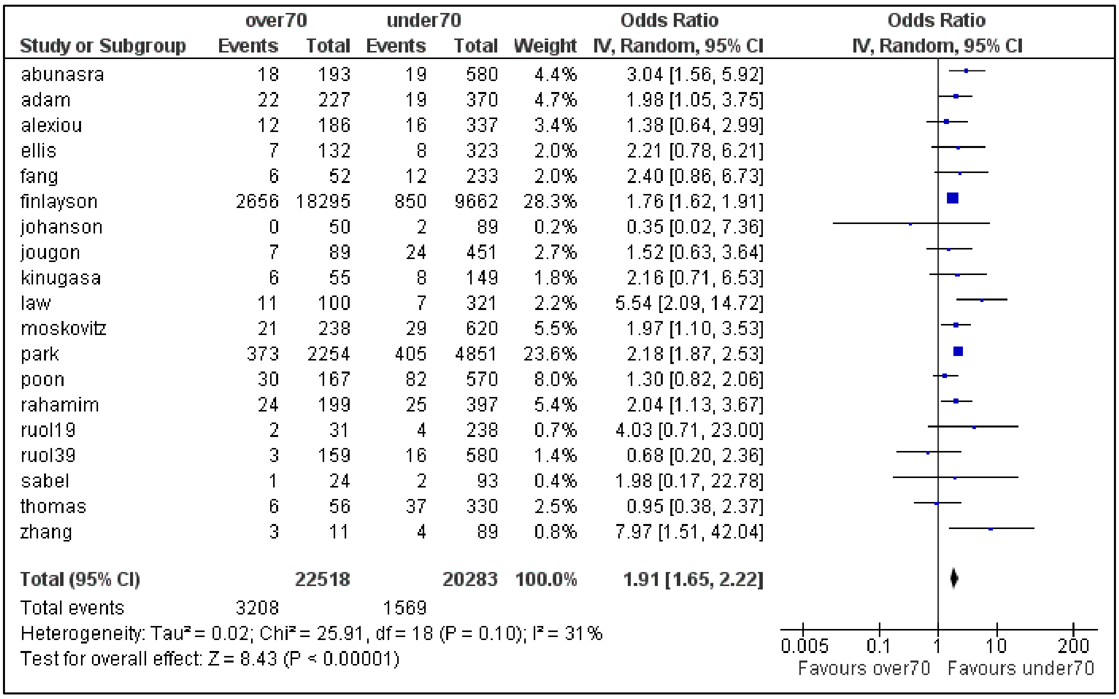
Thirty two studies examined the effect of age on perioperative mortality and their details are summarised in Table 3, Table 4 and Table 5 (Law *et al.*, 1994; Zhang *et al.*, 1994; Adam *et al.*, 1996; Thomas *et al.*, 1996; Ferguson *et al.*, 1997; Jougon *et al.*, 1997; Alexiou *et al.*, 1998; Ellis Jr *et al.*, 1998; Poon *et al.*, 1998; Johansson and Walther, 2000; Fang *et al.*, 2001; Kinugasa *et al.*, 2001; Griffin *et al.*, 2002; Sabel *et al.*, 2002; Bailey *et al.*, 2003; McCulloch *et al.*, 2003; Rahamim *et al.*, 2003; Rentz *et al.*, 2003; Tsai *et al.*, 2003; Atkins *et al.*, 2004; Law *et al.*, 2004; Tekkis *et al.*, 2004; Abunasra *et al.*, 2005; Sauvanet *et al.*, 2005; Moskovitz *et al.*, 2006; Schroder *et al.*, 2006; Steyerberg *et al.*, 2006; Finlayson *et al.*, 2007; Ruol *et al.*, 2007(a); Ruol *et al.*, 2007(b); Ra *et al.*, 2008; Park *et al.*, 2009). Sixteen studies used categorical age groupings and one reported (Griffin *et al.*, 2002) mean for survivors and non-survivors. One divided the sample at 50 (Tsai *et al.*, 2003) and four divided the sample into three age groups (45-63, 63-71, and 71-89) (Rahamim *et al.*, 2003); 65-69, 70-79 and over 80 (Finlayson *et al.*, 2007); under 70, 70-79 and 80-86 (Alexiou *et al.*, 1998); under 50, 50-69 and over 70 (Adam *et al.*, 1996)). These studies are summarised in Table 3.

Table 3 Effect of age on perioperative mortality from studies which compared groups of patients in different age categories

Study design	Author	Mortality %	OR, r, p value	Comments
Comparison <> 70	Sabel	4 vs 2%	na	> 70 "average" 77.1, range 70, 95; < 70 "average" 57.9, range 21, 69. Operative rate in eligible patients 37% of < 70, 18% > 70
	Kinugasa	10.9 vs 5.4	na	
	Fang	7.6 vs 3.3%	p=0.082	
	Ruol(b)	6.5 vs 1.7%	p=0.12	
	Ruol(a)	1.9 vs 2.7%	p=0.778	Distribution: 67.3%, 70-74; 27%, 75-79; 5.7% >=80. Operative rate 57.3 <70, 46.5% >70
	Jougon	7.8% vs 5.3%	p=0.53	> 70, mean 75, range 70-84, 35 patients over 75
	Thomas	10.7% vs 11.2%	"not significant"	% only; "operability" 62.5% < 70, 81.5% > 70
	Ellis Jr	5.3% vs 2.4%	p=0.149	> 70; median 74, range 70-87; operative rate 89.8%, >70; 90.2% <70
	Poon	18% vs 14.4%	p=0.27	Operability 48% > 70, 65% < 70
	Johansson	0 vs 2.7%	na	
Other study designs				
<> 50	Tsai	5.4 vs 3%	na	
45-63, 63-71, and 71-89	Rahamim	12 vs 6.2%		> 70; median 75, (range 71-88)
65-69, 70-79, >=80	Finlayson		p<0.0001 for 3 groups	
<70, 70-79, 80-86	Alexiou (1998)	6.5% vs 4.7%	0.51 for 3 groups	Patients considered unfit for surgery: 2.3% < 70, 8% >70
<50, 50-69 and >70	Adam	na	na	30 day mortality
< 50, 50-59, 60-70, 70-79, >80	Moskovitz	9% vs 4.7%	na	Hazard ratio for mortality >80, 3.9 (p<0.01, CI 1.5, 10.6).
Mean age of survivors and non-survivors	Griffin		p=0.028	Mean for survivors 62.3 vs 68.9 for non survivors

Age: Studies which investigated age over and under 70

None of the ten studies, which investigated patients under and over 70, found statistically significant differences in mortality (Thomas *et al.*, 1996; Jougon *et al.*, 1997; Ellis Jr *et al.*, 1998; Poon *et al.*, 1998; Johansson and Walther, 2000; Fang *et al.*, 2001; Kinugasa *et al.*, 2001; Sabel *et al.*, 2002; Ruol *et al.*, 2007(a); Ruol *et al.*, 2007(b)). This was the most frequently reported design for age effect studies. Five studies recorded mortality rates between 2 and 3.8 times greater in the over 70s (Ellis Jr *et al.*, 1998; Fang *et al.*, 2001; Kinugasa *et al.*, 2001; Sabel *et al.*, 2002; Ruol *et al.*, 2007(b)), but the studies were too small to detect differences at a significance probability of 0.05 and power of 0.8 as calculated on 'statpages.org (Pezzullo, Updated May 2009). Two studies recorded similar mortality rates in the age groups (Jougon *et al.*, 1997; Poon *et al.*, 1998) and three found small, non-statistically significant increases in mortality in younger patients (Thomas *et al.*, 1996; Johansson and Walther, 2000; Ruol *et al.*, 2007(a)).The mortality rates for over and under 70 were also extractable from other study designs. Because of the frequency of this study design and the small sample sizes I decided to attempt a data synthesis of the summary results for over and under 70 year olds using a random effects synthesis, with age category as the intervention (Revman version 5 (Cochrane Information Management System, 2011). The results (Zhang *et al.*, 1994; Adam *et al.*, 1996; Alexiou *et al.*, 1998; Rahamim *et al.*, 2003; Law *et al.*, 2004; Abunasra *et al.*, 2005; Moskovitz *et al.*, 2006; Finlayson *et al.*, 2007; Park *et al.*, 2009) were incorporated into a forest plot (Figure 2 on following page) and a pooled odds ratio calculated. The data synthesis produced an odds ratio of 1.91 (95% CI 1.65, 2.22) for age over 70.



	over70		under70			Odds Ratio
Study or Subgroup	**Events**	**Total**	**Events**	**Total**	**Weight**	**IV, Random, 95% CI**
abunasra	18	193	19	580	4.4%	3.04 [1.56, 5.92]
adam	22	227	19	370	4.7%	1.98 [1.05, 3.75]
alexiou	12	186	16	337	3.4%	1.38 [0.64, 2.99]
ellis	7	132	8	323	2.0%	2.21 [0.78, 6.21]
fang	6	52	12	233	2.0%	2.40 [0.86, 6.73]
finlayson	2656	18295	850	9662	28.3%	1.76 [1.62, 1.91]
johanson	0	50	2	89	0.2%	0.35 [0.02, 7.36]
jougon	7	89	24	451	2.7%	1.52 [0.63, 3.64]
kinugasa	6	55	8	149	1.8%	2.16 [0.71, 6.53]
law	11	100	7	321	2.2%	5.54 [2.09, 14.72]
moskovitz	21	238	29	620	5.5%	1.97 [1.10, 3.53]
park	373	2254	405	4851	23.6%	2.18 [1.87, 2.53]
poon	30	167	82	570	8.0%	1.30 [0.82, 2.06]
rahamim	24	199	25	397	5.4%	2.04 [1.13, 3.67]
ruol19	2	31	4	238	0.7%	4.03 [0.71, 23.00]
ruol39	3	159	16	580	1.4%	0.68 [0.20, 2.36]
sabel	1	24	2	93	0.4%	1.98 [0.17, 22.78]
thomas	6	56	37	330	2.5%	0.95 [0.38, 2.37]
zhang	3	11	4	89	0.8%	7.97 [1.51, 42.04]
Total (95% CI)		**22518**		**20283**	**100.0%**	**1.91 [1.65, 2.22]**
Total events	3208		1569			

Heterogeneity: $\text{Tau}^2 = 0.02$; $\text{Chi}^2 = 25.91$, df = 18 (P = 0.10); $I^2 = 31\%$

Test for overall effect: Z = 8.43 (P < 0.00001)

Figure 2 Forest plot and data synthesis for the effect of age younger and older than 70 years

Age: Other age groupings

Rahamin (Rahamim *et al.*, 2003) found a 1.6 times increase in 30 day mortality in over 71 year olds compared with 63-71, and 2.37 times in 45-63. Finlayson (Finlayson *et al.*, 2007) found operative mortality was 19.9% in octogenarians, 13.4% in 70-79 year olds and 8.8% in 65-69 year olds (statistically significant). Alexiou (Alexiou *et al.*, 1998) found no difference in mortality between under 70s, 70-79 and over 70 year olds, despite an increased incidence of post-operative complications in the older groups. Adam (Adam *et al.*, 1996) also concluded there was no difference in mortality between three age groups (under 50, 50-69 and over 70).

Age: Studies which included age in multivariate studies

Age was included in 15 multivariable studies (Table 4) of outcome (Law *et al.*, 1994; Zhang *et al.*, 1994; Ferguson *et al.*, 1997; Bailey *et al.*, 2003; McCulloch *et al.*, 2003; Rentz *et al.*, 2003; Atkins *et al.*, 2004; Law *et al.*, 2004; Tekkis *et al.*, 2004; Abunasra *et al.*, 2005; Sauvanet *et al.*, 2005; Schroder *et al.*, 2006; Steyerberg *et al.*, 2006; Ra *et al.*, 2008; Park *et al.*, 2009) of which 6 included age as a continuous variable shown in Table 5 (Bailey *et al.*, 2003; Atkins *et al.*, 2004; Tekkis *et al.*, 2004; Abunasra *et al.*, 2005; Schroder *et al.*, 2006; Steyerberg *et al.*, 2006). Eleven of 22 studies reported statistically significant associations between age and perioperative mortality in multivariable designs (Law *et al.*, 1994; Ferguson *et al.*, 1997; Griffin *et al.*, 2002; Rahamim *et al.*, 2003; Rentz *et al.*, 2003; Law *et al.*, 2004; Abunasra *et al.*, 2005; Sauvanet *et al.*, 2005; Finlayson *et al.*, 2007; Ra *et al.*, 2008; Park *et al.*, 2009).

Table 4 Effects of age on perioperative mortality from multivariate studies

Study design	Author	Mortality % or relative	OR, r, p value	Comments
Miscellaneous age groupings	Law 2004	11 vs 2.8%	1.1433 (95% CI 1.0690-1.2229); p=0.002	Not clear whether age category or continuous(OR)
	Law 1994		0.052(r); p<0.001	Categories <> 62; selected from discriminant analysis; not clear if regression coefficient is for category or continuous
	Ferguson	RR 2.8	p=0.001	Relative risk for 67 vs 50
Categorical age group comparisons	Rentz		0.41(r)	30 day mortality for <> 65
	Sauvanet		p=0.001	<> 60; no numerical details
	Abunasra	9.3 vs 3.2%	4.87(95% CI, 1.35, 17.55) for over 73.2; p<0.001	OR adjusted in multiple regression. Age groups <59.5, 59.5-67.8, 67.9-73.2, >73.2 (quartiles). <> 70 derived from results tables.
	McCulloch		na	<60,61-70,71-80,=>81
	Park		P<0.001	<50,,50-59(OR 1.35),60-69(OR1.68),70-79(OR2.64),80+(OR3.84)
	Zhang	27% vs 4.5%	na	
	Ra		30 day mortality; > 80 compared with 65-69: OR 1.88 (95%CI 1.08, 1.36), p= 0.025. OR 70-79 compared with 65-69: 1.54 (CI 1.01, 2.35).	

Table 5 Effects of age on perioperative mortality from multivariable studies with age as a continuous variable

Study design	Author	Mortality % or relative risk	OR, r, p value	Comments
Age as continuous variable	Bailey		OR 1.05 (r 0.05, se 0.01); p=0.0001	Multivariate adjusted (r). 30 day mortality
	Schroder		OR 9.6 (95% CI 2.6-32.7); p=0.001	Unclear whether age category or continuous. Adjusted
	Tekkis		1.06(1.03,1.08)	
	Atkins		0.066 (r); p=0.003	30 day mortality
	Steyerberg		OR 1.4(1.2,1.7)	OR per decade; 30 day mortality
	Abunasra		OR 1.97	Adjusted OR

Age: effect of age over 80

Moskovitz (Moskovitz *et al.*, 2006) showed an increase in perioperative mortality in octogenarians (nearly three times compared with patients between 70 and 79) and confirmed this effect in a multivariable logistic regression, which controlled for various comorbidities and demonstrated acceleration in the effect of age on perioperative mortality in the ninth decade. Other studies also reported a marked effect in octogenarians (Moskovitz *et al.*, 2006; Finlayson *et al.*, 2007; Park *et al.*, 2009)

Age: Distribution of comorbidities

Most studies recorded the distribution of gender, tumour and operative details, and use of neoadjuvant therapy, between the age groups (Thomas *et al.*, 1996; Jougon *et al.*, 1997; Alexiou *et al.*, 1998; Ellis Jr *et al.*, 1998; Poon *et al.*, 1998; Johansson and Walther, 2000; Fang *et al.*, 2001; Kinugasa *et al.*, 2001; Rahamim *et al.*, 2003; Tsai *et al.*, 2003; Alexiou *et al.*, 2005; Finlayson *et al.*, 2007; Ruol *et al.*, 2007(a); Ruol *et al.*, 2007(b)). Cardiac and respiratory morbidities and some other comorbidities (e.g. incidence of diabetes, liver disease and renal disease) were also recorded in some studies (Thomas *et al.*, 1996; Jougon *et al.*, 1997; Alexiou *et al.*, 1998; Poon *et al.*, 1998; Fang *et al.*, 2001; Kinugasa *et al.*, 2001; Moskovitz *et al.*, 2006; Finlayson *et al.*, 2007; Ruol *et al.*, 2007(a); Ruol *et al.*, 2007(b)). The incidence of cardiovascular and respiratory disease was generally higher in the elderly groups (Jougon *et al.*, 1997; Poon *et al.*, 1998; Fang *et al.*, 2001;

Kinugasa *et al.*, 2001; Moskovitz *et al.*, 2006; Ruol *et al.*, 2007(a); Ruol *et al.*, 2007(b)) but the opposite was found in the study by Alexiou (Alexiou *et al.*, 1998).

Cardiovascular comorbidity

Twelve studies investigated the effect of cardiovascular comorbidity on perioperative mortality (Law *et al.*, 1994; Zhang *et al.*, 1994; Ferguson *et al.*, 1997; Alexiou *et al.*, 1998; Bartels *et al.*, 1998; Kuwano *et al.*, 1998; Liu *et al.*, 2000; Griffin *et al.*, 2002; Bailey *et al.*, 2003; Law *et al.*, 2004; Abunasra *et al.*, 2005; Gockel *et al.*, 2005; Schroder *et al.*, 2006). These are listed in Table 6 on the following page. Comorbidity was coded in a large variety of ways and only two (Zhang *et al.*, 1994; Gockel *et al.*, 2005) found an association with perioperative mortality. Five studies (Bartels *et al.*, 1998; Atkins *et al.*, 2004; Tekkis *et al.*, 2004; Steyerberg *et al.*, 2006; Ra *et al.*, 2008) demonstrated a relationship between outcome and composite comorbidity scores, which incorporated some element of cardiac morbidity.

Table 6 Studies which included an investigation of the effect of cardiac comorbidity on perioperative mortality

Study author	Definition of preoperative morbidity	Comments
Law (2004)	Pre-existing cardiac disease, abnormal ECG (ischaemia/arrhythmia), abnormal CXR	No values reported
Alexiou	History (IHD, hypertension, myocardial infarction, peripheral vascular disease, DVT)	No values reported
Ferguson	NYHA heart failure score, hypertension, beta or calcium channel blockers, previous MI (p=0.1)	For NYHA heart failure
Bailey	Congestive heart failure, dyspnoea at rest, history of CVA, (p non-significant)	No values reported
Law (1994)	Abnormal chest xray & cardiograph, (p non-significant)	No values reported
Liu	Mild arrythmia, hypertension, valve disease without symptoms, aortic stenosis, angina, old infarction (p=0.0001 for hypertension only)	For hypertension only
Kuwano	ECG abnormalities requiring treatment, myocardial ischaemia, arrythmia, valve disease, abnormal scintography, (p non-significant)	No values reported
Schroder	Composite score(physician defined cardiac risk, electrocardiograph, chest xray), (p non-significant)	No values reported
Griffin	MI, CABG, hypertension, symptoms, ECG, exercise test, (p non-significant)	Incidence of cardiovascular disease, 44% in non-survivors, 31% in survivors
Gockel	History of coronary heart disease, MI, arterial hypertension, valvular disease, arrhythmia requiring treatment, congestive heart failure, peripheral occlusive disease (p=0.0172)	
Zhang	Abnormal ECG (p=0.06)	r, 3.4 in logistic regression for abnormal ECG
Whooley	Abnormal ECG(ischaemia, arrhythmia); previous cardiac history, (p non-significant)	

Respiratory comorbidity

Nine of 14 studies (Law *et al.*, 1994; Bartels *et al.*, 1998; Liu *et al.*, 2000; Rentz *et al.*, 2003; Abunasra *et al.*, 2005; Alexiou *et al.*, 2005; Sanz *et al.*, 2006; Schroder *et al.*, 2006; Healy *et al.*, 2008) reported an association between pre-existing pulmonary disease or pulmonary function, and mortality. These are summarised in Table 7 and Table 8. Three studies included pulmonary components (Bartels *et al.*, 1998; Tekkis *et al.*, 2004; Steyerberg *et al.*, 2006) in composite scores which were associated with mortality. Healy (Healy *et al.*, 2008) reported an association between preoperative dyspnoea and mortality.

Table 7 Studies of the effect of respiratory comorbidity on mortality: physiological measures

Study author	Definition of preoperative morbidity	Odds ratio(OR), regression coefficient®, relative risk(RR) or mortality rate%	Probability significance (p value) for predictor effect	Notes
Law (2004)	Spirometry & gases		ns	No values
Ferguson	Spirometry, arterial gases, CO diffusion		0.085, FEV1(univariate)	FEV1 ns in multivariate model.
Law (1994)	Spirometry (incentive), arterial gases, chest xray		Incentive spirometry(<0.001), PCO2 (0.032), abnormal chest xray(<0.001)	Incentive spirometry predictive in multivariate model
Liu	Spirometry		0.049	FEV1/FVC; no values
Kuwano	Spirometry		ns	not predictive
Bartels	Spirometry and arterial gases	RR 1.7 for impaired respiratory function (composite)	<0.05 for VC & arterial pO2 and	Discriminant analysis to maximise relative risk for VC<90% predicted, PaO2<70mmHg.
Schroder	As in Bartels	1.56(95% CI 1.01, 3.4)	0.049	Respiratory function score calculated as unweighted Bartels score
Griffin	Spirometry, arterial gases		ns	not predictive
Abunasra	Spirometry	4.72 (1.01, 21.99) for highest vs lowest quartile	FEV1, 0.001; FEV1, 0.004; FVC, 0.014; note %predicted	Lowest relative to highest quartile
Sanz	Spirometry, arterial gases	RR 1.1(95% CI 0.7-3.5), p=0.014 for multivariate model	0.03	
Zhang	Spirometry		ns	
Healy	Spirometry		ns	not predictive

Table 8 Studies of the effect of respiratory morbidity on mortality: clinical history

Author	Description of respiratory disease used for prediction		Comments
Law	Pre-existing pulmonary disease	ns	No values
Alexiou	Pre-existing pulmonary disease	P=0.15	No values
Rentz	Severe COPD, dyspnoea, current pneumonia	Dyspnoea: regression coefficient in logistic regression: 0.41 (p=0.0477)	
Bailey	Severe COPD, dyspnoea at rest		Not predictive
Law	Chronic respiratory disease		Not predictive
Gockel	Chronic obstructive pulmonary disease, use of bronchodilators	0.0059 univariate: p= 0.0099 multivariate	
Healy	Dyspnoea	1.08(1,1.7); multivariate (p=0.041); p<0.001 univariate	
Griffin	History of chronic lung disease	Present in 44% of non-survivors & 22% survivors	

Exercise or activity capacity

These studies are summarised in Table 9. Of three studies of cardiopulmonary exercise capacity (Law *et al.*, 1994; Murray *et al.*, 2007; Forshaw *et al.*, 2008), one (Law *et al.*, 1994) reported an association of 'stair climbing capacity' with mortality. Of five studies of activity or general health, the following four reported an association with mortality. Ferguson (Ferguson *et al.*, 1997) reported the Zubrod (Oken *et al.*, 1982), health and activity score, Bartels (Bartels *et al.*, 1998) reported the Karnovsky health score (Karnofsky, 1984), Healy the EORTC QOL questionnaire (Healy *et al.*, 2008), and Bailey and Rentz used undefined scores (Bailey *et al.*, 2003; Rentz *et al.*, 2003).

Table 9 Studies which investigated exercise or activity capacity as predictors of perioperative mortality

Author	Description of exercise or activity capacity used in study	Relative risk(RR), mortality rate(%), odds ratio (OR), or regression coefficient® with significance (p)	Notes
Cardiopulmonary exercise testing			
Law (1994)	Stair climbing	RR 2.9 in high risk group (p=0.015)	Discriminant analysis to maximise RR for high risk group. Multivariate model
Forshaw	Cardiopulmonary exercise testing (anaerobic threshold, VO2 max)		Not predictive (one death)
Murray	Shuttle walk test		5/8 patients who could not walk 340 metres died; all survived if walked 340 metres
General health or activity scores			
Healy	EORTC QOL	P=0.02	Dyspnoea(p<0.001), fatigue(p=0.003) and nausea & vomiting (p=0.025) are components of QOL associated with perioperative mortality. Only dyspnoea predictive in multivariate model.
Rentz	Undefined "diminished functional health"		
Bailey	Functional status (unclear definition)	r, 0.48(s.e. 0.18), p=0.007 for multivariate model	
Ferguson	Zubrod performance score	P=0.03	Included in multivariate model
Bartels	Karnovsky index	<0.001 for karnovsky index less than 80%	Discriminant analysis to maximise RR of Karnovsky
Composite morbidity scores			
Atkins	Charlson comorbidity score	r=0.89 (p=0.05)	not predictive

Nutritional status

Sixteen authors (Saito *et al.*, 1993; Law *et al.*, 1994; Ferguson *et al.*, 1997; Bartels *et al.*, 2000; Liu *et al.*, 2000; Griffin *et al.*, 2002; Bailey *et al.*, 2003; Rentz *et al.*, 2003; Atkins *et al.*, 2004; Law *et al.*, 2004; Abunasra *et al.*, 2005; Alexiou *et al.*, 2005; Gockel *et al.*, 2005; Sauvanet *et al.*, 2005; Han-Geurts *et al.*, 2006; Sanz *et al.*, 2006; Healy *et al.*, 2008) examined the effect of a variety of measures of nutritional or immune status on perioperative mortality. These are summarised in Table 10 and Table 11. Three of 9, who investigated serum albumen found associations (Law *et al.*, 1994; Rentz *et al.*, 2003; Atkins *et al.*, 2004; Sanz *et al.*, 2006), one found an association with a "host defence index" (including α2 macroglobulin and arm circumference)(Saito *et al.*, 1993), one with arm circumference (Law *et al.*, 1994), and one with a composite "general status" measure, which included weight loss (Bartels *et al.*, 2000). Only one (Law *et al.*, 1994) of 12, who investigated weight loss found any association with mortality.

Table 10 Studies of nutritional status and its effect on perioperative mortality: serum protein, albumen and white cell count

Study author	Description of nutritional measure	Odds ratio(OR), regression coefficient(r), relative risk(RR); (p value)	Notes
	ALBUMEN		
Law(2004)			Not predictive
Ferguson		0.43	Not predictive
Rentz		r=0.056, p=0.0135	Less than 35 gm/L
Law(1994)		0.001	Not predictive in multivariate model
Bartels			Not predictive
Sanz		0.02	p=0.01 in multivariate model
Griffin		ns	Not predictive
Atkins		r=0.078, ns	Not predictive
Saito		ns	Not predictive
	SERUM PROTEIN		
Ferguson	Total serum protein	0.27	Not predictive
	WHITE CELL COUNT		
Ferguson	Lymphocyte count	0.55	Not predictive
Law(1994)	White cell count		Not predictive
Saito	Lymphocyte, T cell, B cell	ns	Not predictive

Table 11 Studies of nutritional state; measures of loss of body mass & miscellaneous other measurements

Study author	Description of nutritional measure	Odds ratio(OR), regression coefficient(r), relative risk(RR); (p value)	Comments
	WEIGHT LOSS		
Alexiou			Not predictive
Ferguson	Weight loss during previous 6 months	0.87	Not predictive
Law(1994)	% weight loss	0.048	Not predictive in multivariate model
Sauvanet	3 groups: <10%, 10-20%,>20%	0.161	Not predictive
Liu	Weight loss during previous 6 months		Not predictive
Abunasra	Body mass index(quartiles)	0.433 for trend in 4 quartiles	Not predictive
Gockel	Body mass index	0.072	Not predictive
Bartels	% Weight loss		Not predictive
Griffin	% weight loss	ns	Not predictive; weight loss 4.7% in survivors, 6.9% in non-survivors.
Healy	% weight loss		
Atkins	% weight loss	r=-0.54, p=0.25	Not predictive
Saito	% of ideal body weight	ns	Not predictive
	OTHER		
Law(1994)	Hand grip strength, tricep skinfold	Hand grip (p=0.003); mid-arm circumference RR 3.0 in high risk group,(p<0.001)	Mid arm circumference included in multivariate model
Gockel	Nutritional score combining alcohol/tobacco use	P=0.222	Not predictive
Saito	21 separate variables measuring aspects of immune and nutritional state		In univariate analysis arm muscle circumference and α2-macroglobulin were significant predictors (p<0.05). A "host defence index" was constructed to predict mortality (included albumen, B cell count, albumen & 7 other serum proteins)
Han-Geurts	PNI (‘prognostic nutritional index’), NRI(‘nutritional risk index’), BMI & weight loss		

Other candidate predictors

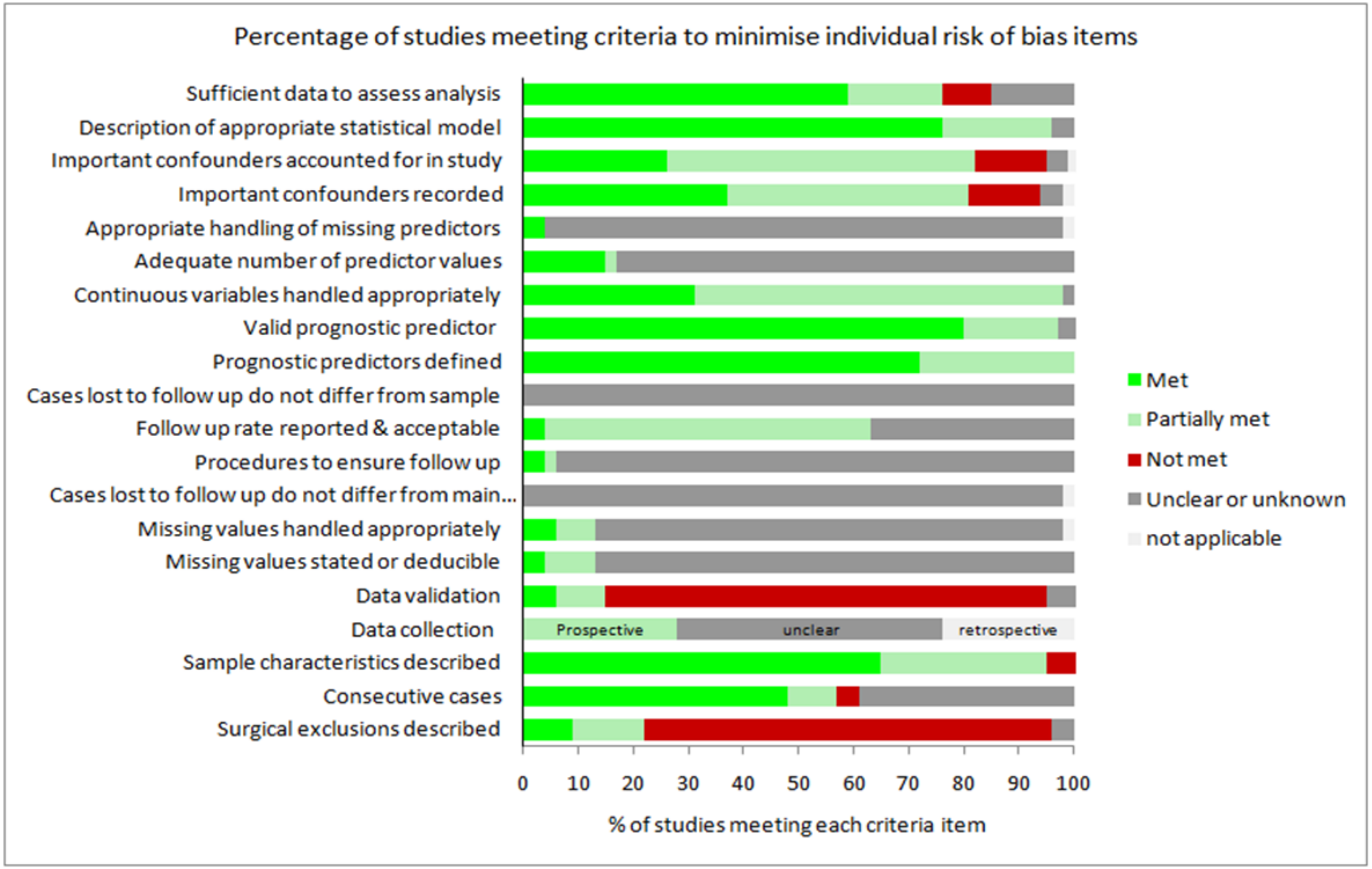
Nine studies investigated the effect of tumour characteristics (Zhang *et al.*, 1994; Whooley *et al.*, 2001; Law *et al.*, 2004; Abunasra *et al.*, 2005; Alexiou *et al.*, 2005; Gockel *et al.*, 2005; Sanz *et al.*, 2006; Steyerberg *et al.*, 2006; Healy *et al.*, 2008). Three (Abunasra *et al.*, 2005; Gockel *et al.*, 2005; Sanz *et al.*, 2006) reported tumour site to influence outcome and one reported tumour stage to influence outcome. Nine studies (Law *et al.*, 1994; Ferguson *et al.*, 1997; Whooley *et al.*, 2001; Rentz *et al.*, 2003; Abunasra *et al.*, 2005; Gockel *et al.*, 2005; Sauvanet *et al.*, 2005; Sanz *et al.*, 2006; Healy *et al.*,

2008) included investigation of surgical approaches to oesophagectomy. Only Gockel (Gockel *et al.*, 2005) reported a perioperative difference between transhiatal and transthoracic procedures but it was unclear which was favoured. Of six studies (Whooley *et al.*, 2001; Bailey *et al.*, 2003; Atkins *et al.*, 2004; Law *et al.*, 2004; Schroder *et al.*, 2006; Steyerberg *et al.*, 2006) of neoadjuvant therapy, only one (Steyerberg *et al.*, 2006) reported an increased perioperative mortality with radiotherapy alone or with chemotherapy, compared to chemotherapy alone.

Other predictors reported to be associated with perioperative mortality included renal impairment (Zhang *et al.*, 1994; Bailey *et al.*, 2003), hepatic impairment (Bailey *et al.*, 2003; Gockel *et al.*, 2005), alcohol use (Bailey *et al.*, 2003), tobacco use (Law *et al.*, 1994; Liu *et al.*, 2000), ASA grade (Griffin *et al.*, 2002), low white cell count (Griffin *et al.*, 2002), gender (Griffin *et al.*, 2002), cholesterol (Sanz *et al.*, 2006), and an "index of medical deprivation" (Leigh *et al.*, 2006). Two (Zhang *et al.*, 1994; Bailey *et al.*, 2003) of eleven studies (Zhang *et al.*, 1994; Ferguson *et al.*, 1997; Bartels *et al.*, 1998; Kuwano *et al.*, 1998; Liu *et al.*, 2000; Bailey *et al.*, 2003; Law *et al.*, 2004; Alexiou *et al.*, 2005; Gockel *et al.*, 2005; Sanz *et al.*, 2006; Steyerberg *et al.*, 2006) found an association between diabetes mellitus and perioperative mortality.

Potential risk of bias in primary studies

The pattern of how well the potential risks for bias, as outlined by Hayden (Hayden *et al.*, 2006), were managed or reported across the primary studies is shown in Figure 3. The risk of bias profile for individual primary studies is shown in Appendix E.



.

Figure 3 Reporting & management of potential for bias across included primary studies

A minority of studies reported potential selection bias, e.g. patients, who were eligible for surgery but were excluded (Law *et al.*, 1994; Jougon *et al.*, 1997; Forshaw *et al.*, 2008; Healy *et al.*, 2008). Reasons for exclusion from surgery were described in two studies (Jougon *et al.*, 1997; Forshaw *et al.*, 2008). Some investigators also chose to study a particular age group or surgical operation (Moskovitz *et al.*, 2006)

In twenty three studies (about 40%) it was not clear whether study samples included all consecutive cases. Twenty six (48%) included all operated cases in their analyses. One study excluded patients who had had neoadjuvant therapy (Tsai *et al.*, 2003) and one selected only those receiving neoadjuvant therapies (Ruol *et al.*, 2007(b)). Twelve (22%) studies fully or partially accounted for patients who were excluded from surgery but were potentially eligible (Law *et al.*, 1994; Jougon *et al.*, 1997; Alexiou *et al.*, 1998; Sanz *et al.*, 2006; Forshaw *et al.*, 2008) and three reported reasons for exclusion (Law *et al.*, 1994; Jougon *et al.*, 1997; Forshaw *et al.*, 2008).Thirty five studies (65%) and sixteen studies (30%) respectively, completely or partially, reported important sample characteristics.

Fifteen studies (28%) collected data prospectively (often into clinical databases), thirteen (24%) retrospectively (for example, extracting data from clinical records) and in 26 (48%) it was unclear. Data validation techniques such as audit were only clearly or partially described in eight studies (Adam *et al.*, 1996; Bailey *et al.*, 2003; Rahamim *et al.*, 2003; Tekkis *et al.*, 2004; Moskovitz *et al.*, 2006; Al-Sarira *et al.*, 2007; Park *et al.*, 2009), with 43 (80%) making no reference to data validation. Only 9 (17%) studies reported missing values and/or their handling (Ferguson *et al.*, 1997; Zafirellis *et al.*, 2002; Tekkis *et al.*, 2004; Sauvanet *et al.*, 2005; Leigh *et al.*, 2006; Steyerberg *et al.*, 2006; Lagarde *et al.*, 2007; Lai *et al.*, 2007; Nagabhushan *et al.*, 2007). Three studies (6%) reported procedures to ensure patient follow up (Adam *et al.*, 1996; Jougon *et al.*, 1997; Takagawa *et al.*, 2008) and thirty four (63%) reported enough information to enable deduction of follow up rate. No studies reported any details of cases, which were lost to follow up.

Thirty six studies reported all the prognostic variables, which I thought important for this review, and the remainder reported varying numbers. Thirty provided definitions of predictors and their methods of measurement, sixteen provided these for some predictors and in two it was unclear. There was considerable heterogeneity in defining cardiorespiratory and nutritional comorbidities. Seventeen studies (31%) treated potentially continuous variables (e.g. age) as continuous and 36 (67%) as categorical.

I defined confounding variables predominantly as predictors likely to be important predictors of mortality and associated in some way with other predictors (e.g. age and cardiovascular disease). Nineteen studies (35%) recorded confounding variables as specified in our methods, and twenty four (44%) recorded some. Fourteen (26%) accounted for confounders in study design or analysis and thirty (56%) partially accounted for confounders, for example, by tabulating their distribution between study groups. In nine (17%), reporting of confounders was unclear.

Data analysis

Forty one (76%) studies described appropriate adequate statistical methods and eleven (20%) described appropriate methods but incompletely. In two studies, the methods were unclear. Thirty one studies (57%) reported enough data to assess analysis. In fourteen studies of binary predictors, notably age categories, the sample sizes were too small to detect statistical significance on measured observations, and it was unclear in twenty six studies, particularly those using regression. One study of the forty where regression techniques were used, reported description of interaction between prognostic factors, collinearity between variables or sensitivity to extreme values.

Ranking studies by ‘risk of bias’

It was difficult to use ‘potential risk of bias’ scores to select the most reliable studies because individual study context can be important. There were a heterogeneous mix of study designs and aims, therefore I did not try and draw conclusions based on ranking scores for having satisfied the criteria for minimising risk of bias

Publication bias

Formal statistical methods (e.g. funnel plots) require directly comparable presentation of study effects (for instance relative risk), similar interventions, and at least five constituent studies (Rothstein *et al.*, 2005). The only studies with comparable results were those comparing outcomes in patients under and over 70 years. The samples in these studies may have been quite heterogeneous because of the unknown prevalence of the very elderly and cardiorespiratory disease and therefore the interpretation of statistical estimation of publication bias should be tempered with caution. The funnel plot of the distribution of effects of the age grouping between studies is shown in Figure 4 (overleaf). This was constructed using in Revman (Cochrane Information Management System, 2011). There is no strong evidence of publication bias in these studies, as the odds ratios are symmetrically distributed around the pooled estimate. The one possible outlier was the study by Johansson (Johansson and Walther, 2000), but this had only two fatalities in a small sample size.

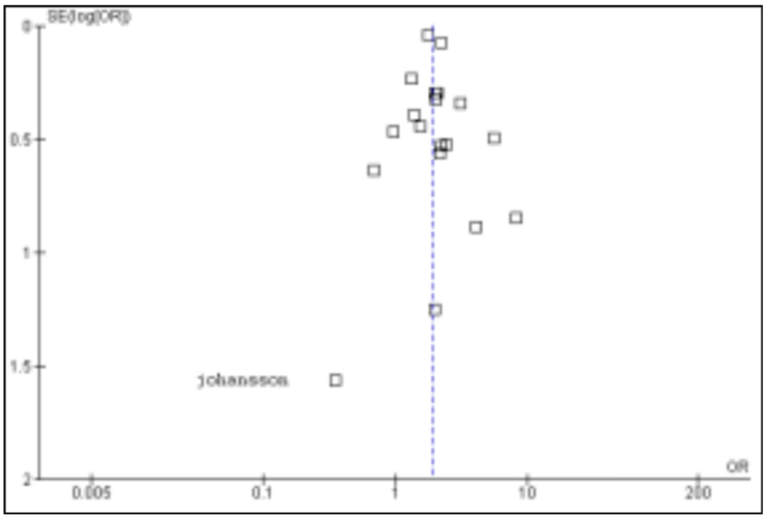


Figure 4 Funnel plot of odds ratios for effect of age over and below 70 on perioperative mortality generated in Revman. See text for explanation.

2.4 Discussion

2.4.1 Summary

Ten studies of prediction models fitted our inclusion criteria. Only the POSSUM based models (Zafirellis *et al.*, 2002; Tekkis *et al.*, 2004; Lagarde *et al.*, 2007; Lai *et al.*, 2007; Nagabhushan *et al.*, 2007), the 'Munich' (Bartels *et al.*, 1998) and 'Rotterdam' models (Steyerberg *et al.*, 2006) had been validated on patient samples from populations outside the development sample. In these studies, performance degraded and overestimation of mortality risk in higher risk groups was common. There were no formal clinical impact studies, but two were reported to reduce perioperative mortality (Bartels *et al.*, 1998; Tsai *et al.*, 2003) after they had been implemented in practise. The lack of consistent definition and 'face validity' of some of the predictors made these models difficult to transfer to new populations.

Forty four studies investigated the effect of comorbidity on perioperative mortality. The reported effects of preoperative comorbidity predictors on outcome were inconsistent but age appeared to be convincingly associated with mortality. Other comorbidity variables were defined inconsistently and it was difficult to draw conclusions about their potential role or effect. Cardiorespiratory disease, nutritional state, activity levels, and site of tumour were variably reported to affect outcome.

In many studies of prediction models and comorbidity effects, potential sources of bias were not reported or addressed. The details of statistical modelling were sometimes difficult to follow, and the age predictor was frequently included as a categorical rather than continuous variable, risking loss of information. Only the 'Rotterdam' model appeared potentially applicable to our patients from the NOGCU and only age appeared conclusively important for prediction models. The effects of other comorbidity predictors are still at an exploratory stage.

2.4.2 Prediction Models

Model development entails a series of stages from initial data exploration, through modelling and validation to clinical application (Wallace *et al.*, 2011). Clinical impact studies are uncommon, with most studies focussing on model development, internal validation and a few on external validation (Bouwmeester *et al.*, 2012). This was the case for the studies included in this review.

The POSSUM based models were the most frequently tested on new populations. The POSSUM (the 'Physiological and Operative Severity Scores for the Enumeration of Mortality and Morbidity') (Copeland *et al.*, 1991) was developed as an audit tool for perioperative outcome in 1991 and has been applied to various surgical groups subsequently. However, the original POSSUM did not perform well in new groups, specifically overestimating mortality in oesophagectomy (Zafirellis *et al.*, 2002). This is not surprising as the POSSUM was developed on a heterogeneous group of surgical patients including emergencies. Some of the higher scoring items in the POSSUM score included items most likely to be found in patients presenting for emergency surgery, rather than for major elective surgery e.g. impaired conscious level, heart failure, renal failure, severe respiratory impairment. These items may bias predictions in an elective surgical population and lack 'face validity' as the patients scoring on these items are unlikely to be presenting for major elective surgery.

The O-POSSUM was developed to focus on oesophagogastric surgery. It was developed by Tekkis (Tekkis *et al.*, 2004) from United Kingdom databases containing data from oesophagogastric cancer centres (McCulloch *et al.*, 2003; Tekkis *et al.*, 2004). The O-POSSUM excluded surgical process variables but incorporated age as an independent variable, although this was represented again in the Physiological Severity Score part of the POSSUM. Performance was acceptable in development and internal validation but deteriorated when applied to new populations (Tekkis *et al.*, 2004; Lagarde *et al.*, 2007; Lai *et al.*, 2007; Nagabhushan *et al.*, 2007).

Overestimation of mortality was a common problem with all POSSUM models (Lagarde *et al.*, 2007; Lai *et al.*, 2007; Nagabhushan *et al.*, 2007). Differences between modelling and validation samples could have impaired performance. The original O-POSSUM was developed on a sample containing 79.5% elective patients (mortality 9.4%), 7.5% emergencies (mortality 26.9%) and 13.1% unknown (19.1% mortality). Differences in validation samples included the use of 30 day instead of 'in hospital' mortality (Nagabhushan *et al.*, 2007), only using elective cases (Lagarde *et al.*, 2007; Nagabhushan *et al.*, 2007), different operations (Lagarde *et al.*, 2007) (Lai *et al.*, 2007) and overall mortality rates (Lagarde *et al.*, 2007). Differences in comorbidities may also have affected Lai's Hong Kong validation study (Alexiou *et al.*, 2006; Lai *et al.*, 2007). The P-POSSUM (Prytherch *et al.*, 1998), which had been developed on a United Kingdom mixed surgical population to improve on the original POSSUM, performed well on the Hong Kong sample but poorly elsewhere (Nagabhushan *et al.*, 2007).

The 'Munich' model (Bartels *et al.*, 1998), which classified patients into three ascending risk groups, successfully identified a high risk group, in a subsequent sample from the same centre, with a similar mortality to that in the modelling sample (25%). It was later introduced into clinical practise and reported to reduce 90 day mortality rate from over 10% to about 5%, but at the expense of excluding 24 patients from surgery, but there were no formal clinical impact studies. However, when validated externally the high risk group had in-hospital mortality of 16.7% (Schroder *et al.*, 2006), again suggesting overestimation when applied in new populations. This model would be unlikely to be applicable elsewhere because of differences in prognostic variable definition e.g. "subjective cardiac assessment by cardiologist" and some risk items (e.g. hepatic impairment) suggest that the model was applied to patients who may not be currently considered for surgery.

Steyerberg's model (Steyerberg *et al.*, 2006) was developed on a population database from the USA and validated on data from the same database as well as population and clinical databases from the Netherlands. The model calibrated well but discrimination was poor. The risk score was simple and reproducible but the comorbidity items were necessarily general and did not discriminate layers of risk e.g. cardiac morbidity scoring. This was the only model that could possibly be applied to data from the NOGCU.

2.4.3 Candidate predictors for perioperative mortality

Age

It might be expected that an increase in age would be associated with increasing perioperative morbidity and mortality because of increasing incidence and severity of comorbidity, and a reduced capacity to respond to physiological stress (Priebe, 2000; Park *et al.*, 2009). However, this was not clearly reflected in the primary studies included in this review.

Most of the primary studies divided patients into age categories. The most common cut-off, by far was 70 years and none found statistically significant associations with mortality. There were frequently reported differences, but the samples were too small to achieve statistical significance. Studies, which used age categories, were unable to show the proportion of very old patients, and therefore it was not clear whether the effect of extreme old age was investigated. Combining the pooled results of studies, which divided groups at 70 years, suggested that age is an important predictor (odds ratio 1.91 for over 70). Ideally it should also be included as a continuous variable to incorporate information from across the whole range of old age.

Confounding could also have influenced the effect of age on outcome. For example, both the incidence and severity of cardiovascular disease are likely to increase in extreme old age (Priebe, 2000) and be associated with perioperative mortality. In the primary studies included in this review, the incidence of cardiovascular and respiratory disease was generally higher in the elderly groups (Jougon *et al.*, 1997; Poon *et al.*, 1998; Fang *et al.*, 2001; Kinugasa *et al.*, 2001; Moskovitz *et al.*, 2006; Ruol *et al.*, 2007(a); Ruol *et al.*, 2007(b)). The distribution of these comorbidities were generally reported but not included in statistical analysis. Differing resection rates and choice of procedure between age groups (Thomas *et al.*, 1996; Sabel *et al.*, 2002; Ruol *et al.*, 2007(a)) and levels of perceived fitness (Jougon *et al.*, 1997; Alexiou *et al.*, 1998; Moskovitz *et al.*, 2006) may also bias the effects of age. Many studies did not record this information.

Nutritional status

The nutritional effects of gastro-oesophageal cancer have been described by Gupta (Gupta and Ihmaidat, 2003). Nutritional compromise is common in gastro-oesophageal cancer, caused by both local mechanical effects and neoplastic systemic effects (Fekete and Belghiti, 1988). Protein-calorie malnutrition has been demonstrated in hospitalised cancer patients (Nixon *et al.*, 1980) and is associated with impaired cardiac muscle, respiratory muscle and skeletal muscle function as well as intestinal muscle atrophy. Protein-calorie malnutrition has also been associated with immune system impairment in oesophageal cancer and may potentially increase the possibility of post-operative infection (Law *et al.*, 1973). Postoperatively, malnutrition has been associated with increased rates of anastomotic breakdown (Fekete and Belghiti, 1988) and postoperative respiratory complications (Windsor and Hill, 1988). In this review, five (Saito *et al.*, 1993; Law *et al.*, 1994; Rentz *et al.*, 2003; Atkins *et al.*, 2004; Sanz *et al.*, 2006) of sixteen studies, which investigated nutritional status, found an association between a variety of measures of nutritional status and perioperative mortality. Reported measures included serum albumen, a variety of measures of loss of body mass and immunological studies. This provided only weak evidence that the nutritional or immune status measures, which were studied here, should be considered as candidate predictors. There is physiological rationale for the inclusion of a nutritional measure, but the heterogeneous nature of the studies, definitions and results means that we are at an early data exploration stage and unable to draw conclusions to guide selection of predictors.

Cardiovascular comorbidity

Serious cardiac complications in a population of mixed major surgical procedures have been reported to be about 2.5% (Lee *et al.*, 1999) and preoperative cardiovascular disease has been associated with increased morbidity and mortality in many studies of non-cardiac surgery (Mangano, 1990; Eagle *et al.*, 1997). Myocardial infarction and cardiac mortality combined, has been reported in 'high risk surgery' to be over 4% in patients who had medically treated coronary heart disease (Eagle *et al.*, 1997). These rates are also influenced by other comorbidities including peripheral vascular disease, where there was an observed cardiac morbidity of 8% (L'Italien *et al.*, 1996), and increasing age (Priebe, 2000). More recently, in a mixed surgical population with cardiovascular comorbidity, total cardiac complications were about 6% and mortality about 1.5% (2008). Of course it may be that other causes of mortality may be associated with cardiac comorbidity, for instance generalised vascular disease may be associated with an increased risk of multiorgan failure or anastomotic breakdown. Cardiac comorbidity may then have a stronger predictive effect.

In the articles included in this review, where the information was available, between 9 % and 15% of perioperative mortality was attributable to cardiac causes (Law *et al.*, 1994; Whooley *et al.*, 2001; Law *et al.*, 2004; Alexiou *et al.*, 2005). This means cardiac mortality might be about 1-2% in studies where all cause mortality was between 5% and 10%. If we assume that deaths, which are primarily attributable to cardiac comorbidity, are cardiac in nature, it is not surprising that of seventeen studies, which investigated this, only four found any association. This is because the sample sizes were small relative to the attributable mortality (Zhang *et al.*, 1994; Bartels *et al.*, 1998; Liu *et al.*, 2000; Tekkis *et al.*, 2004; Gockel *et al.*, 2005; Steyerberg *et al.*, 2006).

One of the problems in this review was the lack of consistent definition for reporting or scoring cardiac comorbidity. Many studies have used stratified scores for cardiac risk for non-cardiac surgery. The most commonly used is probably the Revised Cardiac Risk Index (RCRI) (Lee *et al.*, 1999), which stratified into groups with cardiac event rates from 0.5% to 11%. A recent validation study (Boersma *et al.*, 2005) examined the predictive performance of the RCRI on patients, which included oesophagectomy, with an overall cardiac mortality of 0.8% and all cause mortality of 4%, in keeping with the studies in this review. Predicted mortalities ranged from 0.3% in RCRI Class 1 to 3.6% in Class 4. The authors found that the addition of age and operation detail also added to the predictive performance.

It seems reasonable that, although most studies in our review did not identify preoperative cardiovascular morbidity as an independent risk factor for perioperative mortality, other evidence supports its consideration in prediction models. Despite the relatively low prevalence of attributable outcomes and possibly high risk scores, it may be worth considering the use of a standardised risk score such as the RCRI.

Respiratory comorbidity

Transthoracic oesophagectomy involves chest wall surgery, prolonged operating time, one lung ventilation, mechanical retraction of lung tissue, thoracic lymphadenectomy and potentially large and complex body fluid shifts. Unsurprisingly, respiratory complications are common with rates reported up to 32% (Law *et al.*, 1994; Whooley *et al.*, 2001; Law *et al.*, 2004). These complications are also a major cause of mortality, for example contributing up to 55% of all perioperative fatalities (Whooley *et al.*, 2001; Law *et al.*, 2004). It seems reasonable that respiratory comorbidity should be a candidate predictor for a prediction model.

Predictors for post-operative pulmonary complications in non-cardiothoracic surgery have been examined in a systematic review by Smetana (Smetana *et al.*, 2006), which was used to develop a stratification guideline by the American College of Physicians (Qaseem *et al.*, 2006). They reported that for age over 60, chronic obstructive pulmonary disease, congestive heart failure, functional dependence, ASA grade, and serum albumen less than 35 grams/litre were all considered important predictive factors. Spirometry was also associated with pulmonary complications but no better than clinical examination. Abnormal chest radiography has shown to correlate with pulmonary complications, but was considered to be potentially helpful only in patients with established cardiopulmonary disease or those over the age of 50. Cessation of smoking was also associated with a modest reduction in post-operative pulmonary complications, but only if instigated at least two months before surgery. There was some evidence that acute mental state change (excluding stable mental disease or dementia), alcohol intake, impaired renal function and weight loss also had a moderate correlation. Obesity, asthma and oropharyngeal bacterial colonisation had no predictive effect on pulmonary complication rate. There was not enough evidence to conclude whether exercise capacity, diabetes or HIV had any effect.

In this review about half of the studies, which included respiratory comorbidity as a predictor, found an association with mortality but again there was no consistency in marker definitions, which included a history of respiratory disease, spirometry and smoking history. Given the frequency of post-operative respiratory complications and their association with mortality, together with the published evidence, respiratory comorbidity should at least be considered as candidate predictors. However, the most useful method of including respiratory risk is unclear and therefore this predictor is at an exploratory stage.

Exercise and activity capacity

Seven (Law *et al.*, 1994; Bartels *et al.*, 1998; Bailey *et al.*, 2003; Rentz *et al.*, 2003; Murray *et al.*, 2007; Forshaw *et al.*, 2008; Healy *et al.*, 2008) of eight studies, which investigated exercise or activity capacity found an association with perioperative mortality, however these included a wide range of assessment tools from self-reported levels of daily activity, through to scores of general wellness to objectively measured cardiorespiratory capacity. Although one would expect some form of exercise capacity to be included in a risk score, it is far from clear which is the most appropriate.

Other comorbidity predictors

Diabetes, renal function, liver function, alcohol intake, level of social deprivation, surgical procedure, and tumour site and stage were all variably associated with perioperative outcome. Findings were inconclusive, definitions varied and potential confounders not included in analyses. Interpretation and applicability to other datasets was difficult because the relative incidences of these predictors are likely to be small and their importance unclear.

Potential risk of bias in primary studies

In this section I discuss some issues pertaining to risk of bias in the primary studies and I have grouped these into main categories as described in Hayden's study of systematic reviews for prognostic studies (Hayden *et al.*, 2006).

Does the sample adequately represent the population of interest?

Selection bias in study samples may prevent reliable generalisation to new populations. For example, different approaches between centres, as to what is an acceptable level of 'fitness for surgery' may introduce selection bias. Since this is potentially related to predictors for this prediction model and unlikely to be controllable, the reasons for excluding patients from surgery should be reported.

The ideal unbiased sample would be an unselected consecutive series described as, for example, "100 consecutive cases". Only a minority of studies clearly reported a consecutive case series or case selection procedures and it was frequently unclear whether there could have been selection bias. Because of the lack of reporting clarity, I may have overestimated the real risk of potential for selection bias in some of these studies.

'Loss to follow up' can produce biased underestimates of mortality in longitudinal cohort studies (Butler *et al.*, 2001) but it is not clear what the scale of effect may be. A simulation study (Kristman *et al.*, 2004) demonstrated that small losses (from 5%), which occur 'not at random', i.e. are influenced by the outcome, can bias binary outcome estimates in cohort studies, and large effects can occur if follow up losses exceed 20% and depend on both outcome and prognostic variables. If losses to follow up are completely random or depend only on prognostic factors, considerable losses can be incurred without biasing the results, although precision will be affected.

I specified perioperative mortality as periods of follow up from 30 to 90 days and "in hospital mortality". Adequate follow up rate was considered to be met if there was a clear statement that all fatalities and survivors were accounted for; this was considered to "partially met" if it required deduction from the results. Although follow up rates in this type of study from single centres are likely to reasonably complete, a statement of mortality rate alone cannot exclude the possibility that cases may have been lost to follow up, particularly mortality rates within a set period (e.g. 30 day mortality), when patients could have died after discharge home or to a facility providing a lower level of care. Unless it was reported that all survivors are known to have been accounted for, I considered that follow up was not fully reported.

Does the data represent the sample?

Fourteen studies were retrospective but in twenty, it was unclear whether data collection was retrospective or prospective. Retrospective studies are open to bias from case selection, missing records, and misclassification error (Sackett *et al.*, 2006). A possible solution would be to 'blind' data collectors to outcome but this would be very resource intensive, and impractical for a clinical database. A more realistic solution is to include data validation procedures, however only eight studies gave clear evidence of data validation by data audit or checking with other data sources. This was reported more frequently, in presumably better resourced and planned population database or prospective studies.

Prognostic variables were generally described in enough detail to be reproducible. However definitions (e.g. cardiac and respiratory morbidity) varied between studies, making comparisons difficult. About half the studies recorded other potentially important predictors (comorbidity, tumour histology and stage, neoadjuvant therapy) but considerably less took account of all important variables in analysis.

These studies were observational and randomisation was not part of study design. Without randomisation, confounding variables may distribute in a non-random way. Ideally, known confounders should be reported and accounted for at analysis; for instance, age could be a confounder for cardiac or respiratory disease. Many studies reported the distribution of some but not all important predictors between groups, e.g. age categories, but they were not generally included in the analysis.

Only eight studies reported missing values and their handling methods and no studies reported whether patients with missing values were representative of the main sample in other important characteristics. Reporting missing data is important because it can both impair study efficiency by reducing the effective sample size, and introduce bias if the extent of missing data is associated with outcome (Steyerberg, 2009f).

Data analysis

A majority of studies appeared to use appropriate statistical methods for data analysis, but details were frequently scanty and some difficult to follow. Regression methods were the commonest methods used in generating prognostic models; however it was commonly difficult for the reviewers to find details in the manuscript. Most studies which used regression did not report collinearity between variables, model fitting, or interactions between variables e.g. age and cardiorespiratory morbidity.

Small samples can exacerbate the problems of overoptimistic statistical significance and result in unrealistic effect sizes in multiple regression methods, particularly those which rely on ‘data driven’ selection methods (Steyerberg *et al.*, 1999). Samples should be large enough to account for the multiple comparisons, interactions between variables and the use of categorical variables. Sample size in studies with binary outcomes are driven by the number of outcome events and prevalence of predictor and outcome; it has been suggested that adequate samples should have at least ten (some have suggested twenty) events for each potential variable investigated (Vach, 1997).

In this review, sample sizes were frequently too small to detect potentially important differences in outcomes, particularly when age had been investigated as a categorical variable.

Synthesis of results

Data synthesis of pooled results can be carried out for appropriate summary statistics of predictor effects (Deeks, 2001). However, there was much heterogeneity in study design, reporting of summary statistics, sample characteristics and prognostic variable definition. There was also considerable variation in reporting potential for bias and therefore data pooling was inappropriate for most predictors. However, studies of age categories below and above 70 years were reported in a way to enable data pooling and I did this because all of these studies reported no significant effect of age, but used small sample sizes. Pooling the data resulted in a summary odds ratio of 1.9 for age over 70. However, care should be exercised in interpreting this because the proportion of very elderly and those with other comorbidities in each category was not always obvious.

Strengths and weaknesses of this review

I carried out this systematic review using recommended methods (Altman, 2001; Hayden *et al.*, 2006; Centre for Reviews and Dissemination, 2009; Moher *et al.*, 2009) and focussed on studies which, as far as possible, might apply to patients currently managed in our unit. The strengths and potential weaknesses within the review are discussed below.

Review Methods

The methodology for systematic reviews of prognostic studies is not as well developed as that for interventional studies. There are summary guidelines from the Centre for Reviews and Dissemination , a Cochrane group is developing methodology (Centre for Reviews and Dissemination) and there are reporting guidelines (Moher *et al.*, 2009) but much is unvalidated including methods of optimal search strategies and assessment of potential for bias.

A prospective review protocol carried out in a 'linear' fashion is an ideal method of reducing bias. Inevitably, issues arose during the review process which required protocol revision and the process then became iterative rather than linear. For example, some studies known to the reviewers were not retrieved by the search strategy and during the searches important predictors became apparent (e.g. surgical volume), requiring protocol revision. During data extraction, some items, which had been defined to assess potential for bias, also required redefinition, because of interpretation difficulties for the data extractors. These iterative processes clearly leave room for potential bias within the review, but have been recognised by other investigators and some degree of iteration is considered inevitable (Moher *et al.*, 2009) (Pope *et al.*, 2007).

Article selection

There was a dilemma as to whether to retrieve studies which closely matched our population of interest or whether to search a broader literature ("splitting" vs "lumping") (Pope *et al.*, 2007). I chose the broader approach because this would increase our chances of determining whether comorbidities had a general effect, and we were not clear whether factors, which defined our local population (e.g. type of surgical procedure), were important. A broader approach also potentially reduces the potential for bias, which may be inherent in selecting particular populations (Grimshaw *et al.*, 2003). However, it became clear that certain criteria, which define our local centre, also affected outcome, and these were used as inclusion criteria (e.g. surgical volume). I also confined selection to studies from 'high volume' centres published or mainly carried out after 1990, and the reasons for this have been discussed above.

Publication bias

Publication bias occurs when the published research literature is systematically unrepresentative of the population of completed studies (Rothstein *et al.*, 2005). This occurs in various forms and aspects of this review could have been at risk of this bias. The most well known is the association of favourable outcome, large effects or statistical significance with publication. This is particularly so for randomised trials (Song *et al.*, 2000) but has also been shown in prognostic studies of Barrett's oesophagus (Shaheen *et al.*, 2000). Conversely, studies of prognostic models and their constituent variables, with unfavourable results may not be published. Outcome bias can arise from selective reporting of study methods and results. Several of these biases have been reported in studies of mortality after oesophagectomy, including reporting '30 day' rather than 'in hospital' mortality, variable patient selection for surgery, selective reporting of denominator values, and institutional selection (Jamieson *et al.*, 2004).

Language bias is another potential cause of publication bias. Specifically, studies with statistically significant results are more likely to be published in English (Rothstein *et al.*, 2005). In our review, only articles written in English were included because the reading population of interest is most likely to be English speaking or publish in English.

Studies with statistically significant results are more likely to be cited by others (Rothstein *et al.*, 2005). Part of our search strategy included hand searching book chapters, reviews and the reference lists within the articles selected, and therefore citation bias was a potential risk in this review.

My searches focussed on the electronic bibliographic databases MEDLINE and EMBASE. Searches of electronic databases routinely retrieve only a fraction of the available studies because of imperfections in search algorithms. This is possibly a greater problem for prognostic study searches because they are at an earlier stage of development. Considerable numbers of studies may be published in sources not indexed with the major electronic databases, the so called 'grey' literature. Our searches did not include a full search of all grey literature sources, however we (IW/MC) hand searched book chapters, review articles and the articles retrieved from the electronic searches, even though these may be susceptible to their own biases. We contacted local known experts (Dr Ian Shaw) for further information, and we did not contact original study authors for study details because of resource constraints.

Data extraction

There were differences between the reviewers in interpretation of risk of bias items, requiring discussion and in some cases revision of definitions. This may have been exacerbated by a lack of clarity in item definition, but difficulty in finding important items because of poor reporting caused considerable difficulty.

2.5 Key findings

1. Several prediction models have been developed to predict perioperative mortality after oesophagectomy. Performance on external validation has been disappointing; overestimation of mortality rates has been common and none have been subject to formal prospective clinical impact studies. Only the Rotterdam (Steyerberg *et al.*, 2006) model is transferrable to our data.
2. Age should be considered a candidate predictor for any prediction model and included as a continuous variable, particularly to incorporate the effect of more extreme old age.
3. There is clinical knowledge to support the inclusion of other comorbidity predictors such as cardiorespiratory, nutritional and physical capacity measures as candidate predictors, but the evidence from this review is weak. This was compounded by a variety of inconsistent definitions of predictors and small sample size studies.
4. Several important risks of potential bias were poorly reported or not addressed in the primary studies. These included potential selection bias, data validity, and missing data.

Table 12 Characteristics of studies fulfilling inclusion criteria[1]

	Author **publication date**	**Study design; study period; sample size (n)**	**Perioperative mortality definition; mortality rate (%)**	**Geographical location, number of centres or databases, average annual operative volume**	**Histology;** **use of neoadjuvant therapy;** **surgical procedure**
Prediction models	Bailey (Bailey *et al.*, 2003)	Clinical prediction model, internal validation 1991 to 2000 n= 1777	30 day Mortality 10.0%	USA Population database; 109 centres Operative volume not applicable	Histology: all malignant (no detail) except 268 benign Neoadjuvant na Surgical procedure: na
	Steyerberg (Steyerberg *et al.*, 2006)	Clinical prediction model, external validation 1991 to 2002 n= 3592	30 day Mortality 8%	USA/Netherlands Population database and clinical centre Operative volume na/yr	Histology: 2118 (a), 1307 (s), 164 (o) Neoadjuvant 878 (neo) Surgical procedure: na
	Ra (Ra *et al.*, 2008)	Clinical prediction model, internal validation 1997 to 2003 n= 1172	In hospital & 30 day Mortality 13.7%	USA Population database Operative volume not applicable	Histology: na Neoadjuvant na Surgical procedure: na
	Tekkis (Tekkis *et al.*, 2004)	Clinical prediction model, internal validation 1994 to 2000 n= 538	In hospital Mortality 8.6%	UK Regional & national clinical databases (36 centres) Operative volume 77/yr	Histology: 317 (a), 118 (s), 103 (o) Neoadjuvant na Surgical procedure: 45 TH, 297 Rt 2-stage , 106 thoracoabdominal, 22 3-stage, 68 other
	Law (Law *et al.*, 1994)	Clinical prediction model 1982 to 1992 n= 523	In hospital & 30 day Mortality 15.5%	Hong Kong Single centre Operative volume 63/yr	Histology: All (s) Neoadjuvant na Surgical procedure: 80 TH, 303 LTan, 2 LT, 45 3-phase, 18 split sternum, 43 E, 32 phary'laryngo'esophagectomy

[1] (Abbreviations: Histology; (a) adenocarcinoma, (s) squamous carcinoma, (o) other. Operation; TT transthoracic, T thoracotomy, RT right thoracotomy, LT left thoracotomy, lap laparotomy, Abd abdominal, TH transhiatal, LTan Lewis-Tanner, M McKeown, LTA left thoracoabdominal, IL Ivor-Lewis, E esophagogastrectomy, Tscopic thoracoscopic, CTA cervico-thoracoabdominal)

Author publication date	Study design; study period; sample size (n)	Perioperative mortality definition; mortality rate (%)	Geographical location, number of centres or databases, average annual operative volume	Histology; use of neoadjuvant therapy; surgical procedure
Liu (Liu *et al.*, 2000)	Clinical pred[2]iction model 1994 to 1997 n= 32	In hospital Mortality 13%	Australia Single centre Operative volume 25/yr	Histology: na Neoadjuvant na Surgical procedure: 2 TH, 29 IL, 1 3-stage
Bartels (Bartels *et al.*, 1998)	Clinical prediction model, external validation, clinical application 1982 to 1996 n= 764	30 day & 90 day Mortality 15.2%	Germany Single centre Operative volume 56/yr	Histology: na Neoadjuvant na Surgical procedure: "transmediastinal" for adeno, TT for squamous, no quantities
Sanz (Sanz *et al.*, 2006)	Clinical prediction model 1987 to 1999 n= 114	In hospital Mortality 12.3%	Spain Single centre Operative volume 9/yr	Histology: 39 (a), 73 "epidermoid", 2 (o) Neoadjuvant na Surgical procedure: 13 TH, 101 TT
McCulloch P (McCulloch *et al.*, 2003)	Predictor effect study 1999 to 2002 n= 365	In hospital Mortality 12%	UK Subset of ASCOT National database(multiple centres reporting gastric and oesophageal surgery) Operative volume Centres allocated to one of 3 categories of volume	Histology: na Neoadjuvant 26% Surgical procedure: IL 67.1%, LTA 6.8%, TT 10.4%, M(3 stage) 7.9%, other 7.7%
Zhang (Zhang *et al.*, 1994)	Clinical prediction model, external validation, clinical application 1986 to 1989 n= 100	45 day Mortality 13%	Japan Single centre Operative volume 37/yr	Histology: na Neoadjuvant na Surgical procedure: All Lap & RT (assumed)

[2] (Abbreviations: Histology; (a) adenocarcinoma, (s) squamous carcinoma, (o) other. Operation; TT transthoracic, T thoracotomy, RT right thoracotomy, LT left thoracotomy, lap laparotomy, Abd abdominal, TH transhiatal, LTan Lewis-Tanner, M McKeown, LTA left thoracoabdominal, IL Ivor-Lewis, E esophagogastrectomy, Tscopic thoracoscopic, CTA cervico-thoracoabdominal)

	Author publication date	Study design; study period; sample size (n)	Perioperative mortality definition; mortality rate (%)	Geographical location, number of centres or databases, average annual operative volume	Histology; use of neoadjuvant therapy; surgical procedure
Validation studies	Schroder (Schroder *et al.*, 2006)	Predictor effect study, external validation 1997 to 2002 n= 126	In hospital Mortality 5.6%	Germany Single centre Operative volume 21/yr	Histology: 68 (a), 54(s), 4(o) Neoadjuvant 46 chemotherapy Surgical procedure: 126 IL
	Lai (Lai *et al.*, 2007)	External validation, comparison 2001 to 2005 n= 545	In hospital Mortality 5.5%	Hong Kong Administrative database (14 centres) Operative volume na/yr	Histology: All (s) Neoadjuvant na Surgical procedure: "thoracic"
	Nagabhushan (Nagabhushan *et al.*, 2007)	External validation, comparison 1990 to 2002 n= 110	30 day Mortality 10.2	UK Single centre Operative volume 9/yr	Histology: 80 (a), 29 (s), 1 (o) Neoadjuvant na Surgical procedure: 27 TH, 55 IL, 28 other
	Lagarde (Lagarde *et al.*, 2007)	External validation 1993 to 2005 n= 663	In hospital Mortality 3.6%	Netherlands Single centre Operative volume 52/yr	Histology: 476 (a), 187 (s) Neoadjuvant 114 (neo) Surgical procedure: 424 TH, 239 TT
	Zafirellis (Zafirellis *et al.*, 2002)	External validation 1990 to 1999 n=[3] 204	30 day Mortality 12.8%	UK Single centre Operative volume 21/yr	Histology: 156 (a), 45 (s), 3 (o) Neoadjuvant 39 (neo) Surgical procedure: 9 TH, 158 IL, 7 M, 22 Tscopic, 8 LTA
	Zingg (Zingg *et al.*, 2009)	External validation, comparison 1990 to 2007 n= 346	In hospital & 30 day Mortality Australia 8.0%, Switzerland 4.7%	Australia/Netherlands/Switzerland Two centre Operative volume Australia (unclear); Zurich 9.4/yr	Histology: Aus 76%(a), 16%(s), Switz 73%(a), 42%(s) Neoadjuvant Aus (54.5%), Switz 25.9% Surgical procedure: All TT

[3] (Abbreviations: Histology; (a) adenocarcinoma, (s) squamous carcinoma, (o) other. Operation; TT transthoracic, T thoracotomy, RT right thoracotomy, LT left thoracotomy, lap laparotomy, Abd abdominal, TH transhiatal, LTan Lewis-Tanner, M McKeown, LTA left thoracoabdominal, IL Ivor-Lewis, E esophagogastrectomy, Tscopic thoracoscopic, CTA cervico-thoracoabdominal)

	Author publication date	Study design; study period; sample size (n)	Perioperative mortality definition; mortality rate (%)	Geographical location, number of centres or databases, average annual operative volume	Histology; use of neoadjuvant therapy; surgical procedure
Candidate predictor effect studies	Han-Geurts (Han-Geurts *et al.*, 2006)	Predictor effect study 1996 to 2003 n= 400	In hospital Mortality 5.5%	Netherlands Single centre Operative volume 50/yr	Histology: 277 (a), 118(s), 59(o) Neoadjuvant 174 Surgical procedure: TH for distal tumour, TT and Abd for proximal, no quantities
	Sabel (Sabel *et al.*, 2002)	Pr[4]edictor effect study 1991 to 1998 n= 117	30 day Mortality 2.6%	USA Two Operative volume 15/yr	Histology: 93 (a), 24 (s) Neoadjuvant 104 (adj) (34 neoadjuvant) Surgical procedure: 3 TH, 98 IL
	Tsai (Tsai *et al.*, 2003)	Predictor effect study 1985 to 2000 n= 785	30 day Mortality 5.2% (30 day)	Taiwan Single centre Operative volume 49/yr	Histology: All (s) Neoadjuvant none Surgical procedure: RT ("in most")
	Rahamim (Rahamim *et al.*, 2003)	Predictor effect study 1979 to 1999 n= 596	30 day Mortality 8.2%	UK Single centre Operative volume 30/yr	Histology: 378 (a), 185 (s), 33 (o) Neoadjuvant 71 (adj), 19 (neo) Surgical procedure: 1 TH, 518 IL, 54 LTA, 23 M
	Moskovitz (Moskovitz *et al.*, 2006)	Predictor effect study 1996 to 2005 n= 751	In hospital & 60 day Mortality % 5.8	USA Single centre Operative volume 91/yr	Histology: na Neoadjuvant na Surgical procedure: 569 T, 201 Abdo
	Finlayson (Finlayson *et al.*, 2007)	Predictor effect study 1994 to 2003 n= 27957	In hospital Mortality 12.5%	USA Population database Operative volume not applicable	Histology: na Neoadjuvant na Surgical procedure: na
	Law (Law *et al.*, 2004)	Predictor effect study 1990 to 2001 n= 421	In hospital & 30 day Mortality 4.8%	Hong Kong Single centre Operative volume 35/yr	Histology: All (s) Neoadjuvant 143 (includes chemoradiation) Surgical procedure: 219 LT, 86 3 phase, 44 E, 39 TH, 25 Tscopic, 5 split sternum, 3 staged
	Fang (Fang *et al.*, 2001)	Predictor effect study 1986 to 1998 n= 441	In hospital & 30 day Mortality 4.1%	Japan Single centre Operative volume 34/yr	Histology: about 90% (s), about 7% (o) Neoadjuvant 27 Surgical procedure: All CTA

[4] (Abbreviations: Histology; (a) adenocarcinoma, (s) squamous carcinoma, (o) other. Operation; TT transthoracic, T thoracotomy, RT right thoracotomy, LT left thoracotomy, lap laparotomy, Abd abdominal, TH transhiatal, LTan Lewis-Tanner, M McKeown, LTA left thoracoabdominal, IL Ivor-Lewis, E esophagogastrectomy, Tscopic thoracoscopic, CTA cervico-thoracoabdominal)

Author **publication date**	**Study design; study** **period; sample size (n)**	**Perioperative mortality** **definition; mortality rate (%)**	**Geographical location, number of** **centres or databases, average** **annual operative volume**	**Histology;** **use of neoadjuvant therapy;** **surgical procedure**
Alexiou (Alexiou *et al.*, 1998)	Predictor effect study 1987 to 1997 n= 166	In hospital & 30 day Mortality 6.0%	UK Single centre Operative volume 18/yr	Histology: All (s) Neoadjuvant None Surgical procedure: 9 TH, 101 IL, 45 LT
Kinugasa (Kinugasa *et al.*, 2001)	Pred[5]ictor effect study 1981 to 1999 n= 204	60 day Mortality 6.9%	Japan Single centre Operative volume 11/yr	Histology: na Neoadjuvant 154 (neo) Surgical procedure: 204 tot, 193 R TT , 8 L TT, 3 blunt esophagectomy
Ferguson (Ferguson *et al.*, 1997)	Predictor effect study 1980 to 1995 n= 269	In hospital & 30 day Mortality 13%	USA Single centre Operative volume 17/yr	Histology: na Neoadjuvant na Surgical procedure: 54 TH, 110 LT, 81 Lap & RT, 24 other
Rentz (Rentz *et al.*, 2003)	Predictor effect study 1991 to 2000 n= 945	30 day Mortality 10%	USA Population database; 109 centres Operative volume 95/yr	Histology: na Neoadjuvant na Surgical procedure: 383 TH, 562 TT
Ruol(a) (Ruol *et al.*, 2007(b))	Predictor effect study 1992 to 2005 n= 269	In hospital & 30 day Mortality 2.2%	Italy Single centre Operative volume 59/yr	Histology: 22% (a), 78% (s) Neoadjuvant 182 chemo- & radio-, 83 chemo-, 4 radiotherapy Surgical procedure: McKeown 97, IL 126, Lap+L cervicotomy 46) , IL for mid/lower , McKeown for upper, no quantities
Alexiou (Alexiou *et al.*, 2005)	Predictor effect study 1987 to 1997 n= 523	In hospital & 30 day Mortality 5.3%	UK Single centre Operative volume 60/yr	Histology: 339(a), 166 (s), 18 (o) Neoadjuvant na Surgical procedure: 28TH, 146IL, 276LT, 71 Lthoracolaparotomy, 2 Mckeown
Adam (Adam *et al.*, 1996)	Predictor effect study 1982 to 1992 n= 597	30 day Mortality 6.9%	UK Single centre Operative volume 60/yr	Histology: 370 (a), 216(s), 11(o) Neoadjuvant na Surgical procedure: 13 TH, 584 TT (573 Lap & LT)

[5] (Abbreviations: Histology; (a) adenocarcinoma, (s) squamous carcinoma, (o) other. Operation; TT transthoracic, T thoracotomy, RT right thoracotomy, LT left thoracotomy, lap laparotomy, Abd abdominal, TH transhiatal, LTan Lewis-Tanner, M McKeown, LTA left thoracoabdominal, IL Ivor-Lewis, E esophagogastrectomy, Tscopic thoracoscopic, CTA cervico-thoracoabdominal)

Author publication date	Study design; study period; sample size (n)	Perioperative mortality definition; mortality rate (%)	Geographical location, number of centres or databases, average annual operative volume	Histology; use of neoadjuvant therapy; surgical procedure
Sauvanet (Sauvanet *et al.*, 2005)	Predictor effect study 1985 to 2000 n= 1192	In hospital Mortality 6.4%	France 37 centres Operative volume not available	Histology: All (a) Neoadjuvant 132 chemoradiation, 31 chemo, 3 radiotherapy Surgical procedure: 772 TT (636 Lap & RT, 128 Lap & LT, 8 Lap & TT & cervicot), 420 Lap(some TH)
Kuwano (Kuwano *et al.*, 1998)	Predictor effect study 1[6]989 to 1993 n= 178	In hospital & 30 day Mortality 3.4%	Japan Single centre Operative volume 36/yr	Histology: na Neoadjuvant na Surgical procedure: 173 Lap & RT, 5 non T
Griffin (Griffin *et al.*, 2002)	Predictor effect study 1990 to 2000 n= 228	In hospital & 30 day Mortality 4.0%	UK Single centre Operative volume 23/yr	Histology: 146 (a), 75 (s), 7 (o) Neoadjuvant na Surgical procedure: 228 IL
Abunasra (Abunasra *et al.*, 2005)	Predictor effect study 1990 to 2003 n= 652	In hospital & 30 day Mortality 5.6% (30 day)	UK Single centre Operative volume 58/yr	Histology: 523(a), 238 (s), 37 (o) Neoadjuvant None Surgical procedure: 17 TH, 202 IL, 412 LT, 135 Lap & LT, 7 M
Ruol(b) (Ruol *et al.*, 2007(b))	Predictor effect study 1992 to 2005 n= 739	In hospital & 30 day Mortality 2.6%	Italy Single centre Operative volume 57/yr	Histology: 449(s), 260(a), 30(o) Neoadjuvant na Surgical procedure: 487 IL, 103 Lap & cervicotomy, 149 M
Jougon (Jougon *et al.*, 1997)	Predictor effect study 1980 to 1993 n= 540	In hospital Mortality 5.7%	France Single centre Operative volume 39/yr	Histology: 214 (a), 307 (s), 19 (o) Neoadjuvant 47 (neo) Surgical procedure: 5 TH, 181 IL, 216 Lap & LT, 131 Lap & RT & cervicot, 7 other
Thomas (Thomas *et al.*, 1996)	Predictor effect study 1979 to 1994 n= 386	In hospital & 30 day Mortality 11.1%	France Single centre Operative volume 34/yr	Histology: 132 (a), 254 (s), 51 (o) Neoadjuvant 51 chemotherapy Surgical procedure: 153 Lap & RT & cervicot, 82 Lap & RT, 125 Lap & cervicot, 26 LT,

[6] (Abbreviations: Histology; (a) adenocarcinoma, (s) squamous carcinoma, (o) other. Operation; TT transthoracic, T thoracotomy, RT right thoracotomy, LT left thoracotomy, lap laparotomy, Abd abdominal, TH transhiatal, LTan Lewis-Tanner, M McKeown, LTA left thoracoabdominal, IL Ivor-Lewis, E esophagogastrectomy, Tscopic thoracoscopic, CTA cervico-thoracoabdominal)

Author publication date	Study design; study period; sample size (n)	Perioperative mortality definition; mortality rate (%)	Geographical location, number of centres or databases, average annual operative volume	Histology; use of neoadjuvant therapy; surgical procedure
Ellis Jr (Ellis Jr *et al.*, 1998)	Pre[7]dictor effect study 1970 to 1997 n= 505	In hospital & 30 day Mortality 3.0%	USA Two centres Operative volume 19/yr	Histology: 335 (a), 155 (s), 15 (o) Neoadjuvant 46 (neo) Surgical procedure: 103 TH, 147 IL, 169 LT, TA 11, 25 other, (some unresected)
Poon (Poon *et al.*, 1998)	Predictor effect study 1982 to 1996 n= 737	In hospital & 30 day Mortality 10.6%	Hong Kong Single centre Operative volume 50/yr	Histology: 22 (a), 668 (s), 47 (o) Neoadjuvant na Surgical procedure: 113 TH, 608 TT, 16 Tscopic
Atkins (Atkins *et al.*, 2004)	Predictor effect study 1996 to 200 n= 379	30 day Mortality % 5.80474934	USA Single centre Operative volume 54/yr	Histology: 228 (a), 70 (s), 17 (o) Neoadjuvant na Surgical procedure: 130 TH, 179 IL, 70 other(inc 35 LT)
Forshaw (Forshaw *et al.*, 2008)	Predictor effect study 2004 to 2006 n= 78	In hospital (not explicit) Mortality 1.3%	UK Single centre Operative volume 28/yr	Histology: 58 (a), 13 (s), 7 (o) Neoadjuvant 50 (neo) Surgical procedure: 39 TH, 29 2-stage (23 lap assisted) , 5 3-stage, 5 LTA
Gockel (Gockel *et al.*, 2005)	Predictor effect study 1985 to 2004 n= 424	In hospital ("mortality rate") & 30 day Mortality 11.5%	Germany Single centre Operative volume 23/yr	Histology: 152 (a), 234 (s), 38 (o) Neoadjuvant na Surgical procedure: 186 TH, 231 Lap & TT , 7 free jej graft
Murray (Murray *et al.*, 2007)	Predictor effect study 2002 to 2005 n= 51	30 day Mortality 9.8%	UK Single centre Operative volume 31/yr	Histology: na Neoadjuvant na Surgical procedure: na
Saito (Saito *et al.*, 1993)	Predictor effect study, external validation 1983 to 1991 n= 99	In hospital & 30 day Mortality 13.3% (10/32 first period, 3/67 second period)	Japan Single centre Operative volume 11/yr	Histology: na Neoadjuvant na Surgical procedure: All TT

[7] (Abbreviations: Histology; (a) adenocarcinoma, (s) squamous carcinoma, (o) other. Operation; TT transthoracic, T thoracotomy, RT right thoracotomy, LT left thoracotomy, lap laparotomy, Abd abdominal, TH transhiatal, LTan Lewis-Tanner, M McKeown, LTA left thoracoabdominal, IL Ivor-Lewis, E esophagogastrectomy, Tscopic thoracoscopic, CTA cervico-thoracoabdominal)

Author publication date	Study design; study period; sample size (n)	Perioperative mortality definition; mortality rate (%)	Geographical location, number of centres or databases, average annual operative volume	Histology; use of neoadjuvant therapy; surgical procedure
Whooley (Whooley *et al.*, 2001)	Predictor effect study 19[8]82 to 1998 n= 710	In hospital & 30 day Mortality 11%	Hong Kong Single centre Operative volume 42/yr	Histology: All (s) Neoadjuvant na Surgical procedure: 119 TH, 414 LT, 93 3-phase, 63 E, 21 split sternum
Johansson (Johansson and Walther, 2000)	Predictor effect study 1984 to 1996 n= 139	In hospital & 30 day Mortality 1.4%	Sweden Single centre Operative volume 13/yr	Histology: 57 (a), 60 (s), 22 (o) Neoadjuvant 6 (neo) Surgical procedure: All Lap & RT (+/_ cervicot)
Healy (Healy *et al.*, 2008)	Predictor effect study 1999 to 2005 n= 169	In hospital Mortality 4.3%	Ireland Single centre Operative volume 29/yr	Histology: 118 (a), 52 (s), 6 (o) Neoadjuvant 48% (neo) Surgical procedure: 3 TH, 111 Lap & RT, 41 3-stage, 9 E, 5 total gastrectomy
Leigh (Leigh *et al.*, 2006)	Predictor effect study 2001 to 2004 n= 93	30 day & 90 day Mortality 8.3%	UK Population database Operative volume not applicable	Histology: na Neoadjuvant na Surgical procedure: na
Alibhakshi (Alibakhshi *et al.*, 2009)	Predictor effect study 2000 to 2006 n= 480	In hospital Mortality 2.9%	Iran Single centre Operative volume 77/yr	Histology: 29(a), 451(s) Neoadjuvant excluded from study Surgical procedure: TH (lower third) 286, IL or M 194(mid or upper)
Braiteh (Braiteh *et al.*, 2009)	Predictor effect study 1999 to 2005 n= 621	30 day & in hospital Mortality 3.4%	USA Single centre Operative volume 62/yr	Histology: 539(a), 61(s) Neoadjuvant 400 (chemo, 8 chemo, 1 rad) Surgical procedure: RT & unknown other
Park (Braiteh *et al.*, 2009)	Predictor effect study 1995 to 2007 n= 7277	In hospital Mortality 11%	UK ICNARC database Operative volume not applicable	Histology: na Neoadjuvant na Surgical procedure: na

[8] (Abbreviations: Histology; (a) adenocarcinoma, (s) squamous carcinoma, (o) other. Operation; TT transthoracic, T thoracotomy, RT right thoracotomy, LT left thoracotomy, lap laparotomy, Abd abdominal, TH transhiatal, LTan Lewis-Tanner, M McKeown, LTA left thoracoabdominal, IL Ivor-Lewis, E esophagogastrectomy, Tscopic thoracoscopic, CTA cervico-thoracoabdominal)

Author publication date	Study design; study period; sample size (n)	Perioperative mortality definition; mortality rate (%)	Geographical location, number of centres or databases, average annual operative volume	Histology; use of neoadjuvant therapy; surgical procedure
Tagagawa (Takagawa *et al.*, 2008)	Predictor effect study 1994 to 2004 n= 222	In hospital Mortality 4%	Japan Single centre Operative volume 42/yr	Histology: 200(s), 22(o) Neoadjuvant 190 Surgical procedure: RT 198, TH 24
Takeno (Takeno *et al.*, 2008)	Predic[9]tor effect study 1990 to 2001 n= 70	In hospital (unclear definition) Mortality 6%	Japan Single centre Operative volume 17/yr	Histology: All (s) Neoadjuvant na Surgical procedure: TT 51, other 19

[9] (Abbreviations: Histology; (a) adenocarcinoma, (s) squamous carcinoma, (o) other. Operation; TT transthoracic, T thoracotomy, RT right thoracotomy, LT left thoracotomy, lap laparotomy, Abd abdominal, TH transhiatal, LTan Lewis-Tanner, M McKeown, LTA left thoracoabdominal, IL Ivor-Lewis, E esophagogastrectomy, Tscopic thoracoscopic, CTA cervico-thoracoabdominal)

Table 13 Excluded studies and reasons for exclusion

Study	Reason for exclusion of study from review
Jiao(Jiao *et al.*, 2006), Bonavina(Bonavina *et al.*, 2003), Morgan(Morgan *et al.*, 2007), Alexiou(Alexiou *et al.*, 2006), Di Martino(Di Martino *et al.*, 2005)	Period of follow up for mortality undefined
Lund(Lund *et al.*, 1990), Gulliford(Gulliford *et al.*, 1993), Liedman(Liedman *et al.*, 1995), Charoenpan(Charoenpan *et al.*, 1993)	Data collection before 1990
Mokart(Mokart *et al.*, 2005)	Outcome “sepsis”
Ferguson(Ferguson *et al.*, 1997)	Outcome “respiratory complications”
Nozoe(Nozoe *et al.*, 2002), Karl(Karl *et al.*, 2000), Blazeby(Blazeby *et al.*, 2005a)	Predictor variable and mortality relationship unstated
Chamogeorgakis(Chamogeorgakis *et al.*, 2007)	Outcome in thoracic surgery for various conditions
Cariati(Cariati *et al.*, 2002), Golubovi(Golubovi and Golubovi, 2002)	Low surgical volume centres (average annual caseload 3 & 2.9)
Bartels(Bartels *et al.*, 2000)	Report of previous study with an extra 71 patients

Table 14 Studies of clinical prediction models

Author	Study design	Modelling method	Validation methods	Performance	Comments
Bailey (Bailey *et al.*, 2003)	Clinical prediction model	Multivariate logistic regression	Calibration & discrimination in modelling sample	C-index 0.69, Hosmer-Lemeshow 3.01 (p=0.93)	Predictors of 30 day mortality included age, diabetes, functional status, neoadjuvant, BUN, alcohol intake, ascites, alkaline phosphatase
Ra (Ra *et al.*, 2008)	Clinical prediction model; internal validation	Logistic regression;generation of risk score from SEER data[10]base	Comparison of predicted and observed mortality in modelling sample	Predicted & observed mortality reasonably matched in modelling sample but over predicted by about a quarter in high risk group	Predictors of mortality: age over 80, Charlson score (Charlson *et al.*, 1987), hospital surgical volume. Uncontrolled study of clinical application reduced mortality from 7% to 3%.

[10] Key: BUN, blood urea nitrogen; SEER, Surveillance Epidemiology and End Results National Cancer Institute (2012) *Surveillance Epidemiology and End Results*. Available at: http://seer.cancer.gov/.; ROC, receiver operator curve; AUC, area under curve; HL, Hosmer Lemeshow

Author	Study design	Modelling method	Validation methods	Performance	Comments
Steyerberg "Rotterdam" (Steyerberg *et al.*, 2006)	Clinical prediction model; external validation	Logistic regression model developed on Medicare-SEER database; validated on later SEER sample, Eindhoven Canc[11]er Registry & Rotterdam clinical database	Calibration and discrimination on modelling and validation samples	ROC: AUC for modelling cohort 0.66 (0.65 on internal validation); AUC range 0.56-0.7 in external validation cohorts. Calibration stated to be good for combined data (results not given)	

[11] Key: BUN, blood urea nitrogen; SEER, Surveillance Epidemiology and End Results ibid.; ROC, receiver operator curve; AUC, area under curve; HL, Hosmer Lemeshow

Author	Study design	Modelling method	Validation methods	Performance	Comments
Bartels "Munich" (Bartels *et al.*, 1998)	Clinical prediction model with validation and clinical application study on prospective samples from same centre	Correlation of preoperative predictors (and levels of abnormality) with composite out[12]come ('normal', 'prolonged', 'severe', 'fatal') between 1982 & 1991 (n=432). Discriminant analysis modelled 3 levels of organ dysfunction('normal', 'compromised', 'severely impaired') with postoperative mortality. Validated on 121 patients (1992-1993). Clinical application 1994-1996.	Descriptive comparison of predicted and observed mortality in same centre prospective sample.	3 risk groups for 30 day mortality in modelling sample: 3.6%, 8.7% and 28%. In prospective validation sample mortality was 2%, 5%, & 25%, and 5 of 9 deaths predicted as "high" risk.	30 day mortality in modelling sample predicted by Karnofsky index (Karnofsky, 1984), mental "cooperation", spirometry & arterial pO2, aminopyrine breath test, cirrhosis, cardiac risk.

[12] Key: BUN, blood urea nitrogen; SEER, Surveillance Epidemiology and End Results ibid.; ROC, receiver operator curve; AUC, area under curve; HL, Hosmer Lemeshow

Author	Study design	Modelling method	Validation methods	Performance	Comments
Law (Law *et al.*, 1994)	Clinical prediction model	Discriminant analysis identified risk factors; three level risk stratification based on sum of equ[13]ally weighted risk factors	Sensitivity and specificity in modelling data	Sensitivity 72%, specificity 74%, overall accuracy 74% on modelling data.	Predictors of hospital mortality: age, mid-arm circumference, operative blood loss, spirometry, abnormal chest xray, curative vs palliative resection. Risk scored groups with 7%, 30%, 38% mortality.

[13] Key: BUN, blood urea nitrogen; SEER, Surveillance Epidemiology and End Results ibid.; ROC, receiver operator curve; AUC, area under curve; HL, Hosmer Lemeshow

Author	Study design	Modelling method	Validation methods	Performance	Comments
Liu (Liu *et al.*, 2000)	Clinical[14] prediction model	Multiple regression; composite score stratified 3 levels of risk (mortality 50%, 27% & 8%)	No formal validation procedure		Multivariate regression: hypertension, smoking, spirometry predicted postoperative outcomes. Composite score of levels of predictor abnormality stratified mortality risk groups of 50%, 27% & 8% in modelling sample.
McCulloch (McCulloch *et al.*, 2003)	Clinical prediction model an validation on prospective sample from same data set	Multivariate logistic regression	Calibration (HL) & discrimination(ROC) on prospective sample of dataset	Modelling sample: C-index 0.79(0.03), Hosmer-Lemeshow 7.33 (p=0.5), O:E ratio 1.04 Validation sample: C-index 0.68(0.08), Hosmer-Lemeshow 7.39 (p=0.49), O:E ratio 0.82	For mixed oesophagogastric case-mix physiological POSSUM,surgeon's assessment, tumour stage, operation were predictors of outcome.

[14] Key: BUN, blood urea nitrogen; SEER, Surveillance Epidemiology and End Results ibid.; ROC, receiver operator curve; AUC, area under curve; HL, Hosmer Lemeshow

Author	Study design	Modelling method	Validation methods	Performance	Comments
Sanz (Sanz *et al.*, 2006)	Clinical prediction model	Discriminant analysis to generate 3 level risk score: 'low' (6.8% mortality), 'intermediate' (12.5% mortality), 'high' (50% mor[15]tality)	No validation procedures	na	Mortality associated with: Previous cancer, cirrhosis, abnormal spirometry, cholesterol, albumen. Composite risk score generated from these weighted variables. Discriminant analysis created three risk levels mortality rate 6.8%, 12.5% & 50% in modelling sample.
Zhang (Zhang *et al.*, 1994)	Clinical prediction model, external validation and clinical application	Logistic regression to develop composite risk score. Validated prospectively on same centre sample.	Sensitivity and specificity on modelling and validation samples	Modelling sample: sensitivity 0.75, specificity 0.99. Validation sample: sensitivity 0.33, specificity 0.98.	Mortality predicted by oral glucose tolerance test, tumour stage, age, abnormal ECG, creatinine clearance, operation type.

[15] Key: BUN, blood urea nitrogen; SEER, Surveillance Epidemiology and End Results ibid.; ROC, receiver operator curve; AUC, area under curve; HL, Hosmer Lemeshow

Author	Study design	Modelling method	Validation methods	Performance	Comments
Tekkis (Tekkis *et al.*, 2004)	Clinical prediction mod[16]el; internal validation & comparison with P-POSSUM	Univariate & multiple Bayesian logistic regression(include inter-hospital variation)	Calibration (HL); discrimination(ROC) on random sample (30%) of modelling data	C-index(95% CI): P-POSSUM 74.3(69.4, 79.2), single level O-POSSUM 74.6(69.9, 79.3); multilevel O-POSSUM 79.7(75.6, 83.8). Hosmer-Lemeshow P-POSSUM 28.8 (p=0.001), single level O-POSSUM 10.52 (p=0.23), multilevel O-POSSUM 10.15 (p=0.254).	Physiological POSSUM, age, urgency , POSSUM surgical stage predicted in-hospital mortality.
Schroder (Schroder *et al.*, 2006)	External validation of Bartels model	na	Estimation of mortality in risk groups defined by Bartels model; no statistical testing	Observed mortality 2.9% in predicted "low" , 3.0% in "moderate", 16.7% in "high" risk groups	Multivariate regression of variables used by Bartels identified age, general status and pulmonary function associated with composite outcome in Schroder's sample.

[16] Key: BUN, blood urea nitrogen; SEER, Surveillance Epidemiology and End Results ibid.; ROC, receiver operator curve; AUC, area under curve; HL, Hosmer Lemeshow

Author	Study design	Modelling method	Validation methods	Performance	Comments
Lai (Lai *et al.*, 2007)	External validation & comparison of POSSUM, O-POSSUM[17], P-POSSUM	na	Calibration (Chi-square); discrimination (ROC)	Calibration (Chi-square, for lack of fit): P-POSSUM (p=0.814), POSSUM(p<0.001), & O-POSSUM(p<0.002). POSSUM & O-POSSUM over predicted mortality by factor of 2.7 and 2.0 respectively. Discrimination(AUC, 95% CI): POSSUM 0.776 (0.689, 0.862), P-POSSUM 0.776(0.692, 0.861), O-POSSUM 0.676(0.586, 0.766).	

[17] Key: BUN, blood urea nitrogen; SEER, Surveillance Epidemiology and End Results ibid.; ROC, receiver operator curve; AUC, area under curve; HL, Hosmer Lemeshow

Author	Study design	Modelling method	Validation methods	Performance	Comments
Nagabhushan (Nagabhushan *et al.*, 2007)	External validation & comparison of O- and P-POSSUM	na	Calibration; discrimination	Calibration(HL): O-POSSUM (p=0.011), P-POSSUM (p= 0.019). Observed/expected mortality ratio 0.89 for P-POSSUM, 0.65 O-POSSUM. Mortality overestimated by a factor of 2 to 3 in higher risk group. Discrimination(AUC): P-POSSUM 0.68(0.59-0.76); and O-POSSUM 0.61(0.5-0.72)	
Lagarde (Lagarde *et al.*, 2007)	External v[18]alidation of O-POSSUM model	na	Calibration, discrimination, observed/predicted mortality	Observed/predicted mortality ratio 0.29; Hosmer-Lemeshow, p<0.001. Discrimination: (AUC, 95% CI): 0.6 (0.47-0.72).	

[18] Key: BUN, blood urea nitrogen; SEER, Surveillance Epidemiology and End Results ibid.; ROC, receiver operator curve; AUC, area under curve; HL, Hosmer Lemeshow

Author	Study design	Modelling method	Validation methods	Performance	Comments
Zafirellis (Zafirellis *et al.*, 2002)	External validation of POSSUM	na	Calibration, discrimination, observed/predicted mortality	ROC: AUC 0.62 (0.52-0.71). Observed/expected mortality ratio 0.66(95% CI 0.43-0.97). Hosmer-Lemeshow statistic: p=0.002	
Zingg (Zingg *et al.*, 2009)	Ext[19]ernal validation and comparison of Bartels (Munich), Steyerberg (Rotterdam) and Ra (Philadelphia) models.	na	Logistic regression of each risk model score on hospital mortality. Results given are p value for regression coefficient, Nagelkerke R-Squared, Hosmer-Lemeshow	Investigators concluded that no model could be applied generally	

[19] Key: BUN, blood urea nitrogen; SEER, Surveillance Epidemiology and End Results ibid.; ROC, receiver operator curve; AUC, area under curve; HL, Hosmer Lemeshow

Chapter 3: Preparation of a dataset and candidate predictors from the Northern Oesophago-Gastric Cancer Unit Clinical Database

3.1 Introduction

This chapter focuses on selecting and preparing a set of candidate predictors from the Northern Oesophagogastric Unit Database (NOGCU), with which to develop a clinical prediction model for perioperative mortality. I will describe the history and development of the database and the methods used to prepare the data and select the predictors.

3.1.1 The Northern Oesophagogastric Unit (NOGCU) database

This clinical database contains records on oesophagogastric cancer patients, who have been treated in the unit since 1990. Clinical information was recorded onto data entry forms by medical staff from the Unit, and transferred to a computerised database. Since 1996 the database has been run by a professional database manager, and more recently with the help of additional data entry staff. The recorded information includes a variety of demographic, pathological, comorbidity, treatment and outcome data. The number of data fields has reached as many as 199, however as decisions about relevance and redundancy have been made, modifications have been made and there are now about 130 fields. The database was initially maintained on Paradox software but was then moved to Microsoft Access and in 1996 was moved onto a Dendrite Clinical Systems database.

3.1.2 Selection of predictor variables to consider for the clinical prediction model

In the systematic review I identified several categories of patient specific predictor which should be considered for inclusion in the prediction model. These included age, operative procedure, tumour details, and several comorbidities (including cardiac, respiratory, nutritional, and exercise and activity capacity). I will consider predictors, which map to the categories above, have 'face validity' or have otherwise been reported to be associated with perioperative mortality in other non-cardiac surgery.

3.1.3 Managing Potential sources of bias in the NOGCU database

Potential sources of bias in prognostic and clinical prediction models are well recognised (Altman and Lyman, 1998; Hayden *et al.*, 2006). Several of these were identified in primary studies included in the systematic review and concerned the reporting of case selection, handling of missing data, and the use of categorical variables where continuous ones may have been better, e.g. for age at surgery. I will explore the extent to which these biases may be problematic, or may be addressed in the prediction modelling of NOGCU data.

3.1.4 Ethical considerations and Data protection

The Northern Oesophago-Gastric Unit database is registered with the Newcastle Hospitals Trust and has been considered by the Caldicott guardian, as required by the Data Protection Act, 1998. An individual's right to privacy in respect of personal data are enshrined in common law, The Human Rights Act 1988, the Health and Social Care Act 2001 and the Data Protection Act 1998. The requirements for the use of identifiable personal data have been summarised in a Parliamentary Postnote (Cant, 2005) and the practicalities of applying these requirements to epidemiological data have been described (Iverson *et al.*, 2006). For medical research, and historical or statistical data, ”the fair processing requirement” may be relaxed provided that “the data are not used to take any decision relevant to that particular individual, that subsequent publication does not lead to identification of the subject, and that it is unlikely to cause substantial damage or distress”(Commissioner, 2002).

3.1.5 Aim and objectives

I aim to prepare a set of predictors from the NOGCU database to include in a clinical prediction model of perioperative mortality after oesophagectomy.

3.2 Objectives

1. To ensure ethical and data protection requirements for this project are met.

2. To select and clean a set of candidate predictor variables from the main NOGCU database, for potential inclusion in a clinical prediction model of perioperative mortality after oesophagectomy for cancer.

3. To recognise and, if possible, address the main risks of bias within the data including:

a. Selection bias

b. Information bias.

c. The frequency and patterns of missing values.

3.3 Methods

3.3.1 Ethics, data protection and confidentiality

In November 2005, the local Research & Development Department advised that this study did not require full Ethical Committee consideration, as the project data would be anonymised. Since 2005, the Ethical Approval process was modified, so the study was resubmitted (December 2008), firstly for consideration by the local Research and Development department, and secondly, as recommended, to the National Research Ethics Service. They recommended that formal consideration by the local Ethics Committee was not required, as the project should be considered as "service evaluation and development" (Appendix D.).The data was used with the full cooperation and knowledge of Professor S M Griffin (professor of surgery in the Northern Oesophagogastric Unit).

3.3.2 Data storage

The subset of data for analysis from the NOGCU clinical database was saved as a Microsoft Office Excel (97-2003) spreadsheet. Patient names, addresses and medical record numbers were removed from the records, however, an NOCGU database 'key' was retained. All data was subsequently stored or transported in this form, on removable data storage devices. All copies of the project data were encrypted with Truecrypt v4.3a (www.truecrypt.org). The encrypted files and their backups were password protected with randomly generated 64 digit passwords, which were in turn stored on separate password protected storage devices.

3.3.3 General data management

1. Each record was allocated a unique identifier for the study. The database key and the medical record number were removed to maintain confidentiality.
2. The data was 'cleaned' using Microsoft Office Excel (97-2003) functions and custom code (Appendix F.). Whenever a field variable was moved for

analysis, integrity was maintained by checking alignment with the corresponding record identifier and 'date of birth field'.

3. Reports summarising fields and records from the original NOGCU dataset were generated.
4. Variables, which mapped to the following main categories, were selected for further examination and study: age, gender, tumour characteristics, surgical procedure, cardiorespiratory morbidity, other comorbidities (diabetes, renal or liver disease), exercise capacity, markers of nutritional state, all cause 'in-hospital' and '30' day mortality. Identifier fields, which were required to calculate new variables or maintain data integrity, were also selected (date of birth, unique key). These were deleted when the relevant data operations had been performed.

3.3.4 Data validation

Data validation was performed in SPSS (Release 17.0.0.; August 23, 2008). SPSS Frequency Analysis was used to examine data ranges and summary statistics.

1. Values of continuous data, which lay outside a prespecified plausible range were identified (*SPSS Data Editor/Validation/Validate*) and considered as missing. The limits of the acceptable ranges were set to represent values outside which, it was unlikely that patients would have been considered for surgery, or that the values could only have been errors.
2. Missing value patterns and frequencies were analysed (*SPSS Missing Value Analysis*).
3. Categorical values were validated using the Excel PivotTable function, which summarises all values that have been entered into a particular field. Only clearly erroneous values were excluded.
4. Where possible, free text field contents were recoded into appropriate categorical codes, which varied depending on individual variables.
5. Data validation was carried out between fields where possible. For example, related variables were cross checked (e.g. date of birth and reported age).

3.4 Results

The initial subset of data from the NOGCU included all patients who had undergone surgery between 5/4/1989 and 24/01/2006. This contained 199 field variables and 1246 patient records. Further data became available from 4/1/2006 to 27/01/2009 and contained 330 records. The merged final dataset contained 1576 records.

3.4.1 Out of range values (Table 15)

Acceptable ranges for variables were set to reflect values which were compatible with patients in reasonable health undergoing major elective surgery, and were set to allow a wide margin of error. Some values clearly fell outside these ranges and there were several possible explanations for these apparent errors. One example was patients who apparently survived uneventful surgery but had reported preoperative arterial pO_2 less than 5 kPa and pCO2 greater than 10 kPa. These values are barely compatible with life and appear to be the result of data transposed into the wrong columns. These values were labelled as missing. The option to trace medical records and 'patch' data was discounted because of resource limitation and previous experience of large scale retrospective audit with these records had proved very difficult, labour intensive and only moderately successful. Selective 'data patching' of obviously incorrect values may also have introduced bias as apparently normal values may also have been incorrectly entered and not checked.

Table 15 Description of variables from NOCGU database and handling of out of range (OOR) values

Variable Name(original database label)	Data procedures & comments	Out of range Rule (acceptable range given)	Number of out of range values (record identifier)
DEMOGRAPHIC DATA			
MyIndex(pid)	Unique key		NA
Gender(GENDER)	male=1, female=0		0
Age at surgery (AGE)	Calculated new numeric variable: 'operation date' minus 'date of birth': (OPDATE(numeric)-DOB(numeric))/365	< 17 years	0
PERIOPERATIVE MORTALITY			
In hospital mortality (INHOSRIP)	All cause 'in hospital mortality'(survivor=0, non-survivor=1)	0,1 only	0
Thirty day mortality(30DMort)	All cause 30 day mortality (survivor=0, non-survivor=1)	0,1 only	0
TREATMENT			
Neoadjuvant therapy (NEO-ADJUVANT THERAPY)	Free text;		NA
Surgical procedure: New variable (Operation_Classfn1)	Recoded from free text	All entries valid	NA
GRADE SURGEON	Free text		NA
GRADE ANAES	Free text		NA
TUMOUR CHARACTERISTICS			
Tumour histology (NewHist1Cln)	Tumour histology from biopsy; recoded from free text to category		NA
Final T classification (OVERALL T)	Final T classification; category		NA
Final N classification (OVERALL N)	Final N classification; category		NA
Final (OVERALL M)	Final M classification; category		22 unclear entries
NUTRITION STATUS VARIABLES			
Weight loss (Kg) at presentation estimated by patient		None specified	NA
Blood white cell count x 109/litre (WCC)		2 to 20	7 plausible values > 20; none excluded
Serum albumen gm/litre (ALB)		25-70 gm/litre	25
Weight (kg)		40-154	15(Excluded values Ian140, 1552, not plausible, 12.7 & 11.4; probably 'imperial')
Height (m)		NA	NA

Variable Name(original database label)	Data procedures & comments	Out of range Rule (acceptable range given)	Number of out of range values (record identifier)
Body surface area(BSA)	Calculated field:(Mosteller) sqrt[Ht(cm)*Wt(kg)/3600]	NA	NA
Body mass index(BMI)	Calculated field: wt(kg)/[ht(m)]2	NA	NA
COMORBIDITY FIELDs			
COMORBID	Categories of comorbidity including ‘cardiac’ & ‘respiratory’	NA	NA
OTHER	Free text description of comorbidity	NA	NA
DETAILS	Free text qualifying details of 'OTHER'	NA	NA
ASA	American Society of Anaesthesiologists physical status classification	Grade 1 to Grade 5 allowed	5 'E' for emergency, 2 zeros, 175 not specified
ALCOHOL		NA	
ALCOHOL TEXT	Free text	NA	
CARDIAC MORBIDITY			
CARDIAC	Category: cardiac diagnoses	NA	NA
ECG	Category: normal, abnormal, not done	NA	NA
ECGDETAILS	Free text ECG abnormality	NA	NA
REVISED CARDIAC RISK INDEX (Lee *et al.*, 1999; Poldermans *et al.*, 2009) and variations	New calculated variables		
TotalRCRI	Total score for the ‘Revised Cardiac Risk Index’ calculated from 'cardiac', 'comorbid', 'other' & 'details', 'ECG’ variables	NA	NA
RESPIRATORY MORBIDITY			
Pulmonary disease(PULM)	Free text respiratory diagnosis	NA	NA
Lung disease category (4 new categories below)	New derived field classifying respiratory disease; categories listed below		
Chronic obstructive pulmonary disease (COPDNewCode)	New categorical code derived from 'PULM' or 'OTHER' & 'DETAILS' fields	NA	NA
Other chronic pulmonary disease (ChronicNewCode)	New categorical code derived from 'PULM' or 'OTHER' & 'DETAILS' fields	NA	NA
Asthma (AsthmaNewCode)	New categorical code derived from 'PULM' or 'OTHER' & 'DETAILS' fields	NA	NA

3. Dataset from NOGCU (Table: Description of variables from NOGCU database)

Variable Name(original database label)	Data procedures & comments	Out of range Rule (acceptable range given)	Number of out of range values (record identifier)
Recent acute respiratory disease (AcuteNewCode)	New categorical code derived from 'PULM' or 'OTHER' & 'DETAILS' fields	NA	NA
SMOKING HISTORY			
Smoker or non-smoker (SmokerYesOne)	Yes/no (1/0)	NA	NA
Smoking category (SMOKERCODE)	Never/ex- for more than 12 months/current (code 3,2,1)	NA	NA
Arterial PO2 (kPa)		9-20 kPa	Excluded OOR 997, 166, 472, 695, 2684, 1543, 412 (probably transposed), 1736 (probably different units)
Arterial PCO2(kPa)		3-7 kPa	Excluded 133, 166, 472, 2844, 695, 2625, Ian 127, 1165, 2684, 1543, 1494, 1104 (probably transposed), 1956 (probably unit error)
FEV1	Forced expiratory volume in 1 second(litres)	<6 lit/sec	Ian285 excluded(out of range)
FVC	Forced vital capacity (litres)	1 to 7 litres	
FEV1/FVC	Calculated ratio FEV1/FVC (%)	NA	
CXR	Chest xray report: Normal, not done or free text description of abnormality	NA	NA
EXERCISE CAPACITY TEST			
Exercise capacity test (EXTOL)	Categorical	Not done, completed satisfactorily, not completed satisfactorily	NA
Pulse rate before test (PRPRE)	Note: acceptable ranges for exercise testing set according to observed range of reported values	40 to 150/min	6
Pulse rate after test (PRPOST)		40 to 170/min	5
Respiratory rate before test (RRPRE)		5 to 30/min	8
Respiratory rate after test(RRPOST)		5 to 60/min	8
Pulse oximetry before test (OXY SATS PRE)		85 to 100 (%)	2
Pulse oximetry after (OXY SATS POST)		85 to 100 (%)	5

3. Dataset from NOGCU (Table: Description of variables from NOGCU database)

Variable Name(original database label)	Data procedures & comments	Out of range Rule (acceptable range given)	Number of out of range values (record identifier)
Time to complete test (TIME TO COMPLETE)		0.5 to 10 (min)	233
Time to return heart rate to pre-test min(RETURN TO BASELINE)		0.5 to 10 (min)	133
OTHER VARIABLES			
Haemoglobin gm/dL (HB)		5 to 20 gm/dL	8
urea mmol/litre (UREA)		2 to 15 mmol/lit	28
creatinine µmol/litre (CREAT)		40 to 150 µmol/litre	26
glucose mmol/litre (GLUC)		3 to 20 mmol/lit	11

3.4.2 Missing data

Surgical procedure, age at operation, operation date, tumour stage and histology, gender, weight, RCRI, ‘in-hospital’ survival status, smoker status, respiratory comorbidity status, and several biochemistry and haematological results all had low percentages (less than 5%) of missing values. However, spirometry and arterial pO_2 had between 15 & 20% missing, weight loss 27%, and all exercise testing variables, ASA grade, grade of operating surgeon and anaesthetist, had between 15 and 50% missing values (Table 16 on following page).

Table 16 Frequency of missing values by variable for whole database

Missing value frequencies	N	Missing	
		Count	%
Operation date	1574	2	.1
Age at surgery	1574	2	.1
In hospital mortality	1572	4	.3
Weight Loss Kg	1145	431	27.3
White cell count	1550	26	1.6
Serum albumen	1527	49	3.1
Weight loss (% bodyweight at surgery)	1104	472	29.9
Weight Kg at surgery	1510	66	4.2
Height m	1365	211	13.4
P02	1328	248	15.7
Hb	1553	23	1.5
Serum K	1535	41	2.6
Urea	1549	27	1.7
Creatinine	1549	27	1.7
Glucose	1303	273	17.3
Pulse rate pre exercise	1326	250	15.9
Pulse rate post exercise	1324	252	16.0
Resp rate pre exercise	1270	306	19.4
Resp rate post exercise	1269	307	19.5
O2 saturation pre exercise	803	773	49.0
O2 saturation post exercise	801	775	49.2
Time to complete exercise test	758	818	51.9
Return to baseline after exercise	1260	316	20.1
FEV1	1457	119	7.6
FVC	1454	122	7.7
FEV/FVC ratio	1454	122	7.7
ASA grade	1313	263	16.7
Operating surgeon grade	966	610	38.7
Anaesthetic grade	897	679	43.1
Total RCRI	1575	1	.1
Gender	1575	1	.1
Histology	1576	0	.0
Comorbidity	1518	58	3.7
Respiratory comorbidity	1518	58	3.7
Smoker (yes/no)	1558	18	1.1
Smoker category(yes/stopped/never)	1548	28	1.8
Surgical procedure	1573	3	.2

3.4.3 Patterns of missing values

Using *SPSS Missing Value Analysis*, I examined missing patterns if more than 5% of the data was missing, if means of missing and non-missing data were statistically significantly different (p less than 0.05 for Student's t test using separate variances) and if data was potentially clinically important. Percentage of missing values for survivors and non-survivors were reported and data missing in groups were tabulated (Table 17 and Table 18 on following pages).

Table 17 Missing values by perioperative survivor status. Four (0.2%) outcomes are missing

Variable	Missing % in survivors	Missing % in non survivors
Operation date	0	1.1
Age at surgery	0	1.1
Weight Loss Kg	27.3	26.4
White cell count	1.5	2.3
Serum albumen	3.0	4.6
Weight Kg at surgery	4.0	6.9
Height m	12.8	23.0
P02 kPa	16.2	8.0
Hb	1.3	2.3
Serum K	2.6	1.1
Urea	1.7	1.1
Creatinine	1.7	1.1
Glucose	17.3	16.1
Pulse rate pre exercise	15.9	14.9
Pulse rate post exercise	16.0	14.9
Resp rate pre exercise	19.5	17.2
Resp rate post exercise	19.7	16.1
O2 saturation pre exercise	48.5	59.8
O2 saturation post exercise	48.6	59.8
Time to complete exercise test	51.6	58.6
Return to baseline after exercise	20.1	18.4
FEV1	7.4	9.2
FVC	7.5	10.3
FEV1/FVC ratio	7.5	10.3
Gender	.0	.0
Comorbidity		
Respiratory comorbidity	3.4	8.0
Smoker (yes/no)	.9	3.4
Smoker (yes/stopped /never)	1.5	3.4
ASA grade	16.5	19.5
Operating surgeon grade	38.0	50.6
Anaesthetic grade	42.4	54.0
Surgical procedure	.1	1.1

Table 18 Patterns of missing variables: commonest groups of missing data and variable means for each group

Missing patterns																	Means for each pattern of missing data					
Number of Cases	Comorbidity	Respiratory comorbidity	Pulse rate pre exercise	Pulse rate post exercise	Respiratory rate pre exercise	Respiratory rate post exercise	Return to baseline after exercise	Height m	ASA grade	Operating surgeon grade	Anaesthetic grade	O2 saturation pre exercise	O2 saturation post exercise	Time to complete exercise test	P02	Weight Loss	OPDATE	Weight Kg	Pulse post exercise	respiratory rate post exercise	Time to complete	Return to baseline
240																	14-Apr-2004	71.58	93.00	19.36	1.038	1.66
37															X		19-Jul-2006	65.58	96.59	20.38	1.239	1.22
19															X		06-Jan-2006	70.19	92.63	24.63	1.012	1.22
20																	03-Aug-2003	74.38	93.90	19.20	1.063	1.41
43								X		X	X	X	X	X			07-Aug-1996	66.97	104.86	25.37	.	2.95
153										X	X	X	X	X			01-Jun-1997	70.47	102.93	21.70	.	2.58
18												X	X	X			12-May-1998	77.78	109.44	19.94	.	9.34
76									X	X	X	X	X	X			29-Mar-1997	69.13	103.49	22.95	.	2.55
189																X	24-Mar-2005	78.24	96.12	20.11	.899	1.38
36															X	X	02-Jun-2007	76.34	87.75	19.00	1.404	1.40
25																X	24-Sep-2004	76.62	101.12	22.48	1.307	1.60
17	X	X							X	X	X	X	X	X			09-Aug-1998	76.12	99.71	19.29	.	1.78
31			X	X	X	X	X			X	X	X	X	X			29-Nov-1994	70.65	.	.	.	.
22			X	X	X	X	X		X	X	X	X	X	X			18-Jul-1996	69.55	.	.	.	.

Outcome

The definition of perioperative mortality for this study was any death occurring during the hospital admission associated with the primary surgical procedure or within 30 days of surgery. The binary ‘in hospital’ field (yes/no) was validated against the ‘date of death’ field and compared with the ’30 day mortality’ binary field to ensure that none of the latter were missed. Survivor status was validated against the ‘follow-up outpatient date’, to ensure that they had actually left hospital and been followed up. Three records of the total 1575 had missing data for mortality outcome and were excluded from the study. No formal data auditing procedures were used in this database but in 2010 mortality data was verified against NYCRIS data (Northern and Yorkshire Cancer Registry and Information Service).

Period of data collection

There were more missing weight loss and exercise testing values in earlier stages of data collection. It appears exercise testing was not started regularly until later in the 1990s and data collection may also have been less rigorous. Similarly, routinely collected data on grade of surgeon, anaesthetist and ASA score were scanty in earlier periods (Table 18).

Body Weight and weight loss

Body weight was associated with the frequency of missing estimated weight loss, and inversely with the frequency of missing exercise testing, pO_2 and height. Perhaps it is easier to miss weight loss in heavier individuals. However, it is not obvious why exercise testing values and height were more likely to be missing in lighter individuals. However, as missing frequencies for both were greater in earlier periods , and measured body weight also increased over time (68Kg for the first quartile of data collection against 75Kg for the fourth, confirmed in multiple regression, $p<0.001$), this may be explained by variations in data collection over time.

Exercise testing and other cardiopulmonary investigations

The exercise capacity test, which was used, incorporates various cardiorespiratory measures made before and after walking up flights of stairs. This has not been validated and is generally supervised by trainee medical staff. There was more missing data in earlier years and considerable amounts missing overall. Patients with missing pO2 and exercise testing results were associated with lower RCRI scores, suggesting perhaps that lower estimated cardiac risk was associated with less investigation (Table 18). Shorter time to completion of exercise testing was also associated with missing spirometry suggesting less comprehensive investigation of fitter patients. These findings would be expected in this mixed database if exercise was not deemed standard for all patients irrespective of estimated fitness.

It appeared that these investigations were done less frequently in certain categories of operation and in the small number of patients with non-malignant conditions, who had been entered onto the database. Again this might be expected if they did not follow the usual cancer staging. As expected, this data was missing more frequently in the cases which may have been urgent.

Frequencies of missing oxygen saturations before and after exercise and time to completion were associated with longer post-exercise recovery times and higher respiratory and pulse rates, and possibly with poorer spirometry results (FEV1 and FVC). It is possible that there may have been a problem completing the tests or obtaining observations such as oxygen saturation in less fit individuals following exercise.

Age at surgery

Age at surgery was calculated from: ‘operation date’ minus ‘date of birth’. This data was almost complete.

Operative procedure

Oesophagectomy is the focus of this study; in our unit it is almost exclusively the Ivor Lewis procedure which entails major intra-abdominal and intrathoracic surgery. Total gastrectomy is a major surgical procedure, but exclusively intra-abdominal, which has a frequently reported higher mortality than oesophagectomy (Cromwell *et al.*, 2010), and is also a frequent procedure in this unit. There were also a mix of other procedures, which generally incurred a lesser morbidity and mortality. Therefore I mapped free text descriptions of surgical procedures into these three main categories. The 'other' group (including partial gastrectomy, palliative procedures, laparotomies and a small number of emergencies), would be the reference group for the oesophagectomies, and total gastrectomies.

Tumour histology

I recoded this free text field into adenocarcinoma, squamous cell carcinoma and 'other' conditions (including benign).

Nutritional markers

Weight, height, weight loss (as reported by the patient in kilograms and as a percentage of body weight at surgery), white cell count and serum albumen were all subjected to routine data checking.

Cardiovascular morbidity

Preoperative cardiac morbidity was coded in a variety of ways in several free text and categorical fields in the original database. One field coded cardiac disease as present or absent ('COMORBIDITY'), one allowed a free text description ('CARDIAC') with a free text qualifying field ('DETAILS'), and one allowed free text comorbidity descriptions in a field containing any comorbidity ('OTHER'). 'ECG' recorded the pre-operative electrocardiograph as normal or abnormal, and 'ECGDETAILS' allowed a free text description of any abnormalities.

The free text fields were difficult to analyse because of the great range of free text entries. For instance there were 184 distinctly different entries in the ‘CARDIAC’ field, 96 in the ‘COMORBID’ field, 751 in ‘OTHER’, 700 in ‘DETAILS’ and 676 in ‘ECG’. Many of these were accounted for by variations in definition, description and spelling. Therefore I decided to try and map these to terms which would form the basis of the ‘Revised Cardiac Risk Index (RCRI)’, a validated cardiac risk score (Lee *et al.*, 1999; Fleisher *et al.*, 2007). Using Excel ‘PivotTable’, I summarised all possible entries in the cardiac fields and mapped these to terms used in the RCRI using a ‘lookup’ table. This enabled each case to be scored using an appropriate form of the RCRI.

Respiratory comorbidity

Respiratory comorbidity was also represented by several continuous, as well as categorical and free text field variables. Continuous variables included spirometry (FEV_1, FVC and their calculated ratio) and arterial blood gases (oxygen saturation measured by pulse oximetry, arterial oxygen and carbon dioxide partial pressures, and pH and bicarbonate concentration). Respiratory comorbidity was also represented in free text fields which described various comorbidities (‘COMORBID’, ‘OTHER’, ‘DETAILS’). The field ‘PULM’ contained 96 distinct free text descriptions of pulmonary diseases; three fields described tobacco use (‘yes’/’no’, current/past smoking habit and a free text description of smoking history). As in the cardiac data fields, there was a large amount of heterogeneous free text information. I mapped free text data terms which defined four main groups: chronic obstructive pulmonary disease (COPD), asthma, other chronic lung disease, or a history of other acute respiratory illness.

3.5 Discussion

The NOGCU has twenty years worth of data and should be a rich source of information. I have prepared a 'cleaned' dataset of fields from this clinical database, in preparation for exploring a clinical prediction model of perioperative mortality after oesophagectomy. However, there are several sources of potential bias, which can arise from the data in prediction modelling, and which were identified in the systematic review and have been summarised by Hayden (Hayden *et al.*, 2006). I discuss their significance in relation to this database under headings from Hayden's recommendations below (Hayden *et al.*, 2006).

3.5.1 Does the sample represent the population of interest?

Patients may be excluded from surgery for a variety of reasons including medical fitness, age, and variation in surgical indications between centres. This type of selection bias is well recognised (Delgado-Rodriguez and Llorca, 2004) and its potential was noted in some primary studies in the systematic review (Thomas *et al.*, 1996; Sabel *et al.*, 2002; Ruol *et al.*, 2007(a)). Its effect may be to bias results and make application to other groups less reliable. The NOCGU database was set up with the intention of including every case of oesophagogastric cancer referred to the unit, and there is no evidence to suggest that the data does not include a consecutive set of operated cases, and should be relatively free of 'loss to follow up' selection bias. I internally validated survivor status within the database against 'discharge from hospital' and 'outpatient follow up appointment' fields, and the NOGCU team validated survivor status with NYCRIS (Northern and Yorkshire Cancer Registry and Information Service). However, it is possible that patients may have been completely omitted from the database, or may have been excluded from surgical treatment on the basis of perceived 'unfitness' for surgery (e.g. comorbidity, old age); the database did not hold data to allow conclusions about this aspect of management until fairly recently.

3.5.2 Does the data represent the sample?

Data validity

There were several potential sources of information bias in our data. In the early stages much clinical data was entered by various grades of junior staff, and therefore subject to their interpretation and errors of which several types were reported in the results. This was compounded by the use of many free text fields and lack of consistent definition of certain variables e.g. cardiovascular comorbidity. Further transcription from data entry form to computer and in this study, my recoding of some fields only adds to this risk. As the database has developed, many of these issues have been resolved with senior medical staff completing data entry forms, and the use of standardised and categorical data.

Data audit is a possible solution to maintaining data integrity. This was not formally used in the NOGCU database until the more recent use of NYCRIS to validate mortality (Northern and Yorkshire Cancer Registry and Information Service) and was infrequently reported in the systematic review of primary studies (Adam *et al.*, 1996; Bailey *et al.*, 2003; McCulloch *et al.*, 2003; Rahamim *et al.*, 2003; Rentz *et al.*, 2003; Tekkis *et al.*, 2004; Moskovitz *et al.*, 2006).

Clinical database studies can also be prone to 'observer expectation bias', when data is entered by an investigator aware of the study aim or hypothesis (Delgado-Rodriguez and Llorca, 2004). This can be partially alleviated by 'blinding' data entry staff to the outcome, but this is not a practical solution in most cases. Entering prognostic data prospectively, before the outcome of interest (survival or non-survival) has occurred accounts for this problem and data collected for the NOGCU database mainly fulfils this criterion.

Missing data

Missing data can reduce study efficiency by reducing effective sample size, and can bias the effects of predictors (Steyerberg, 2009f). The latter is more likely if missingness of data (outcome or predictor) depends on outcome, e.g. non-survivors have more missing data than survivors. It is also possible if the missing and non-missing predictor data have different outcome rates (Steyerberg, 2009f). Outcome (mortality) data was virtually complete and had been validated so was unlikely to be 'missing not at random'. However, mortality rate was higher in patients with missing values on height, arterial pO_2, and pre and post exercise test oxygen saturations. Higher mortality in association with missing pO_2 could have been explained by the inclusion of emergency cases, which may not have had this measured. There was also a suggestion that patients with poorer cardiorespiratory reserve, for example those with poorer spirometry and higher pulse and respiratory rates after exercise testing, had more missing pre and post exercise oxygen saturations. This could perhaps be explained by an inability to obtain measurements in patients possibly struggling with exercise. Mortality was also higher in patients with missing oxygen saturations during exercise testing, perhaps a reflection that patients unable to manage exercise may have been at higher risk. Patients treated earlier in the data collection period were also less comprehensively investigated, particularly in respect of cardiorespiratory and exercise capacity. This would be expected as the unit was developing but makes application to future groups of patients difficult. These explanations are all speculation, but together with amount of missing data for some exercise testing measurements, decreases confidence for including them as candidate predictors. These patterns suggest that for some predictors missing data is not missing completely at random, which is the usual safe assumption for missing data. This was suggested as Little's MCAR test applied to all potential variables, provided evidence against the null hypothesis that data was MCAR (Chi-Square 5119.695, df 3889, $p<0.001$).

3.5.3 Strengths and weaknesses of this dataset

1) The NOGCU clinical database is a moderately large database with a set of records from a reasonably homogenous case-mix and set of surgical

procedures, which has been managed by a professional database manager for a long period.

2) The outcome mortality data and some simpler data such as gender, operation, operation date, and tumour histology were nearly complete and had been subjected to some degree of validation.
3) For the purposes of this study it is a 'convenience' sample (Harrell, 2001b), which was not prospectively set up for this purpose. Some of the fields of interest lacked definitions, much data was free text, and there were several steps where potential bias in data entry could have occurred. Some of the fields required a considerable amount of recoding into useable predictors, which added to the risk of information bias (Delgado-Rodriguez and Llorca, 2004).
4) Some fields (e.g. weight loss, exercise testing measurements) had considerable amounts of missing data.

3.6 Key findings

1) I have prepared a set of fields containing candidate predictors from 1575 cases from the NOGCU database, with which to explore a clinical prediction model of perioperative mortality after oesophagectomy.
2) Perioperative mortality outcome, age, gender, operation and tumour histology were nearly complete and reliable. There is no reliable information to study patients who may have been excluded from surgery because of medical unfitness.
3) Other predictors of potential interest, for instance comorbidities, were entered into the database in extensive free text, without prior definition and requiring considerable recoding. These must be open to potential bias.
4) Some fields had considerable amounts of missing data including weight loss, measures recorded during exercise testing, grade of surgeon and anaesthetist and ASA score. There was more missing data from the earlier years of data collection and as expected possibly in fitter patients and those undergoing emergency surgery, who may not have undergone the test.

Chapter 4: Developing a clinical prediction model of perioperative mortality after oesophagectomy from the NOGCU clinical database

4.1 Introduction

In this section of the thesis I aim to develop a clinical prediction model of perioperative mortality using a 'cleaned' subset of data from the NOGCU. General considerations about the modelling methods are summarised in this introduction and have been drawn predominantly from Steyerberg's 'Clinical Prediction Models' (Steyerberg, 2009e) and Harrell's 'Regression Modelling Strategies' (Harrell, 2001e).

4.1.1 Background to modelling methods

Perioperative mortality has a binary outcome, for which there are several modelling methods. Logistic regression is a flexible and widely used method allowing the incorporation of continuous, categorical and non-linear predictors and the interactions between them (Kleinbaum, 1994; Steyerberg, 2009l). The outcome is modelled as the natural logarithm of the odds against a linear function of the predictors. Individual predictor and model performance can be compared with formal statistical tests.

Other methods include discriminant analysis, Bayesian methods, classification and regression trees and neural networks (Steyerberg, 2009l). Neural networks are probably the most familiar of these and are well suited to identifying non-linear effects, interactions and unspecified effects. However, they identify relationships which are wholly data driven (the 'black box' analogy), may be less acceptable to medical practitioners, and model performances are more difficult to compare. In contrast, logistic regression requires pre-specification of data relationships and requires some knowledge of how the predictors are linked to the outcome, perhaps increasing its acceptability to clinicians. Although extensively used in medical applications, neural networks have not been shown to perform better than logistic regression in classification problems (Tu, 1996; Sargent, 2001; Dreiseitl and Ohno-Machado, 2002). Consequently, I have chosen to use logistic regression to model perioperative mortality.

4.1.2 Selection of candidate predictors

Age at surgery

There was strong evidence from the systematic review that age was associated with perioperative outcome, and had strong supporting rationale. There was also some evidence that its effect may be more marked in extreme old age and therefore it may be worth considering a transformed age predictor to account for this. A simple example is the 'squared' transform of age, which I will consider.

Surgical operation

Surgical procedure was included as a 3 category predictor because the study focuses on oesophagectomy, and the database contains a mixed surgical caseload. The other category was total gastrectomy, which has been reported to have a higher mortality (Cromwell *et al.*, 2010), and the reference category was "other", which included operations known to have a lower mortality (e.g. subtotal gastrectomy).

Cardiovascular comorbidity

The evidence from the systematic review for including cardiovascular comorbidity as a predictor was weak, however there is a physiological rationale for its inclusion, and the Revised Cardiac Risk Index (RCRI) (Lee *et al.*, 1999) has been validated in large samples of other major surgical procedures. This allocates one point for each of high risk surgery, ischaemic heart disease, history of congestive heart failure, history of cerebrovascular disease, IDDM, creatinine above 2.0 mg per dl. The points sum categorises patients into risk classes 1 to 4 depending on number of risk factors present. It has also been used as a two level score (Ford *et al.*, 2010); therefore I explored the total RCRI score, and the two and four level scores.

Respiratory comorbidity

About half the studies in the systematic review found associations between various markers of respiratory disease and perioperative mortality. In oesophagectomy the lung is subjected to a variety of physiological insults and pulmonary complications are common and associated with mortality, therefore it seems reasonable to include some marker of respiratory comorbidity. The NOGCU database contains multiple measures and descriptors of respiratory comorbidity but there is no clear consensus on the most useful.

With the large number of fields representing pulmonary disease in the database some degree of data reduction was desirable. It is unlikely that patients with acute illness will undergo surgery without appropriate treatment, or that 'burnt out' chronic disease would impair outcome, therefore I focussed on identifying patients with COPD, because it is common, progressive and subject to exacerbations. In this dataset there was no relationship between categorical or free text descriptions of respiratory disease and spirometry and therefore did not allow simple combination of these variables. The database did not use a clear definition of COPD, but spirometry is central to its diagnosis (National Clinical Guideline Centre, 2010), therefore I selected this predictor. There is debate about the best diagnostic spirometric measure for COPD, for instance, whether spirometry should be recorded before or after inhaled bronchodilators (National Clinical Guideline Centre, 2010). NICE have recently recommended using post-bronchodilator FEV1, but most studies use FEV1 without bronchodilator, and as this was the reported measure in our database, I selected this as our predictor (National Clinical Guideline Centre, 2010).

Nutritional status

Protein-calorie malnutrition is associated with poorer outcomes after major surgery (Law *et al.*, 1973; Fekete and Belghiti, 1988; Windsor and Hill, 1988). About 30% of studies in the systematic review reported associations between mortality and measures of loss of body mass and serological markers of malnutrition or immunosuppression. The considerable heterogeneity of marker definitions made appropriate selection of predictors difficult but three studies identified serum albumen as important (Saito *et al.*, 1993; Rentz *et al.*, 2003; Atkins *et al.*, 2004). I selected this as the main nutritional marker although weight loss and white cell count were also explored.

Other candidate predictors

Various estimates of exercise (Law *et al.*, 1994) or activity capacity (Ferguson *et al.*, 1997; Bartels *et al.*, 1998; Bailey *et al.*, 2003) were reported to be associated with mortality in the systematic review. However, in the NOGCU database general activity was not routinely recorded and the exercise test had much missing data, was not consistently standardised and not likely to be accepted as a standard test. Other predictors such as renal disease, diabetes, and liver disease are also potentially important, but their prevalence was very low. These potential predictors will be explored but it is unlikely that I will consider them for the prediction models for the reasons given above.

4.1.3 Handling missing data

Missing values pose a particular problem in modelling studies because they may reduce study efficiency by loss of information, and may bias regression coefficient estimates, because data may be missing “systematically” rather than randomly (Little, 1992). The mechanism of “missingness” is central to the effect on the study and how it may be managed. Values which are missing due to random factors outwith the study (e.g. administrative error) are “missing completely at random”(MCAR) and do not bias the study. Observations may be missing at random after controlling for values of other variables e.g. more missing exercise data in earlier study periods. This is “missing at random” (MAR), and is particularly problematic when “missingness” depends on the outcome variable (in this case mortality), resulting in biased regression coefficients (Steyerberg, 2009f). Missing values may depend on the values which are missing (for instance non-survivors missing more than survivors) or other unobserved variables and are “missing not at random” (MNAR). The mechanisms, effects and handling of missing values have been described in various sources (Carpenter *et al.*; Little, 1992; Vach, 1997; Steyerberg, 2009f).

Complete case analysis is the common approach to this problem and excludes any case with missing values in the outcome or predictors and is therefore relatively inefficient. The resulting reduction in sample size results in a reduced event to variable ratio leading to overfitting and chance associations (Harrell, 2001c; Steyerberg, 2009n). Comparisons between models can also be difficult to interpret as differences between univariate and adjusted coefficients may be due to varying patterns of missing data, rather than correlation between predictors.

A potential solution is to replace missing values by multiple imputations and thereby maximise study efficiency (Steyerberg, 2009f). However, this may lack face validity for some clinicians, and I have chosen complete case analysis initially and will only consider statistical data replacement later depending on initial patterns of missing data.

Replacement of missing values by multiple imputations is based on the idea that the original observations are a random sample from the overall population and therefore, the same conclusions should be reached if they were replaced by other random observations from that population. The missing values can be replaced under the assumption that they are ‘missing at random’. That is ‘missingness’ is random after controlling for other variables (Carpenter *et al.*; Howell, 3/7/2009). For example, perhaps albumen may have more missing values in younger patients. Under ‘missing at random’ we would assume that albumen is missing at random after controlling for age. The statistics program (SPSS) would replace the missing values with random values from a distribution based on the non-missing values from the predictor to be replaced and other auxiliary predictors in the dataset.

4.1.4 Model validation

Clinical prediction models should be capable of predicting outcome accurately (calibration) and allocating the correct outcome to patients at high and low risk (discrimination) (Altman and Royston, 2000). Models, which have performed well in development, often fail to deliver satisfactory performance when applied to new patients (Justice *et al.*, 1999). Validation phases include examining performance of the model on the sample on which it was developed (apparent validation), on a separate portion of the sample under study (internal validation) and on a new and unseen sample of relevant subjects, who have not been used in model building (external validation) (Steyerberg, 2009o). This study is on existing data from a regional clinical database, which will be used to develop and internally validate a prediction model. External validation on "unseen" data is beyond the scope of this study.

Traditional methods of internal validation include split sampling (one for modelling and one for development) but newer methods such as bootstrapping (Harrell, 2001e), are increasingly used to utilise the entire data sample. Split sample validation inevitably leads to reduced size of modelling and validation samples. This can lead to random imbalances in outcome and predictors, and to unreliable assessment of model performance. However, I selected split sampling for internal validation because it is still widely used and accepted in clinical studies, and has face validity. It also gives some scope to explore data and validate findings in the validation sample. I also decided to use random samples, which were balanced for mortality outcome.

4.1.5 Statistical measures to compare and validate models

The statistical measures used in this study have been described in several sources (Justice *et al.*, 1999; Altman and Royston, 2000; Steyerberg *et al.*, 2010) and I have summarised the main categories below.

4.1.6 The amount of information in a model

This is the amount of variation explained by a model and gives some idea of how well it will predict outcome compared to a model with just the mean sample outcome. The maximised likelihood value, L, (the probability of obtaining the observed data given the stated model and parameters) is the basis of the likelihood ratio statistic (-2LL), which has a chi-square distribution and is used to compare the predictive ability of two models (Kleinbaum, 1994; Steyerberg *et al.*, 2010). The Wald statistic has a standardised normal distribution and tests whether individual regression coefficients differ from zero.

Nagelkerke's R^2 is generally reported in logistic regression output. However, this is not directly comparable to the R^2 of ordinary linear regression, which is assessing how well the model minimises the difference between predicted values and actual values. Nagelkerke's R^2 is based on the ratio of likelihoods of the model with and without predictors. Although it is scaled between 0 and 1, and independent of sample size the values can only reliably be used to compare models on the same datasets. It is also usually small in logistic regression and is not a reliable measure of goodness of fit (Steyerberg *et al.*, 2010; Statistical Consulting Group, October, 2011).

4.1.7 Calibration and goodness of fit

This reflects how accurately a model's prediction of "x%" mortality is observed in the sample of interest. Because individual outcomes can only be 0 or 1, it is only possible to compare mortality rates in groups, and to compare predicted and observed means. These values can be demonstrated in plots (typically in 10 groups of ascending predicted values) and compared using the Hosmer-Lemeshow statistic (Lemeshow and Hosmer, 1982).

4.1.8 Discrimination

The ability to correctly allocate outcome is commonly quantified by a receiver operator curve (ROC), and in this study represents the probability that a randomly selected patient, who died, had a higher predicted risk than a randomly selected, one who survived. The value of interest is the area under the ROC curve and is the same as the c statistic for binary outcomes (Hanley and McNeil, 1982).Another estimate of goodness of fit includes the Brier statistic, which uses a scaled score for a quadratic function of the predictions errors (Steyerberg *et al.*, 2010).

4.1.9 Classification

The potential impact of a prediction model can be gauged by how many cases it can correctly classify into high or low risk groups. Models can be compared by assessing how many patients could have benefitted from the use of the models. I will examine what impact could have been made on treatment decisions had selected levels of predicted mortality been acted on.

4.1.10 Aim and goals

I aim to develop, internally validate and assess the performance of a clinical prediction model of perioperative mortality after oesophagectomy using data from the NOGCU database.

Goals

1. To select a set of candidate predictors for inclusion in a clinical prediction model of perioperative mortality. I will use current clinical knowledge, the systematic review in Chapter 2, and secondarily the results of univariate analysis and stepwise regression methods to inform the choice of predictors.
2. To develop a clinical prediction model using complete case logistic regression on a random sample from the NOGCU database. I will consider the use of imputation methods to optimise study efficiency if appropriate.
3. To assess the performance of the prediction models on a random sample of data from the NOGCU database.

4.2 Methods

4.2.1 Selection of candidate predictors

Age at surgery

Age and 'age squared' were examined as candidate predictors. Age was coded as 'age minus 30' in decades to give a clinically useful zero reference (30 years) and regression coefficients that were large enough to manage in SPSS.

Surgical operation

Surgical procedure was included as a 3 category predictor:

a. Thoracoabdominal oesophagectomy

b. Total gastrectomy

c. 'Other' operations (laparotomy, subtotal gastrectomy etc) was the reference group.

Cardiovascular comorbidity

The Revised Cardiac Risk Index (RCRI) (Lee *et al.*, 1999) was examined as a candidate predictor both as a total score, and two and four level categorical scores.

Respiratory comorbidity

Respiratory comorbidity was represented by the spirometric measure of forced expiratory volume in 1 second, as a percentage of that predicted (FEV1) for age gender and height (National Clinical Guideline Centre, 2010). Predicted values for FEV1 were calculated from the following equations from the European Coal and Steel Community (Quanjer *et al.*, 1993).

Males: $(0.043\ \times height\ cm) - (0.029 \times age\ years) - 2.49$

Females: $(0.0395\ \times height\ cm) - (0.025 \times age\ years) - 2.6$

Nutritional status

Nutritional candidate predictors for examination were serum albumen, white cell count and estimated weight loss at surgery.

Other candidate predictors

I explored various predictors from exercise capacity tests but did not consider them for inclusion in the prediction model

4.2.2 Data exploration

Summary statistics, distributions, and missing values were reported for candidate predictors in the modelling sample, which contained oesophagectomies, gastrectomies and a group of other operations. Distributions of preselected predictors and their univariate associations with mortality were explored with logistic regression. Mortality rates and confidence intervals for selected quantiles were plotted using statistics packages based on R (Appendix G iii.).

4.2.3 General Modelling strategy

1. The full dataset (n=1575) was split into two approximately equal random samples with similar mortality rates, one for modelling and one for validation.

2. I used complete case analysis and multiply imputed datasets to develop the prediction model.

3. I pre-specified a ‘Clinical’ model to reduce selection bias and overfitting inherent in data driven methods, which could lead to poor performance in a new sample (Steyerberg, 2009k). Predictors included age, revised cardiac risk index (RCRI), spirometry (FEV1 % predicted), surgical procedure, and serum albumen. I explored the effect of adding or removing certain predictors from the main models.

4. I used the univariate associations and ‘stepwise’ elimination methods to explore a range of other candidate predictors, which could be important and

considered for inclusion in the ‘Clinical’ model. I also used stepwise elimination to generate a ‘Statistical’ model for comparison.

5. I explored the modelling assumptions of linearity of response on predictors, additivity of predictor effects and data fit to model (residuals and effects of any extreme values).

6. Selected models were tested on the random validation sample (50% of the sample). The performance of models was also compared with the Steyerberg ‘Rotterdam’ model (Steyerberg *et al.*, 2006). This model developed a risk score from logistic regression, which I used on our dataset. This score allocated a score (from -2 to 1.5) depending on age category, the presence of pulmonary, cardiovascular, liver or renal disease and diabetes. Points were also allocated for hospital surgical volume, and chemo- or radiotherapy. Using this model required some recoding in the NOGCU database, and therefore direct comparisons with the models derived from the NOGCU should be made with caution.

7. Properties used to examine model performance included:

- The data variance accounted for by the models was compared using the chi-square statistic for -2LL (minus double the log likelihood) and Nagelkerke’s R ^{2}. The latter is a logarithmic score of difference between predicted and observed outcome, scaled to between 0 and 1 (Steyerberg *et al.*, 2010).
- Discrimination between survivors and non-survivors using plots of Receiver Operator Curves (Hanley and McNeil, 1982) and the areas under the curves.
- The fit of predicted to observed values was assessed using the Hosmer-Lemeshow (Lemeshow and Hosmer, 1982) statistic and calibration plots of average predicted risk against average observed mortality for each ascending decile of predictions. There is debate around the utility of the Hosmer-Lemeshow test, particularly the optimal data groupings and its power to detect poor calibration and overfitting (Steyerberg, 2009h).

8. Deviations from the initial protocol, which arose during the analysis, are described in the results as they occurred.

4.2.4 Generation of multiple imputation datasets

Multiple imputation was carried out according to methods and recommendations reported by van Buuren (van Buuren *et al.*, 1999). Patterns of missing data were examined and the multiple imputation carried out in SPSS using the following classes of predictor from the original data were used to estimate missing values (van Buuren *et al.*, 1999; Clark and Altman, 2003):

i. Predictors, which were associated with 'missingness' of predictors to be replaced. This was determined using logistic regression (dependant variable 'missing' or 'present') against a range of predictors.
ii. Predictors which may be correlated with the predictor to be replaced.
iii. The outcome variable and all the predictors in the original full 'clinical' model were included.

The 'Fully Conditional method' was used with iterations set to a maximum of 10. Constraints on imputed values (FEV1 and serum albumen) were set to their original sample ranges. Predictors used to impute values included all predictors and outcome in the prespecified model, predictors whose values were associated with missingness in the target variables (operation date and pO2 for FEV1), and variables which were correlated with the target predictors (weight, height, gender).

4.2.5 Logistic regression

1. Logistic regression (`Analyze, Binary regression`) was carried out using SPSS 17.0. (Release 17.0.0. 23 August 2008). Predictors were 'forced' into the model for pre-specified models and also for 'data driven' stepwise elimination methods, after selection of predictors. Predicted probabilities, standardised residuals, Cook's, leverages, and DfBeta and Hosmer-Lemeshow statistics were saved for later examination. Correlation between predictors was examined and multicollinearity checked by running the

model using SPSS ordinary multiple linear regression as suggested by Field (SPSS:`Analyse, Regression, Linear`)(Field, 2000).

2. For the prespecified 'Clinical' model, age, FEV1, serum albumen and RCRI were entered as continuous variables and surgical procedure as a categorical variable. For categorical variables the 'indicator' contrast (SPSS terminology) was used and the reference group was generally the lowest risk category (e.g. 'other' operation).

2. A 'Statistical' model was generated for comparison and to explore other potential candidate predictors. Stepwise regression is prone to overestimate coefficients and underestimate p values and confidence intervals, and therefore an initial global test of no regression was carried out with all candidate predictors in the model (Harrell, 2001d). Candidate predictors included all from the prespecified 'Clinical' model plus smoker status, presence of respiratory disease, preoperative arterial pO_2, percentage of reported weight loss, white cell count, gender, tumour histology and stage, and operation date. Backward elimination was used to exclude apparently unimportant predictors ($p<0.1$), and those selected were forced into a model to reduce the possible effects of missing data resulting from the initial inclusion of all candidate predictors.

4.2.6 Validation

1. The regression coefficients from selected models were 'back substituted' into the logistic function using SPSS, to calculate mortality probabilities for individuals in the validation dataset.

2. Calibration of individual models was examined using the val.prob.ci function, which is a modification by Vergouwe (Vergouwe and Steyerberg, 2009) of the 'val.prob' (Harrell, 2012) function from the Regression Modelling Strategies ('rms') software package (Harrell Jr, 24/03/2011). The function 'val.prob.ci' adds confidence intervals to the observed outcomes, and I inactivated the histogram output of predicted probabilities in favour of graphics rendered in ggplot2. I communicated by email with Professor E.

Steyerberg for help in using the function and interpreting the output (Warnell, 2012(unpublished communication)).

3. Discrimination was assessed by plotting and calculating the area under the receiver operator curves for each model. This was done using the ‘plotROC’ function (Appendix G vi.) from the R based PredictABEL package (Kundu *et al.*, 2011).

4. True and false positive rates were generated in SPSS and reported for various cut-offs of predicted mortalities for selected models in order to give an idea of the potential utility of models as classifiers of risk in practice.

4.3 Results

4.3.1 Distributions of candidate predictors and their association with mortality

The distributions of candidate predictors and their associations with mortality are reported in Table 19 and Table 20 and follow on the next pages. The unadjusted odds ratios from the logistic regression is the multiplier of the odds of the outcome, which results from a one unit change in the predictor. For instance, in this sample the effect of a one year increase in age is to multiply the odds of perioperative mortality (defined by me as ‘in hospital’ mortality), by 1.047(95% CI 1.011, 1.084). Age was also studied by decade with a reference of 30 years, to give a more practical interpretation, and the odds ratio per decade above 30 was 1.047^{10} , which is 1.583(95% CI 1.119, 2.441). The intercept from the logistic equation gives the baseline odds for the outcome, given no predictors in the equation. For this sample the odds ratio for the overall ‘in hospital’ mortality was 0.058, or a mortality rate of 5.5%.

Other predictors which were statistically significantly (p<0.1) associated with perioperative ‘in hospital’ mortality were, weight loss (OR 1.005, p=0.0.075), white cell count (OR 1.006, p=0.075), RCRI as a continuous variable (OR 1.251, p=0.049), RCRI as a four level categorical variable (OR 2.059, p=0.084), male gender (OR 2.238, p=0.05) and thoracic oesophagectomy, referenced to ‘other’ procedures (OR 2.190, p=0.074).

Table 19 Distribution and mortality rates for outcome and continuous predictors

Variable	**cases (missing)**	**Out of range values**	**Sample mean(sd)** ***[median (min,max)]***	**Survivor mean(sd)** ***[median(min,max)]***	**Non-survivor mean(sd);median (min,max)**	**Unadjusted odds ratio (OR), 95% CI**
In hospital mortality	787(0)	nil	Survivors 743, non-survivors 44 (5.6%)			0.058, p=0.00
Age at surgery	786(1)	nil	65.4(10.2) *[67(30, 90)]*	65.21(10.174) *[67(30,90)]*	69.37(9.757) *[70(31,86)]*	1.047[1.011,1.084],p=0.01‡
Age decade	786(1)	nil	3.544(1.019) *[3.7(0,6)]*	3.521(1.0174) *[3.7 (0,6)]*	3.937(0.9756) *[4(0.1, 5.6)]*	1.583(1.119,2.441), p=0.01
RESP						
FEV1lit/sec	722(65)	1 (95 lit/sec)	2.5(0.8077) [2.5(0.6, 5.4)]	2.34(0.7235) [2.325(0.7, 4)]	2.51(0.8118) [2.5(0.6, 5.4)]	0.760[0.505, 1.145], p=0.189
FEV1 (% predicted)	636(151)	nil	0.871(0.255) *[0.8686(0.24, 2.02)]*	0.874(0.255) *[0.875 (0.26, 2.02)]*	0.819(0.252) *[0.827(0.24, 1.31)]*	0.415[0.096, 1.79], p=0.238
FVC lit	721(66)	1(as FEV1)	3.36 (0.991) [3.37(0.2, 6.9)]	3.374 (0.991) [3.395(0.2, 6.9)]	3.164 (0.974) [3.25(1, 5)]	0.804[0.577, 1.121], p=0.199
pO2 kPa	647(140)	8 (<5,>30; ?data entry & unit error)	12.28 (2.22) *[12(5.1, 25.8)]*	12.27 (2.157) *[12(5.2, 25.5)]*	12.46 (3.066) *[12.2(5.1, 25.8)]*	1.038[0.906, 1.188], p=0.591
NUTRITION						
Height m	679(108)	nil	1.692 (0.093) *[1.7(1.43, 1.95)]*	1.69(0.0926) *[1.7(1.47, 1.95)]*	1.695(0.102) *[1.7(1.43, 1.91)]*	1.441[0.039, 53.53] p=0.843
Weight Kg	749(38)	2 (11.7,12.4)	71.90(15.44) *[71(35.8, 140)]*	71.31(14.79) *[70(47, 108)]*	71.93(15.49) *[71(35.8, 140)]*	0.997[0.997, 1.018], p=0.801
Body surface area	670(117)	nil	1.836(0.2286) *[1.823(1.221, 2.668)]*	1.837 (0.2286) *[1.823(1.22, 2.67)]*	1.827(0.2319) *[1.807(1.42, 2.37)]*	0.839[0.192, 3.670], p=0.816
Body mass index	670(117)	nil	25.203(4.615) *[24.957(13.89, 46.20)]*	25.224(4.634) *[24.953(13.89, 46.20)]*	24.832(4.315)*[24.959(16.26, 32.86)}*	0.981[0.911, 1.057], p=0.620
Weight loss	565(222)	nil	5.96(5.63) *[5(0, 32)]*	5.86(5.621) *[5(0, 32)}*	7.79(5.734) *[8(0, 20)]*	1.005[0.995, 1.119], p=0.0.075‡
White cell count	778(9)	nil	8.29(4.37) *[7.7 (2.9, 74)]*	8.291(4.448) *[7.65 (2.9, 74)]*	8.414(2.904) *[7.95 (4.1, 19.3)}*	1.006[0.944, 1.071], p=0.075‡
Serum albumen	755(32)	nil	41.54(5.454) *[42(15, 93)*]	41.60(5.508) *[42(15, 93)]*	40.38(4.333) *[42(28, 48)]*	0.96[0.908, 1.015], p=0.148

4. Clinical prediction model of perioperative mortality (Table: continuous predictors)

Variable	**cases (missing)**	**Out of range values**	**Sample mean(sd) *[median (min,max)]***	**Survivor mean(sd) *[median(min,max)]***	**Non-survivor mean(sd);median (min,max)**	**Unadjusted odds ratio (OR), 95% CI**
CARDIAC						
RCRI	787(0)	nil	1.41(1.262) *[1(0,6)]*	1.38(1.247) *[1(0,6)]*	1.77(1.461) *[2(0,6)]*	1.251[1.001, 1.665], p=0.049
OTHER PREDICTORS						
Hb gm/dL	774(13)	5(0, 117, 906)	13.4(2.141) *[13.7(5, 19)]*	13.43(2.121) *[13.7(5.9, 19)]*	12.99(2.432) *[13.55(5, 16.7)]*	0.915[0.800, 1.046],p=0.194
Urea mmol/l	776(11)	nil	5.33(3.7533) *[4.9(1, 73)]*	5.33(3.832) *[4.9(1, 73)]*	5.35(2.049) *[5.15(2.3, 14.5)]*	1.001[0.924, 1.084],p=0.984
creatinine µmol/l	775(12)	6 (6,21,8,11, 8,38) not excluded	90.86(20.404) *[89(6, 313)]*	90.65(20.613*)[89(6, 313)]*	94.25(16.380) *[93(66, 137)]*	1.007[0.995, 1.020],p=0.255
Glucose mmol/l	651(136)	2 (115, 97)	6.22(1.869) *[5.7(2, 18)]*	6.2(1.858) *[5.7(2, 18)]*	6.52(2.066) *[5.7(4, 12)]*	1.081[0.926, 1.262],p=0.324

Table 20 Distribution and mortality for categorical predictors

Categorical predictors	Missing n (%)	Excluded values	Distribution of categories (%)	Mortality rates (%)	Unadjusted odds ratio, 95% CI
GENDER	nil				
Male			559(71)	37(6.6)	2.238[0.983,5.096],p=0.055‡
Female			228(29)	7(3.1)	
AGE					
Age <70			518(65.9)	25(4.8)	
>70			268(34.1)	18(6.7)	OR 1.42[0.760,2.652], p=0.271
OPERATION	2(0.3)				
Other			261(33.2)	9(3.4)	reference
Oesophagectomy			331(42.1)	20(6)	2.190[0.928, 5.170], p=0.074‡
Total Gastrectomy			193(24.5)	14(7.3)	1.801[0.806, 4.024], p=0.153
HISTOLOGY	73 (9.3)	nil			
Adenocarcinoma			546(76.5)	34/546(5.9%)	0.661[0.295, 1.484], p=0.316
Benign			35(4.9)	0/35(0%)	0[0], p=0.998
Other			40(5.6)	1/40 (2.5%)	0.272[0.033,2.254], p=0.228
Squamous cell carcinoma			93(13)	8/93(8.6%)	Reference
TUMOUR (TNM) STAGE	146(18.6)	(119 (TX) &	26 "< equal 2")		
<=T2					Reference
T0			11(1.7)	(0.8)	0[0,], p=0.999
T1			55(8.5)	2(3.6%)	0.943[0.082, 10.901], p=0.963
T2			75(11.7)	6(8%)	2.174[0.249, 18.961], p=0.482
T3			437(68.2)	25(5.7%)	1.250[0.124, 12.603], p=0.850
T4			63(9.8)	3(4.8%)	1.517[0.197, 11.657], p=0.690
N0			325(41.3)	14(4.3)	Reference

Categorical predictors	Missing n (%)	Excluded values	Distribution of categories (%)	Mortality rates (%)	Unadjusted odds ratio, 95% CI
N1			387(49.2)	27(7)	1.666[0.858, 3.233], p=0.131
N2			7(0.9)	0	0[0,], p=0.999
RESPIRATORY					
Smoker	6(0.8%)				
Yes			533(68.2)	13/248(5.2%)	1.078[0.552, 2.105], p=0.826
No			248(31.8)	30/533(5.6%)	Reference
Smoker category	12(1.5%)				
Smoker			212(26.9)	10/212(4.7%)	0.845 [.363, 1.970] p=0.697
Ex-smoker			328(41.7)	20/328(6.1%)	1.109 [.540 2.276] p=0.778
Non-smoker			235(29.9)	13/235(5.5%)	Reference
Respiratory disease	22(2.8%)				
Yes			159(20.8)	5/159(3.1%)	0.499[.193,1.292],p=0.152
No			606(79.2)	37/606(6.1%)	Reference
Lung disease category	na	nil			
COPD	na		64(8.1)	4/64(6.3%)	Logistic regression not performed as no valid comparator
Chronic	na		31(3.9)	1/31(3.2%)	
Asthma	na		73(9.3)	0/73(0)%	
Acute	na		13(1.7)	1/13(7.7%)	
CARDIAC					
Cardiac disease	2(0.2%)	nil			
Not present			448(57)	23(5.1)	Reference
Present			338(43)	20(5.9)	1.162[0.627, 2.153], p=0.633
RCRI category (4 level)	nil	nil			
0			231(29.4)	11(4.8)	Reference
1			221(28)	9(4.1)	0.849[0.345,2.049], p=0.722

4. Clinical prediction model of perioperative mortality (Table: categorical predictors)

Categorical predictors	Missing n (%)	Excluded values	Distribution of categories (%)	Mortality rates (%)	Unadjusted odds ratio, 95% CI
	2		185(23.5)	10(5.4)	1.1143[0.474, 2.753], p=0.766
	3		150(19.1)	14(9.3)	2.059,[0.908, 4.666], p=0.084‡
RCRI category (2 level)	nil	nil			
	1		452(57)	20(4.4)	Reference
	2		335(42.6)	24(7.2)	1.667[0.905, 3.071], p=0.101‡

4.3.2 Selected plots of candidate predictor distributions and their relation to mortality

Mortality was plotted against predictors from the 'Clinical' model and those with p values less than 0.05 in univariate analysis (Figure 5 and Figure 6); these follow on the next pages. With the exception of the 61 to 64 year old age group, the relationship between age and mortality appeared non-linear with a greater effect in 7th, 8th and 9th decades. It was not clear whether this represented the typical logistic function or may instead benefit from the addition of a squared 'age' term. The addition of 'age squared' only changed the -2 log likelihood by 0.570 (p=0.45) for the loss of one degree of freedom providing evidence against any improvement in data fit. This lack of added predictive power for 'age square' has also been reported in 4080 patients from the Medicare system (Finlayson and Birkmeyer, 2001; Steyerberg, 2009d) and thereafter I used the simple linear age transformation.

Mortality also appeared to increase with and possibly accelerated with higher scores for the RCRI, as might be expected. However, it can be seen from the distribution that the numbers in the high scores were very low, and therefore it was possible that this observation could be random.

There were also hints of increasing mortality patterns as weight loss increased, and albumen decreased, although these were clouded by isolated deviations from the observed trends. Increasing white cell count was also associated with increasing mortality in univariate logistic regression, but on the plot it could be imagined that the pattern was U-shaped. This could be plausible, as one might imagine that patients with acute illness (high white cell count) or immunosuppression after chemotherapy (low white cell count) might have more post-operative complications. I ran a univariate association for a three category white cell count (low, normal and high), which was not statistically significant.

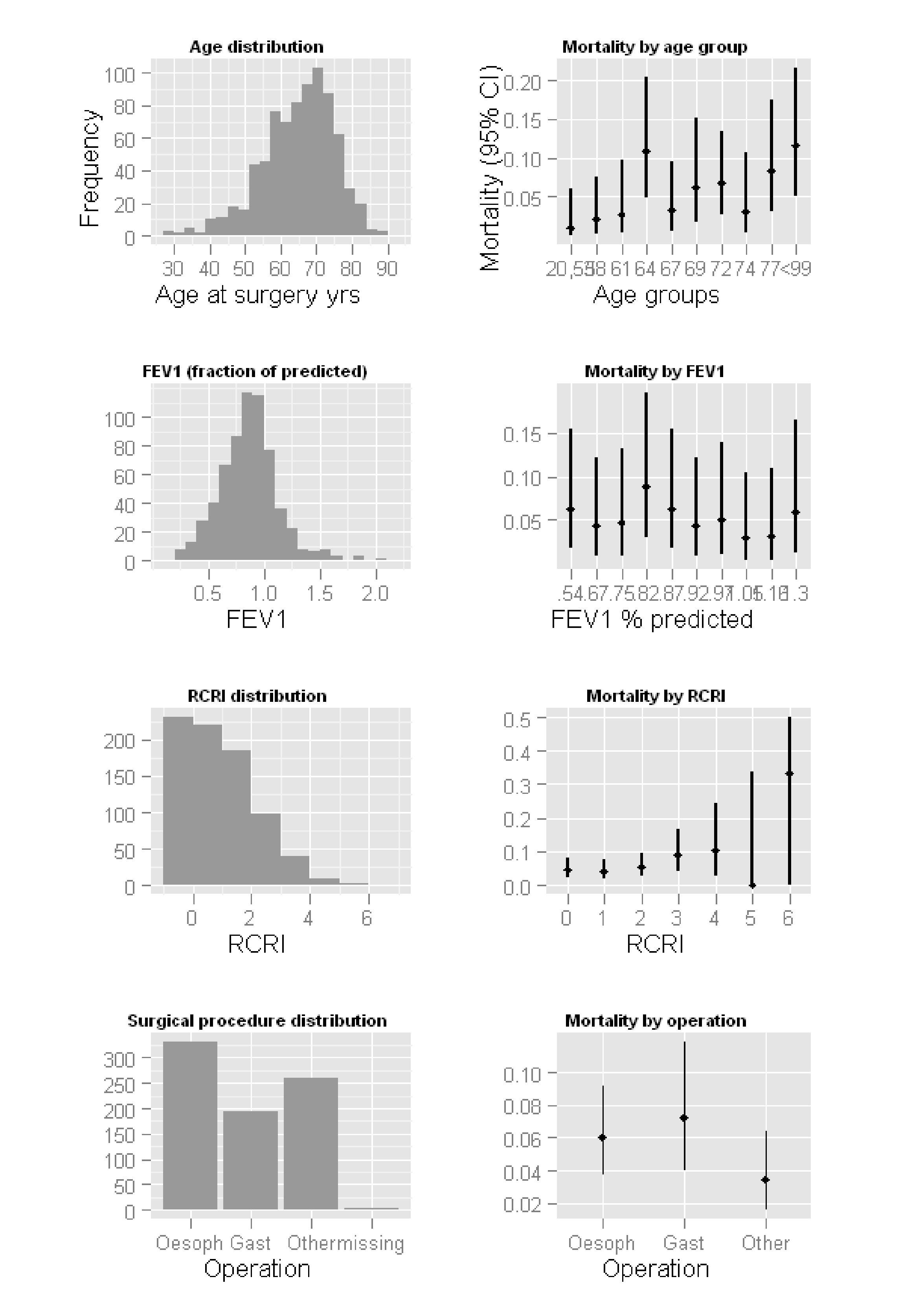


Figure 5 Distribution of candidate predictors and their association with mortality [mean & 95% CI] (a)

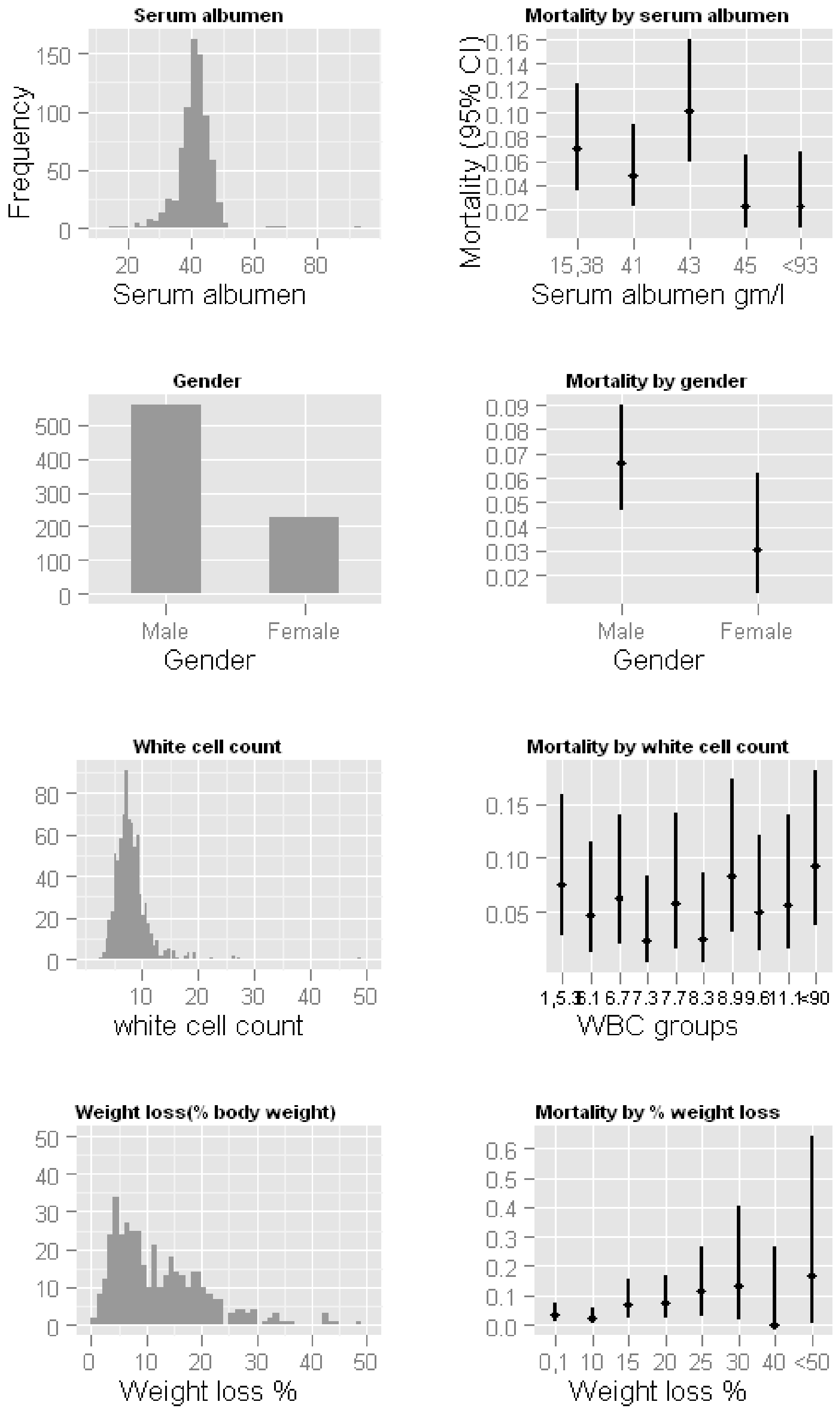


Figure 6 Distribution of candidate predictors and their association with mortality [mean & 95% CI)(b).

4.3.3 Selection of prediction model

To aid identification and referencing in text and tables, the individual models and their component predictors are listed in Table 21.

Table 21 Key to models and their constituent predictors

Model Name	Model description and constituent predictors
Clinical	age, operation, albumen, RCRI, FEV1(% predicted)
Clinical(I)	Based on imputed datasets with age, operation, albumen, RCRI, FEV1(% predicted)
Clinical_sex	Clinical model with gender
Clinical(I)_sex(R)	Based on clinical model with gender but excluded RCRI; based on imputed datasets
Clinical(I)_ sex	Based on clinical model from imputed datasets with gender
Clinical(I)_sex(INT)	Based on clinical model from imputed datasets with gender and RCRI*age interaction
Statistical1	Based on stepwise elimination; final model contained age, operation, gender, weight loss %
Statistical2	Based on stepwise elimination which did not contain weight loss at outset; final model contained age, gender
Prag	A ‘Pragmatic’ model based on simple complete dataset with reliable predictors (age, gender, operation)
Prag(INT)	Pragmatic with operation*age interaction

Clinical model

The prespecified ‘clinical’ model with age, operation, RCRI, FEV1, and serum albumen explained the data better than a constant only (Likelihood ratio test: χ_2 =16.038, 6 df, p=0.014). The Hosmer-Lemeshow statistic (p=0.547) did not provide evidence against a reasonable fit to the data (Table 22). Both age and operation remained significant for mortality after adjustment (Wald test for coefficients, P<0.1) (Table 23). Serum albumen, FEV1 and RCRI were not significantly associated with mortality after adjustment for the other predictors. Age was correlated with RCRI (r=0.5), so a reduction in significance values for both coefficients was expected, however there were 175 missing cases in the clinical model compared to the univariate analysis sample, therefore differences between the modelling samples may have contributed. Sample sizes and mortality rates for each model are shown in Table 22. Odds ratios generally ranged from 0.3 to 3.5 (thoracoabdominal oesophagectomy), and discrimination was moderate (AUC 0.696). Odds ratios for the predictors in each model are summarised in Table 23.

Statistical models

I have named ‘statistical’ models those which were developed using ‘data driven’ stepwise elimination. A global test of ‘no regression’ was performed on a dataset containing all predictors which were entered into a backward elimination process, as suggested by Harrell (Harrell, 2001d). The following predictors were initially included: all predictors from the clinical model, smoker status, respiratory disease, preoperative arterial $pO_{2,}$ percentage of reported weight loss, white cell count, tumour histology, T and N tumour stage (Deans and Patterson-Brown, 2009), gender and operation date. The initial model with all predictors explained variation in the dataset better than the constant (χ_{2} =35.698, 14 df, p=0.001). The final stepwise statistical model (excluding variables for p<0.05) contained age, percentage weight loss, operation and gender, and when run on a full dataset predicted mortality in the modelling sample better than the ‘clinical’ model (χ^2=36.216, 5 df, p<0.001, Nagelkerke’s R ^{2} 0.197, AUC 0.831). However, 30% of weight loss data was missing, which resulted in the stepwise procedure using only 375 cases with 18 deaths. Repeating the stepwise analysis without weight loss resulted in a model with only age and gender (43 deaths) and a reduced Nagelkerke's R ^{2} of 0.043 and AUC 0.667 (χ^2 =11.68, 2 df, p<0.003). These findings lend support to the inclusion of age and surgical procedure in the pre-specified 'clinical' model. Gender and weight loss were not initially considered for the pre-specified clinical model, but it would not be surprising if they were important, and they will be explored later in the section.

Table 22 Characteristics of prediction models developed from modelling sample (see Table 21 for key to models)

Prediction model	Model summary							Hosmer-Lemeshow		ROC	
	Cases(missing)	deaths	χ2	df	p value	-2LL	Nagelkerke's r2	χ2	p	AUC(se)	95% CI
Clinical	612(175)	30	16.038	6	0.014	223.399	0.08	6.998	0.537	0.696(0.047)	0.605, 0.788
Clinical(I)	785(2)	43	14.180 to 15.739	6	0.015 to 0.028	318.627 to 319.206	0.052 to 0.057	3.603 to 13.345	0.101 to 0.891	0.680 to 0.682	0.599 to 0.604, 0.760 to 0.763
Clinical_ sex	612(175)	30	19.843	7	0.006	219.953	0.099	4.38	0.821	0.719(0.04)	0.633, 0.805
Clinical(I)_sex(R)	785(2)	43	20.246 to 22.738	7	0.002-0.005	310.648 to 313.104	0.074 to 0.083	4.152 to 10.358	0.241 to 0.843	0.698 to 0.706	0.623 to 0.631, 0.773 to 0.779
Clinical(I)_ sex	785(2)	43	20.034 to 22.594	6	0.001 to 0.003	310.792 to 313.351	0.073 to 0.082	1.571 to 6.375	0.605 to 0.991	0.696 to 0.703	0.620 to 0.631, 0.772 to 0.778
Clinical(I)_sex(INT)	785(2)	43	23.307 to 26.280	8	0.001 to 0.003	308.901 to 310.079	0.085 to 0.089	1.571 to 6.375	0.667 to 0.991	0.696 to 0.705	0.620 to 0.631, 0.772 to 0.778
Statistical1	546(241)	27	36.216	5	0.001	178.792	0.197	6.345	0.609	0.831(0.03)	0.772, 0.891
Statistical2	786(1)	43	11.68	2	0.003	321.819	0.043	3.976	0.859	0.667(0.04)	0.594, 0.740
Prag	785(2)	43	17.121	4	0.002	316.265	0.062	10.47	0.234	0.691(0.038)	0.616, 0.765
Prag(INT)	785(2)	43	22.608	6	0.001	310.778	0.082	5.686	0.682	0.712	0.640, 0.784

Table 23 Odds ratios (95% CI) and statistical significance for predictors in each prediction model

	Clinical	Clinical_sex	Clinical(I)	Clinical(I)_sex	Statistical1	Statistical2	Pragmatic	Clinical(I)_sex(R)	Prag(INT)	Clinical(I)_sex(INT)
Predictors										
RCRI	0.997(0.715,1.392); p=0.988	0.963(0.687,1.351); p=0.829	1.084; p=0.565	1.063(0.665)	NA	NA	NA	NA		0.37; p=0.1
Albumen	0.945(0.871,1.025); p=0.174	0.953(0.878,1.034); p=0.246	0.956; p=0.136	0.958(0.169)	NA	NA	NA	0.96; p=0.18		0.961; p=0.191
FEV1 (%predicted)	0.304(0.065,1.409); p=0.128	0.337(0.071,1.594); p=0.170	0.438; p=0.29	0.47(0.334)	NA	NA	NA	0.468; p=0.322		0.425; p=0.285
Age	1.994(1.157,3.434); p=0.0129	2.068(1.190,3.596); p=0.010	1.665; p=0.022	1.707(0.017)	2.147(1.273,3.622); p=0.004	1.608(1.131,2.286); p=0.008	1.729(1.195, 2.503); p=0.004	1.79; p=0.003	3.115; p=0.014	1.249; p=0.401
Operation		p=0.064			p=0.011		p=0.09	p=0.015	p=0.053	
Thoracoabdominal oesophagectomy	3.504(1.248,9.841); p=0.017	3.347(1.193,9.392); p=0.022	2.813; p=0.02	2.68(0.026)	6.595(1.914,22.720); p=0.003		2.160(0.946, 4.934); p=0.068	2.641; p=0.028	144.631; p=0.032	2.905; p=0.019
Total gastrectomy	2.655(0.903,7.807); p=0.076	2.746(0.931,8.097); p=0.067	2.691; p=0.028	2.69(0.028)	3.677(0.976,13.848); p=0.054		2.563(1.071, 6.135); p=0.035	2.698; p=0.027	8.828; p=0.409	2.907; p=0.02
Gender	NA	2.665(0.899,7.905); p=0.077	NA	2.097(0.083)	12.841(1.700, 97.025); p=0.013	2.245(0.981, 5.140); p=0.056	2.253(0.980, 5.175); p=0.056	2.123; p=0.077	2.251; p=0.057	2.15; p=0.075

	Clinical	Clinical_sex	Clinical(I)	Clinical(I)_sex	Statistical1	Statistical2	Pragmatic	Clinical(I)_sex(R)	Prag(INT)	Clinical(I)_sex(INT)
% weight loss	NA	NA	NA	NA	1.058(1.017,1.102); p=0.006		NA	NA		NA
Constant	0.047; p=0.122	0.013; p=0.041	0.047; p=0.039	0.021(0.014)	0.001; p=0.00	0.005; p=0.00	0.002; p=0.00	0.018; p=0.01		0.061; p=0.087
Interactions										
Age*RCRI interaction										1.289; p=0.066
Age*operation interaction									p=0.07	
Age*oesophagectomy interaction									0.355; p=0.05	
Age*gastrectomy interaction									0.776; p=0.668	

4.3.4 Models based on imputation datasets

Complete case analysis in the pre-specified ‘Clinical’ model resulted in 175 missing cases and the loss of about 30% of events. Therefore I tried to optimise available data using multiple imputations to replace missing values from serum albumen (4% missing) and FEV1 (8 % missing), both of which had reasonably ‘normal’ distributions.

SPSS generated 5 imputed datasets for each predictor, all of which had means, standard deviations and ranges, which were similar to the original data. This increased the number of deaths available to study from 27 to 43 and reduced the missing cases to 2. Logistic regression (in SPSS) was performed on the imputed datasets and SPSS produced pooled averaged regression coefficients for each model. The performance measures (Nagelkerke’s, -2LL etc.) for the models were not pooled in SPSS and were given as ranges for imputation based models in the summary tables.

The regression results for the ‘Clinical’ models were similar for both the original and the imputed datasets, but the odds ratios were reduced for age (1.994 to 1.665) and the oesophagectomy operation category (3.504 to 2.813) in the imputed data, suggesting possible overfitting in the original smaller sample (Table 23).

4.3.5 Exploration of weight loss, gender, RCRI and operation

Gender and ‘weight loss’ were identified in the stepwise elimination model as potentially important predictors in the model and warranted further investigation.

Gender

In the stepwise elimination model, which included ‘weight loss’, gender was highly significant (p=0.013) with a large OR (12.84, 95% CI 1.700, 97.025) and confidence interval. The odds ratio reduced to 2.665 (95% CI 0.899, 7.905) when weight loss was excluded from the stepwise model. This was possibly caused by an imbalance in deaths between males and females in missing and non-missing data, caused by the 30% missing ‘weight loss’ predictor (Table 24). Female mortality rate was 0.6% (one death) when cases with ‘weight loss’ were non-missing, and 8.8% (6 deaths) in missing cases. This imbalance may also account for the extremely large odds ratio (12.84) for gender.

Table 24 Mortality count (%) within gender for missing and non-missing weight loss data

	Male	Female	Total
Non missing	26(6.7%)	1(0.6%)	27
Missing	11(6.4%)	6(8.8%)	17

It is possible that there was a systematic difference between genders for other important predictors, however this did not appear to be the case. Males had a modestly higher rate of oesophagectomy relative to other operations, a higher rate of adenocarcinoma, and about 15% lower non-smoker rate. The latter was not a predictor of outcome and there were no other significantly unbalanced predictor distributions to explain the results. It is possible that other unobserved predictors could have been unevenly distributed between genders. The means for continuous variable predictors are given in Table 25 and the distributions of categorical variable predictors are given in Table 26.

Table 25 Continuous predictor means by gender

	Male	Female
Age	**65.38**	**65.59**
Serum albumen	**41.45**	**41.76**
Weight loss (%)	**8.86**	**9.60**
pO2 (kPa)	**12.31**	**12.22**
FEV1 (% predicted)	**0.86**	**0.91**

Gender is considered important for many outcomes and perhaps I should have considered including it at the outset, but there was no strong evidence for its inclusion from the systematic review, and there was a limit on how many predictors could have been included given the relatively small sample size. This modest evidence of effect together with the background of its importance suggests that it should be included in the 'Clinical' model.

Table 26 The distribution of categorical predictors between genders

		Male	Female
Operation	Other	180(32.3%)	81(35.5%)
	Thoracic Oesophagectomy	246(44.2%)	85(37.3%)
	Total gastrectomy	131(23.5%)	62(27.2%)
Histology	Adenocarcinoma	403(81.1%)	143(65.9%)
	Benign	22(4.4%)	13(6.0%)
	Other	21(4.2%)	19(8.8%)
	Squamous cell carcinoma	51(10.3%)	42(19.4%)
T stage	< or equal	21(4.4%)	5(2.6%)
	T0	7(1.5%)	4(2.1%)
	T1	33(6.9%)	22(11.5%)
	T2	51(10.7%)	24(12.5%)
	T3	318(66.9%)	119(62.0%)
	T4	45(9.5%)	18(9.4%)
N stage	N0	224(44.1%)	101(47.9%)
	N1	279(54.9%)	108(51.2%)
	N2	5(1.0%)	2(0.9%)
Smoker status	1	148(27.0%)	64(28.3%)
	2	256(46.6%)	72(31.9%)
	3	145(26.4%)	90(39.8%)
Respiratory disease	None	425(78.3%)	181(81.5%)
	Present	118(21.7%)	41(18.5%)

Weight loss

Weight loss was apparently a strong predictor in the statistical model (Nagelkerke's R ^{2} 0.197, AUC 0.831) and clinical knowledge would support this. However, nearly 30% of the data was missing and its effect could have been random, therefore it was not considered credible to use a model with weight loss. It was also not clear in the database whether blank cells meant missing or zero and therefore the replacement of this quantity of data by imputation methods seemed to lack 'face validity'.

Revised Cardiac Risk Index (RCRI)

There was no missing data for this predictor, but I also investigated its apparently weak effect in imputed datasets for the 'clinical' model because cases were excluded as a result of other missing predictors. Its correlation coefficient with age was 0.5 (Table 27), which probably explains its loss of predictive effect in the multivariate model (χ^2 ranged from 0.08 to 0.136, 2 degrees of freedom, p values all >0.7 in the imputed datasets) and clinical models with and without RCRI had similar predictive power. The inclusion of RCRI did not add predictive power but in view of the supporting evidence I kept it in the model.

Surgical procedure

The similar odds ratios and confidence intervals for the different operations, thoracoabdominal oesophagectomy and total gastrectomy (2.812 vs 2.691), compared with 'other' operations suggest a two category surgical predictor could be used ('major' cancer resection and 'other'). In all imputation sets the inclusion of 'operation' (χ^2 6.498 to 7.233, 2 degrees of freedom, p<0.04) improved the clinical model. 'Operation' results were similar whether coded as two or three categories, and therefore, although this used another 'degree of freedom', I included the three category 'operation' predictor to allow an 'oesophagectomy specific' model.

Table 27 Correlation between predictors in the modelling sample (Spearmans rho)

	Gender	RCRI	albumen	FEV1 (% predicted)	Operation	Weight loss (%)
RCRI	-0.043; p=0.232 (n=787)					
albumen	0.053; p=0.143 (n=755)	0.000; p=0.990 (n=755)				
FEV1 (% predicted)	0.056; p=0.155 (n=636)	0.010; p=0.800 (n=636)	.156; p=0.000 (n=613)			
Operation	0.054; p=0.129 (n=785)	.133; p=0.000 (n=785)	-.169; p=0.000 (n=753)	-.174; p=0.000 (n=635)		
Weight loss (%)	0.044; p=0.307 (n=541)	0.025; p=0.567 (n=541)	-.235; p=0.000 (n=523)	-0.049; p=0.303 (n=437)	.135 p=0.002 (n=540)	
Age	0.023; p=0.520 (n=786)	.560; p=0.000 (n=786)	-.137; p=0.000 (n=754)	.081; p=0.042 (n=636)	.152; p=0.000 (n=785)	.100; p=0.020 (n=541)

4.3.6 A 'Pragmatic' model

I subsequently considered a model based on simplicity, reliable complete data, and with reasonable evidence of 'face validity' for its constituent predictors, based on the results of this study and other published information. This 'Pragmatic' model included age (strong evidence), gender (some evidence) and surgical procedure (to focus on 'oesophagectomy'). The 'clinical' models with more predictors explained more variance in the modelling sample, and discriminated better then the 'Pragmatic' model, as would be expected.

4.3.7 Modelling assumptions

Logistic regression assumes a binomial distribution for outcome, and additivity of predictor effects but makes no distributional assumptions about the predictors. Linearity of response is not essential but desirable for a stable model (Harrell, 2001a). The natural log of the odds ratio for mortality was linearly related to age but not obviously with FEV1, serum albumen and RCRI were less clear. There appeared to be no obvious data transformations which could be applied to the predictors.

4.3.8 Interactions between main predictors

The relatively few degrees of freedom compared to the number of events gave little scope to investigate interactions between the main effects without 'overfitting'. One might expect the effects of cardiac, respiratory and nutritional comorbidity to be different in younger and older age groups so I focussed on these in the 'clinical' model and I also examined interactions between the main predictors of the 'Pragmatic' model. These are summarised in Table 28 and Table 29.

Table 28 Effect of adding interactions to the 'clinical' model for original dataset (n=612) and full imputed datasets. Chi-square & p value for likelihood ratio tests and OR are given for the addition of each interaction. Medians and ranges are given for each of the imputed datasets.

	Original dataset		Imputed datasets		
Interaction	**Chisq (p)**	**Odds ratio (p)**	**Median chisq (min,max)**	**Median p value(min,max)**	**Pooled OR (p)**
Age*fev1	0.002(0.962)	0.995(0.962)	0.123(0.01,3.26)	0.726(0.07,0.91)	1.578(0.617)
Age*albumen	1.069(0.301)	0.952(0.31)	3.05(2.56,4.35)	0.081(0.04,0.11)	0.940(0.083)
Age*rcri	0.039(0.843)	0.96(0.843)	2.964(2.85,3.29)	0.085(0.07,0.09)	1.273(0.075)
Fev1*rcri	0.058(0.809)	1.162(0.808)	0.129(0.01,1.5)	0.719(0.22,0.92)	1.231(0.72)
Age*gastrectomy		1.467(0.617)			0.817(0.739)
Age*oesophag		0.731(0.617)			0..369(0.06)

Table 29 Effect of adding interactions between main effects in the 'pragmatic' model; chi square and p value for the likelihood ratio addition step in SPSS.

Interaction	Chisquare(p value)	OR (p value)
Age*gender	0.013(0.918)	1.057(0.907)
Gender*operation	3.491(0.175)	
Gender*oesophagectomy		0.118(0.091)
Gender*gastrectomy		0.487(0.467)
Age*oesophagectomy		0.355(0.05)
Age*gastrectomy		0.776(0.668)

There was weak evidence that the effect of serum albumen (chronic illness or malnutrition) and cardiovascular comorbidity may be altered by age. This was only apparent in the nearly complete imputation sets (Table 28). The mortality rates by age group and RCRI score are shown in Table 30, where it can be seen that most of the cells for old age and high risk are empty because of the low prevalence of high risk scores in the elderly. This does not give good evidence to conclude an interaction.

Table 30 Mortality grouped by RCRI score and age quintile (fatalities/total cases per cell (%) in original modelling dataset

RCRI score	0	1	2	3	4	5	6
Age ascending quintiles (yrs)							
<=58	3/97(3.1)	0/51(0)	0/23(0)	0/5(0)	0/3(0)	0/0(0)	0/0(0)
<=64	4/72(5.6)	4/44(9.1)	2/17(11.8)	0/8(0)	0/2(0)	0/0(0)	0/0(0)
<=69	2/44(4.5)	2/48(4.2)	1/39(2.6)	2/15(13.3)	0/3(0)	0/0(0)	0/0(0)
<=74	2/13(15.4)	0/46(0)	3/59(5.1)	2/37(5.4)	2/10(20)	0/2(0)	0/1(0)
>74	0/0(0)	3/32(9.4)	4/47(8.5)	4/33(12.1)	2/21(9.5)	0/5(0)	1/2(50)

There was some evidence that the effect of age varied in different operations in both Clinical and Pragmatic models (Table 29). For example, the odds ratio for the age*oesophagectomy interaction is 0.355 i.e. the ratio of the effect of a unit change in age on mortality for oesophagectomy relative to the same change in 'other' operations. The odds ratio for the age*gastrectomy interaction is 0.776, so the ratio of the two suggests that the effect of age for gastrectomies is about twice that for oesophagectomies. Calculating the same ratios for patients under 64 and over 74 from cross tabulation (Table 31) gives a ratio of an age effect about three times greater in gastrectomies. Comparisons were on slightly different data as the tables were based on age group quintiles whereas the logistic regression was on age as a continuous variable, but the overall trend was apparent. The explanation is not clear, but one could speculate that thoracoabdominal oesophagectomy is seen as a more stressful operation, and therefore selection is more conservative in the elderly, so only the fittest elderly patients undergo this operation.

Table 31 Mortality (count (%)) in age groups for each type of operation in the modelling dataset

	Operation (count (%))		
	"Other"	Total gastrectomy	Oesophagectomy
Age ascending quintiles			
<=58	0/49(0)	0/52(0)	3/81(3.7)
<=64	2/38(5.3)	1/26(3.8)	7/80(8.8)
<=69	0/50(0)	5/41(12.2)	2/60(3.3)
<=74	1/63(1.6)	2/36(5.6)	6/69(8.7)
>74	6/61(9.8)	6/38(15.8)	2/41(4.9)
Total	9(3.4)	14(7.3)	20(6)

4.3.9 Outliers and influential values

Standardised residuals represent the difference between the predicted probability and outcome, which is either 0 or 1, and are therefore difficult to interpret. They are normally distributed with a mean of zero and standard deviation of 1, therefore one would expect about 95% to be less than 2 and 99% to be less than 2.5. Ninety five per cent of standardised residuals were less than 2 in the clinical model and 98.5% less than 2.5 in the clinical model, which included gender. Unsurprisingly all cases with residuals greater than 2 were fatalities (low prevalence of mortality and weakly predicting model). Only about 1% of residuals exceeded 2 in the Pragmatic model.

Leverage, which should lie between 0 and 1 and gives an estimate of the effect on the overall model of that case, was less than 0.1 for all models, except for one case in the Clinical model. This (case 557) value was 0.7 and its DfBeta (gives an estimate of that case's effect on the regression coefficient and should be less than 1) for the constant was about 1.2 in both Clinical models. Case 557 was an elective surgical case, who had a subtotal gastrectomy and a hemicolectomy, was obese, chronically ill (serum albumen 28 gm/lit and white cell count 19.3), had poor respiratory function and various cardiac comorbidities and did not survive. This case was at the extreme end of poor health before elective surgery, but I could see no reason to exclude it from the model.

4.3.10 Validation

I decided to test four base models. Firstly, the prespecified Clinical model derived on original and imputed datasets. Secondly the Clinical model (imputed) including gender and a 'Pragmatic' model which contained reliable predictors from a complete dataset i.e. age, gender and operation. Finally, I decided to apply the 'Rotterdam' model to the validation sample.

The validation sample

The number of cases for continuous predictors and their means are shown in Table 32. The number of cases and distributions of categorical predictors are shown in Table 33. There were 785 cases with 43 deaths in the validation sample. Missing values were similar in both modelling and validation samples. There were considerably more high RCRI scores in the modelling sample (Table 33).

Table 32 Case summary and means of predictors in development & validation samples

	Development		Validation	
	cases	Mean (sd)	cases	Mean (sd)
Age	786	65.44 (10.19)	785	65.16 (10.43)
RSRI (score)	787	1.41 (1.26)	785	1.75 (1.77)
Albumen	755	41.54 (5.45)	769	41.68 (4.69)
% weight loss	547	9 (9)	561	9 (9)
FEV1 (% predicted)	636	0.87 (0.25)	641	0.93 (0.28)

Table 33 Cases, mortality and distribution of categorical predictors in development and validation samples

Predictor	Development (number of cases)	Validation (number of cases)
Total cases	787	785
Deaths	44	43
RCRI (total score)		
0	231	233
1	221	196
2	185	140
3	99	85
4	39	62
5	9	36
6	3	24
7	0	6
8	0	1
9	0	1
13	0	1
Operation		
Oesophagectomy	331	356
Total gastrectomy	193	184
Other operation	261	245

Predictor	Development (number of cases)	Validation (number of cases)
[20]**Gender**		
Male	228	231
Female	559	554

The prediction models and their associated validation sample

Descriptions of the prediction models, the number of complete cases and mortalities (observed and predicted), with which each model was associated, are shown in Table 34 on the following page. Models based on the 'Clinical' model were validated on smaller samples (635 cases and 29 deaths) than the 'Pragmatic' or 'Rotterdam models (785 cases and 43 deaths) because of the missing data on their comorbidity predictors in the validation sample. The 'statistical' model was validated on 561 cases (33 deaths) because of missing data in weight loss.

Cases, mortality and distribution of categorical predictors in development and validation samples

20

Table 34 Description of validated models, associated case numbers and observed and predicted mean mortality

Model	Description	Cases (missing)	Deaths	Observed mortality	Mean predicted mortality	Range of predicted mortality (2 dec places)
Clinical	Prespecified with age, operation, RCRI, FEV1, albumen	635(150)	29	0.051	0.051	<0.01, 0.23
Clinical(I)	As above from imputed datasets	635(150)	29	0.051	0.058	<0.01, 0.22
Clinical(s)	Clinical model with shrinkage factor applied	635(150)	29	0.051	0.046	0.01, 0.13
Clinical(INT)	Clinical with interaction between age & RCRI (imputed)	635(150)	29	0.051	0.057	<0.01, 0.3
Statistical	Data driven from stepwise regression; age, weight loss, operation, gender	561(224)	33	0.059	0.031	<0.01, 0.18
Pragmatic	Age, operation, gender	785(0)	43	0.055	0.055	<0.01, 0.2
Prag(INT)	As above with operation * age interaction	785(0)	43	0.055	0.055	<0.01, 0.29
Rotterdam	Rotterdam model	785(0)	43	0.055	0.061	0.02, 0.17

Model Calibration

Calibration plots were generated using the function val.prob.ci (Vergouwe and Steyerberg, 2009). The main components of the output are summarised below.

1) Calibration in the large' gives a measure of whether the prediction model is accurate for the overall mean mortality for the whole validation sample. It is reported as an odds ratio for the average under or overestimation of the mortality (Steyerberg, 2009h, p. 272) and is given as the 'Intercept' on the calibration plot. A negative intercept implies the model is overestimating mortality in general and vice versa for positive values. I confirmed that the 'Intercept' values from val.prob.ci tally with the calculation which gives the odds ratio for overall calibration (in SPSS) (Steyerberg, 2009e, p. 272). This is calculated by running a logistic regression of the observed outcomes in the validation sample against logit of their predicted probabilities, which have been generated from the modelling sample equation. The odds ratio for over or under calibration is then the odds (mean [predicted mortality])/odds (mean [new predictions]).
2) The solid line ("Logistic calibration") is generated from the logistic regression of the observed outcomes against the logit of their predicted mortalities in the validation dataset. The plot is of the new mortality predictions against the logit of the original predictions, representing how well the predictions from the model explain the outcomes in the validation sample (Harrell, 2001e, p. 250; Steyerberg, 2009h, p. 272). The slope quantifies this; a perfect predictor would have a slope of one, represented by the dashed ("Ideal") line in the plot.
3) Local regions of poor calibration can be shown graphically. One method is to plot the mean observed mortality (y axis) against the mean predicted mortality (x axis), for deciles of patients after sorting predicted probabilities in ascending order, as in the Hosmer-Lemeshow 'goodness of fit' statistics (Lemeshow and Hosmer, 1982). This plot is represented by triangles on the plot and each has confidence intervals (95%) for the

observed mortality. The dotted line ("Nonparametric") uses the LOWESS smoother, a non-parametric algorithm which fits a smoothed trend for individual points and allows visualisation of local areas of poor fit (Cleveland, 1979).

4) The histograms, which are adjacent to the calibration plots, were constructed using 'ggplot2' (Wickham, 2009) and show the distributions of the predicted mortalities for survivors and non-survivors (Appendix G iv.).

Calibration plots and histograms for the predicted mortalities of selected models are shown in Figure 7 to Figure 10 on the following pages. The odds ratios for overall mis-calibration ranged from -0.01 to -0.2 for the 'Clinical' model and its variants to -0.13 for the 'Rotterdam' model and 0.7 for the 'Statistical' model. The 'Pragmatic' models were best calibrated overall with an 'Intercept' of -0.01. The range of predicted mortalities was mostly from zero to about 0.2, but up to 0.3 in models, which included interactions (age with RCRI in the 'Clinical' model and age with operation in the 'Pragmatic' model). The predictions explained observed mortality best in the 'Pragmatic' models (slope of 0.85) and worst in the 'Statistical' (slope 0.37), with 'Clinical' based models intermediate. These values reflected a tendency to overestimate mortality in the higher prediction ranges, which was also demonstrated in the Lowess plots for predictions over about 10%. The 'Rotterdam' model was mis-calibrated in the opposite direction with underestimation at predictions over 10%.

Running the 'Clinical' model on imputed datasets improved the prediction slope (increased from 0.44 to 0.62) but 'calibration in the large' deteriorated (from -0.1 to -0.22). Adding the age*RCRI interaction in the 'Clinical' (imputed) model improved 'calibration in the large' but the calibration slope deteriorated. All the 'Clinical' models overestimated mortality at the higher predictions, so I also investigated the application of a shrinkage factor to the original 'Clinical' model. Prediction was examined with and without application of a 'shrinkage' factor, which is a method of dealing with overfitting. I used a simple uniform shrinkage factor after coefficient estimation using the formula:

s =(model X^2 -df)/ model X^2 (Copas, 1983; Steyerberg, 2009j).

This considerably solved the problem of overestimation of mortality but at the cost of narrowing the prediction range to between 1% and 13%, which was unlikely to be practically useful. The calibration plots for the models are shown in Figure 7 to Figure 10 on the following pages.

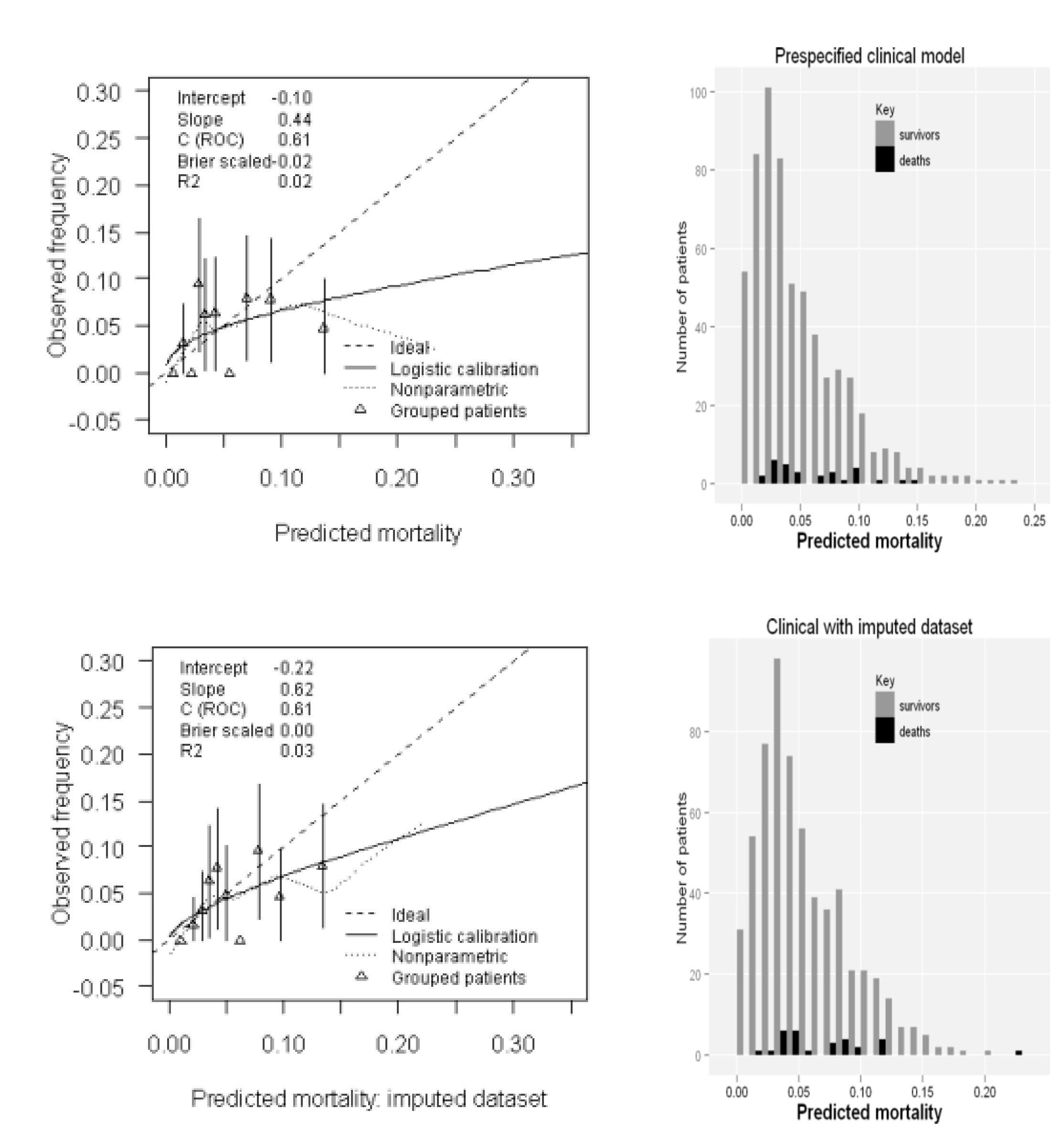


Figure 7 Calibration and distribution of predicted mortalities for the prespecified 'clinical' model and the ‘clinical’ model derived from imputed datasets.

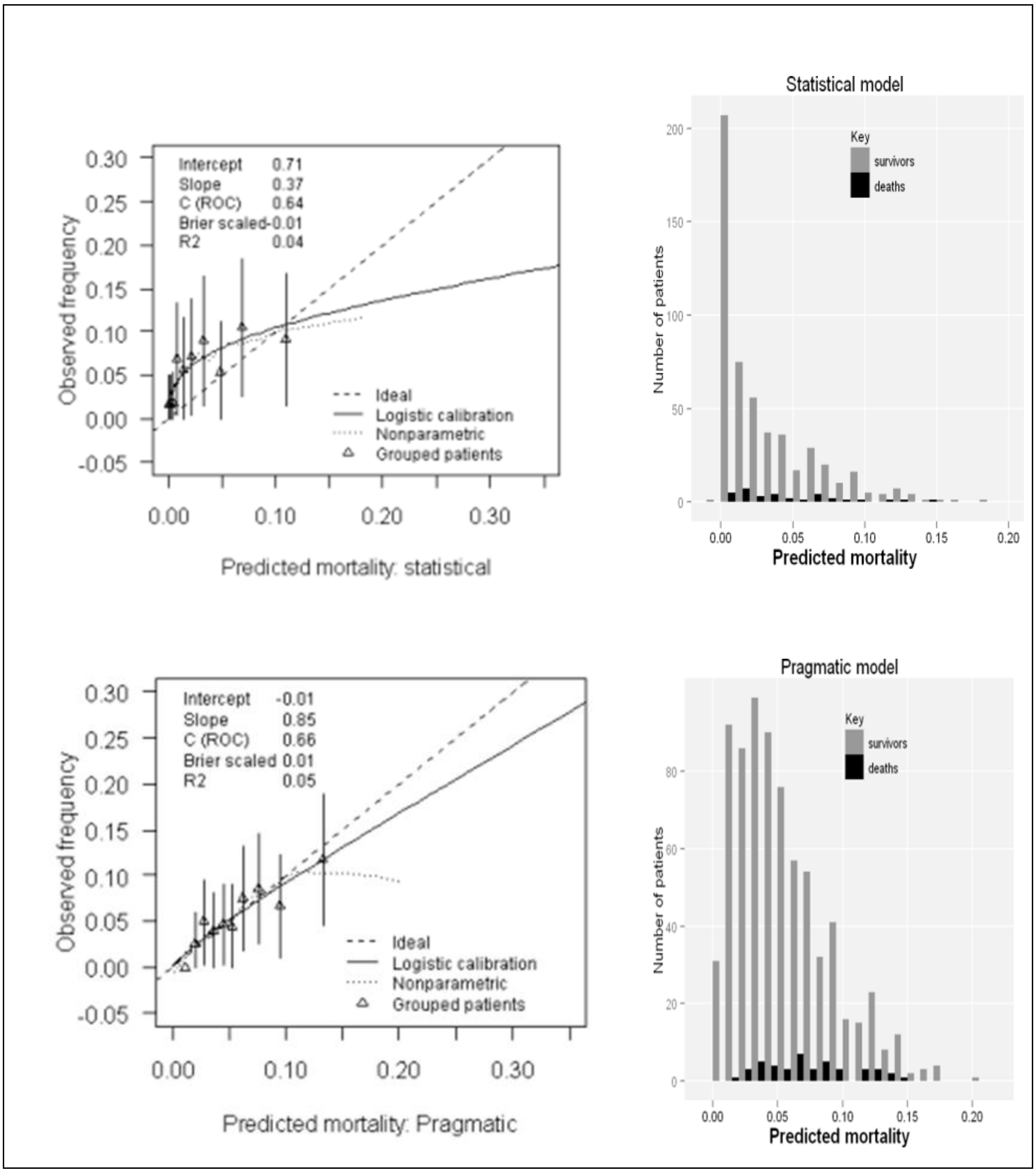


Figure 8 Calibration and distribution of predicted probabilities for the 'statistical' and 'pragmatic' models

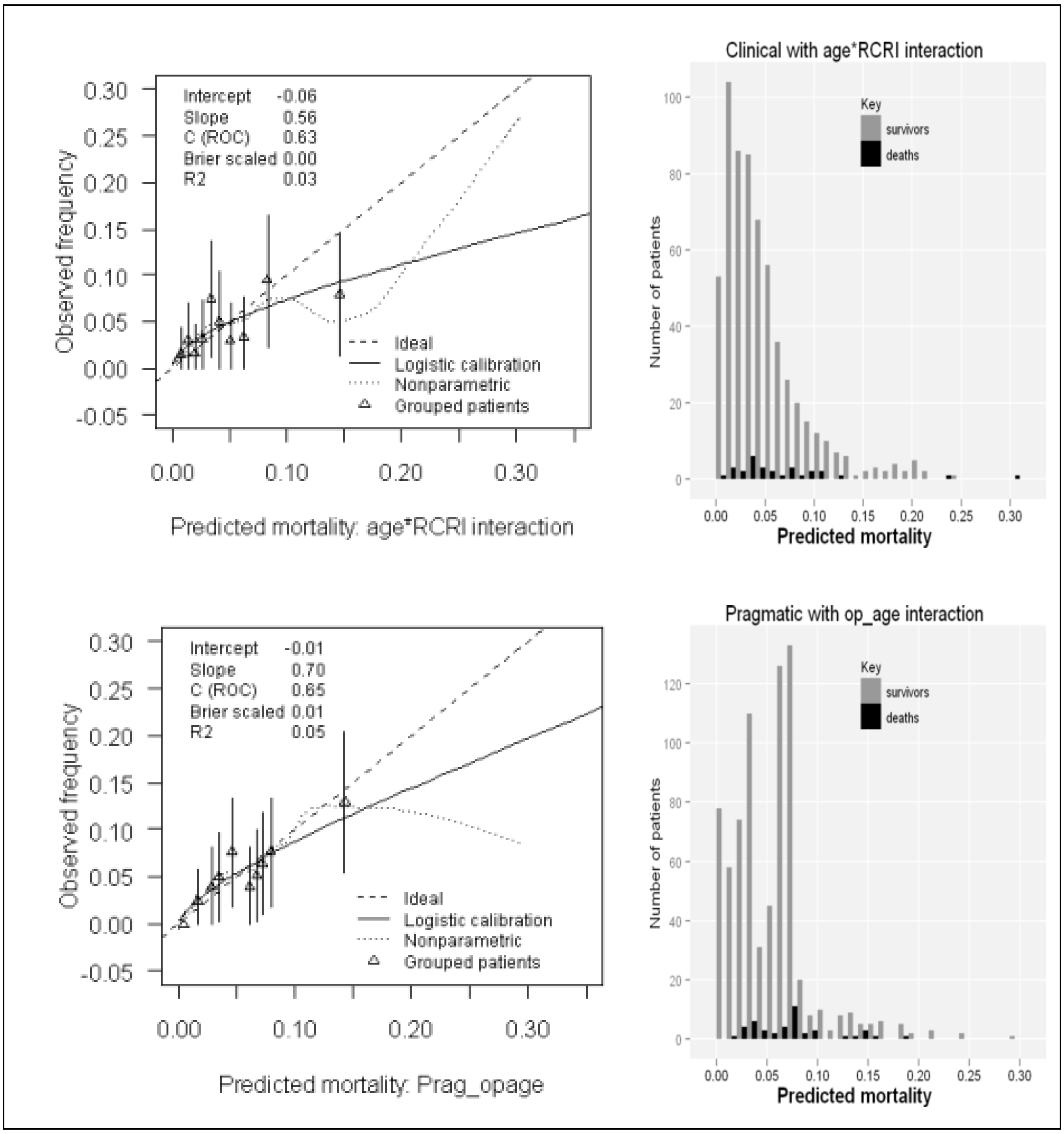


*Figure 9 Calibration plots and distribution of mortality predictions for the 'clinical' model with age*RCRI interaction and 'pragmatic' model' with age*operation interaction.*

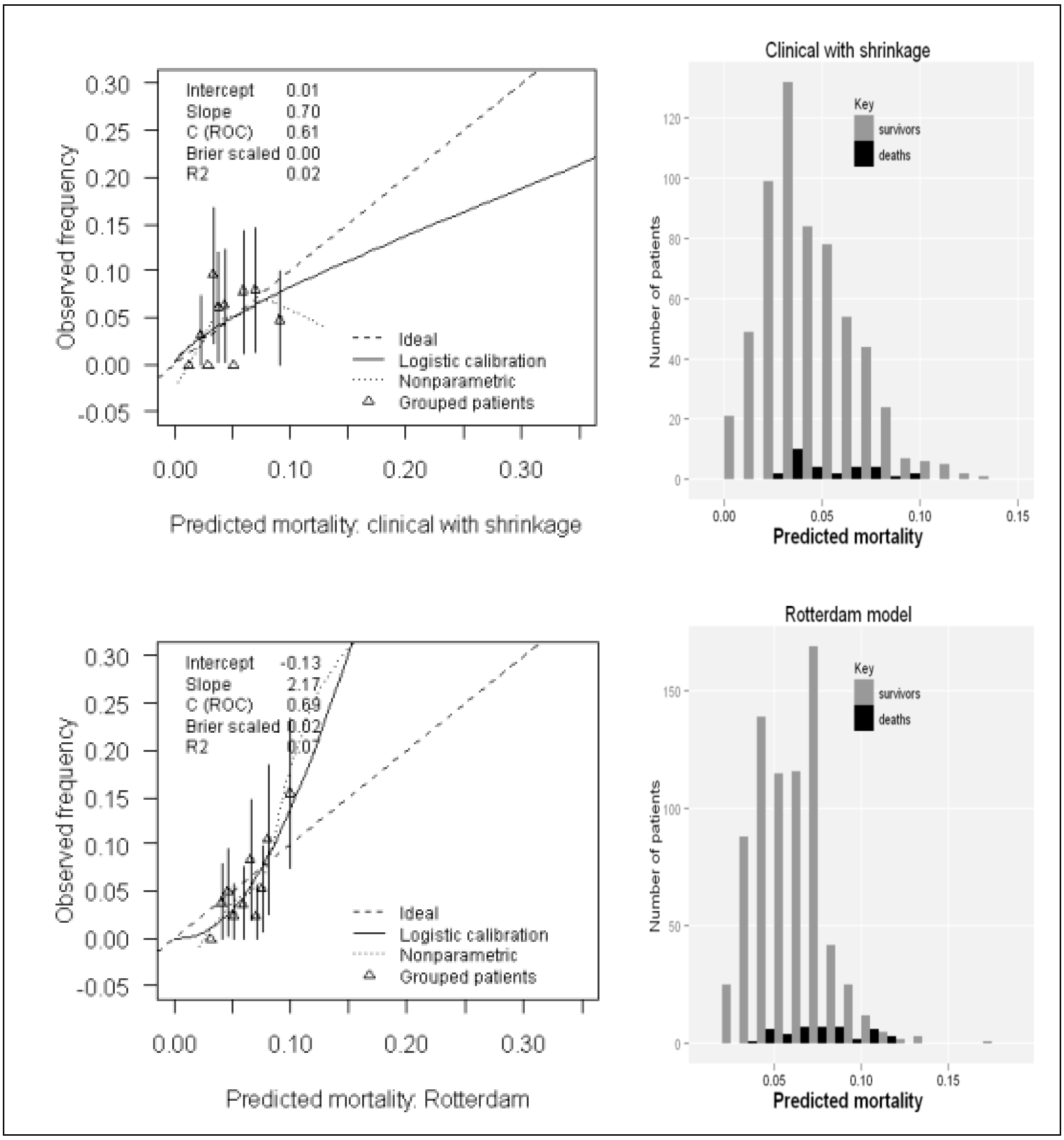


Figure 10 Calibration plots and distribution of mortality predictions for the 'Rotterdam' model and the 'clinical' model with shrinkage applied

Model discrimination

The c statistic for the receiver operator curves ranged from 0.603 to 0.664 in the models developed in this study, and was 0.721 for the Rotterdam model. This indicates at best a moderate capacity to discriminate between survivors and non-survivors.

ROC curves generated from the 'plotROC' function are shown below (Appendix G vi.). Both sets of plots show that discrimination for these models is not particularly good, consistent with the predictors not explaining much of the variation in the datasets. The first plot shows the effects of modifications to the pre-specified 'Clinical' model by using imputed datasets and including an 'age*RCRI' interaction, which produced marginal small improvements.

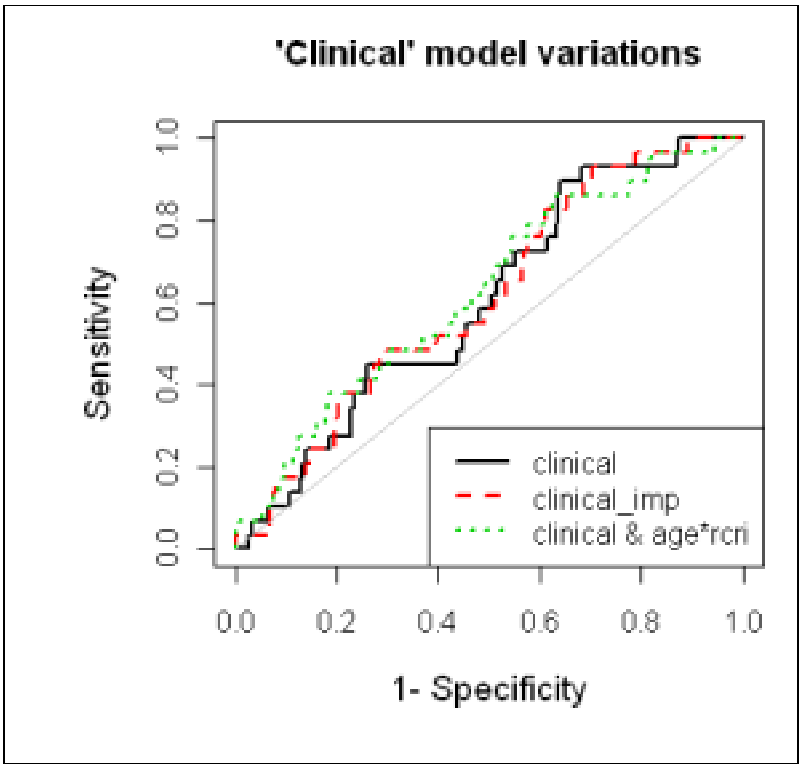


*Figure 11 Receiver operator curves for the 'clinical' model derived from original data, imputed data and with an age*RCRI interaction.*

The following ROC plot shows generally poor discrimination for all models, but the simple ‘Pragmatic’ model, and the ‘Rotterdam’ model both fared better than the ‘Statistical’ model which was developed and validated on datasets with much missing data in the weight loss variable.

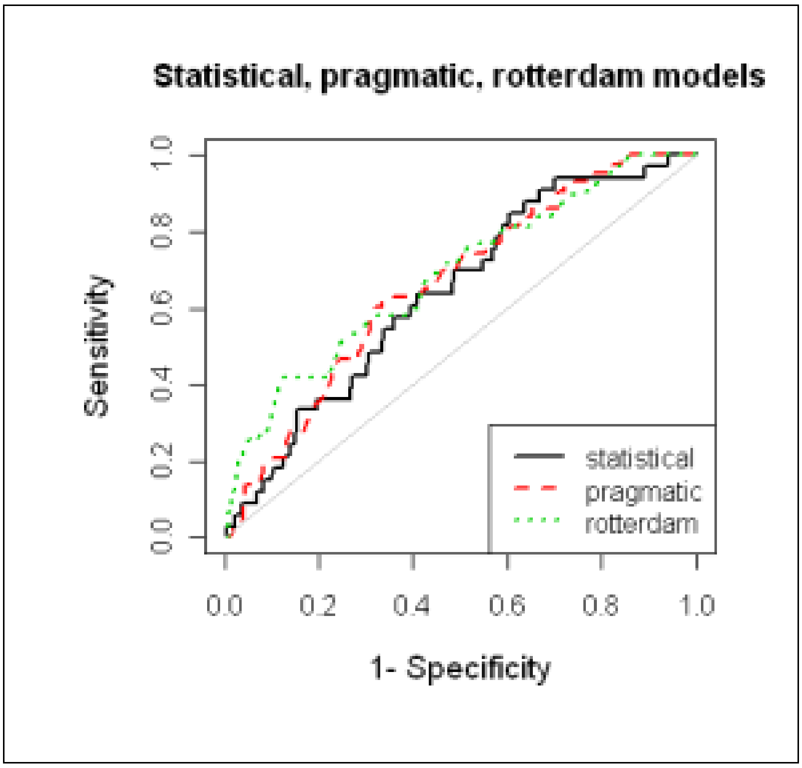


Figure 12 Receiver operator curves for the 'statistical', pragmatic and Rotterdam models.

Prediction models as classifiers of high and low risk

I attempted to summarise the capacity of these models to act as classifiers of risk. That is, can they successfully allocate patients to high or low risk groups? Selecting practically useful cut-offs is somewhat speculative, but possibly 20% predicted mortality might be an appropriate minimum. We are already hampered as the maximum mortality predictions for any model was around 20%, so this limits the capacity at the outset. Therefore I selected 10% and 20% as possible cut-offs for high risk. The sensitivities, specificities, and incorrect allocations are summarised in Table 35. The ROC values in the table were from the identical dataset to that used for the calibration, but were generated in SPSS. There are slight differences in absolute AUC values generated by SPSS and by the val.prob.ci function, although they are close and in the same order. Classification was poor with sensitivity reaching a maximum of only 28% in the Rotterdam model and high numbers of false positives, making these models of no practical value for clinical decision making. The best performers were those based on fewest predictors (the ‘Pragmatic’ models), and the Rotterdam model.

Table 35 Sensitivity, specificity, true/false positive & negatives given predicted mortality cut-offs of 10% and 20%

					Cut 10%					Cut 20%				
Model	cases	missing	deaths	c statistic	TP	FP	FN	sens	spec	TP	FP	sens	FN	spec
Clinical	635	150	29	0.603[0.49,0.715]	3	63	26	0.045	0.95	0	4	0	29	0.99
Clinical(s)	635	150	29	0.603[0.49,0.715]	0	14	29	0	0.95	0	0	0	29	1
Clinical(I)	635	150	29	0.612[0.519,0.741]	5	79	24	0.06	0.95	1	1	0.04	28	0.99
Statistical	561	224	33	0.620[0.513,0.727]	3	24	30	0.09	0.95	0	0	0	33	1
Pragmatic	785	0	43	0.664[0.55,0.777]	9	84	34	0.20	0.88	0	1	0	43	0.99
Clinical(INT)	635	150	29	0.626[0.527,0.726]	5	57	24	0.08	0.96	2	7	0.07	27	0.98
Prag(INT)	785	0	43	0.637[0.533,0.740]	7	59	36	0.16	0.92	0	6	0	43	0.99
Rotterdam	785	0	43	0.721[0.608,0.834]	9	23	34	0.28	0.96	0	0	0	43	0.95

Key: TP, true positive; FP, false positive; FN, false negative; sens, sensitivity; spec, specificity

4.4 Discussion

Ideally this study would have been a straightforward problem of estimation, whereby an established model(s) or known predictors and their weights could have been validated on our dataset. However, the lack of a consensus on predictor selection, and with only one potentially applicable existing prediction model, the 'Rotterdam' model (Steyerberg *et al.*, 2006), model exploration and selection became central to the study.

4.4.1 Selection of prediction models

I pre-specified a main 'Clinical' model and its predictors to try and reduce overfitting and subsequent overoptimism, which can occur when the model is matched to the modelling data and its idiosyncrasies too closely (Steyerberg, 2009k). The result is that small and large predicted values are exaggerated (Harrell, 2001c). The use of imputed datasets increased sample size, reduced regression coefficients, and improved the calibration slope. The same was true in applying shrinkage measures to the Clinical model, but at a cost of reducing the prediction range to very close to the overall mean mortality.

A stepwise elimination model confirmed the importance of age and operation as a predictor and also identified weight loss and gender as potentially important. As a measure of nutritional state one would expect weight loss to be predictive. However, there was no evidence to support the inclusion of gender from the systematic review. But it appeared as a significant predictor with men having twice the mortality of women. Disease incidences vary between men and women e.g. coronary artery disease is higher in men (Gabriel *et al.*, 2009) but the reported effects on perioperative mortality vary (Hayashida *et al.*, 2012; LaPar *et al.*, 2012). In our sample about 70% of the sample was male and there were small between gender differences in tumour histology and non-smoking rate, but these were not predictors of outcome. This could be a random finding or there may be an unmeasured predictor distributed unevenly between genders.

The use of stepwise elimination also highlighted the problems of 'data driven' models, missing data and small samples. The exaggerated effect of gender resulted from missing data on another predictor, which was also associated with mortality, and the statistical model based on this data performed poorly in validation.

The most important predictor was age, which was supported by published evidence from the systematic review, was significant in univariate and all the prediction models (Table 23). This is the most reliable predictor we have and should be adjusted for in studies of perioperative mortality.

The Revised Cardiac Risk Index (RCRI) (Lee *et al.*, 1999) was associated with mortality in univariate analysis but became non-significant in the 'Clinical' model. The effect of cardiac morbidity is likely to be complex because its incidence increases with age (Fleisher *et al.*, 2007) and aging also produces cardiovascular changes which could magnify the effect of cardiac disease in the elderly (Priebe, 2000). Consequently there could be interaction and confounding with age. In this study the RCRI was correlated with age. However, the prevalence of the high scores was very low, possibly accounting for its overall weak effect (Figure 5). The odds ratio from the imputed datasets of 1.084 for RCRI gave plausible predicted mortalities. For instance, predicted mortality was 2% with an RCRI score of zero but increased to 3% at high risk (RCRI 5) for patients in their forties, and from 12% to 17% in their eighties, suggesting a possible interaction with age. The general size of these effects seems plausible and the inclusion of some estimate of stratified cardiac morbidity would seem worth investigating. The frequencies of the highest and potentially most important scores were also very small and therefore the categorical groupings of counts may be more appropriate. A recent systematic review (Ford *et al.*, 2010) confirmed that a two level RCRI can distinguish moderately between a low risk (0 or one risk factor) and high risk (more than two factors) for cardiac events but not for all cause mortality.

Evidence for including serum albumen and spirometry was equivocal and neither was statistically significant in this study; substituting values within plausible ranges into the derived logistic equations produced negligible changes in the mortality rate.

4.4.2 Performance of clinical prediction models

The use of imputed datasets increased sample size for the development of the ‘Clinical’ models. This coincided with a general reduction in size of coefficients to possibly more realistic values in development, and improved discrimination and calibration in validation samples. Other models based on complete datasets (the ‘Pragmatic’ and Rotterdam models) and larger event to predictor ratios (the ‘Pragmatic’ model) also showed better performance, so that the ‘Pragmatic’ model was the best of those developed on the NOGCU database. Overestimation of mortality at higher predictions was common to all models to some degree, as it was in validated models from the systematic review.

The ‘Pragmatic’ model with only age, gender and operation calibrated fairly well up to about 15% predicted mortality and could be considered for use as a risk adjuster, for instance in provider profiling. However, overall the models performed poorly in discrimination and classification. The Rotterdam model had the best AUC of 0.721 and sensitivity of 28%, which is not of practical value. The best performing NOGCU model was the ‘Pragmatic’ model with a sensitivity of about 20%. The narrow range of predictions reaching only about 20% confirms that these models are only capturing a relatively small amount of variance in the sample. Clearly, stronger predictors are required if a practically useful model is to be developed.

The Rotterdam model compared favourably with the other models for discrimination, and had the best sensitivity, but is not directly comparable because the comorbidities required recoding in a different way to the ‘Clinical’ models (Steyerberg *et al.*, 2006). Its comorbidity score is a composite score which increases for each system involved with disease.

Perhaps this simple way of stratifying an overall level of 'illness' may be capturing a general measure of 'frailty' as well as quantitative relationships with some individual measures of comorbidity. This model was unsurprisingly inaccurate overall as it was developed on a different population, using different scoring, but perhaps recalibration on a larger dataset could be possible.

4.4.3 Strengths and weaknesses of this study

Data quality

I endeavoured to verify the reliability of the predictors and outcome before attempting to construct prediction models. However, ultimately the database sample was a "convenience" sample (Harrell, 2001b) because although the data was collected before the outcome was known, it was not designed specifically for this study. Important predictors were not defined prospectively (e.g. cardiac and respiratory comorbidity) and considerable amounts of important data were missing (e.g. weight loss). The use of free text data entry without reference to prior definition left considerable scope for subjective interpretation, and therefore information bias. I consolidated and recoded predictors, which were represented by several fields. I also attempted to define the new predictor categories prospectively (e.g. the revised cardiac risk index) before exploring their relationships with the outcome to avoid overfitting. However, since there was inevitably subjective interpretation during recoding, there will have been some scope for misinformation bias.

Sample size

Sample size in studies with a binary outcome is driven by the number of outcome events, and it has been suggested that adequate samples should have an event per variable (EPV) ratio of at least ten (or even twenty). A low EPV ratio is a cause of bias in estimating regression coefficients and variances, and can cause overoptimistic statistical significance and effect sizes. This is because the model may closely represent a relatively small set of data points, with a large variance. This is particularly so in stepwise elimination models (Peduzzi *et al.*, 1996; Steyerberg *et al.*, 1999). The result is 'overoptimism' in predictive capability in a new sample, and a tendency to exaggerate high predictions and underestimate low predictions.

Recommendations to minimise such problems in small datasets include selecting a small set of predictors from prior knowledge rather than from 'data driven' methods and minimising the inclusion of interactions (Steyerberg *et al.*, 2001a). This modelling dataset was relatively small (43 outcome events) and therefore at risk of overfitting and 'overoptimism'. Consequently I had little scope to investigate main effects and their interactions. Clinical knowledge suggests that age may alter the effects of other predictors e.g. cardiac morbidity. Introducing an interaction between age and RCRI (p=0.06) produced plausible predictions especially in the very elderly, and I examined a model, which included this, but at only small overall improvement in model performance.

In this study, the 'Clinical' and 'Statistical' models all had an EPR less than 10 and their performance on discrimination and calibration was inferior to the models derived from complete datasets with an EPR greater than 10. This was particularly so for the 'Statistical' model as it is recommended that in stepwise models the 'degrees of freedom' include all possible included predictors, models and interactions. However, I used this mainly as a screening tool, added to the evidence that age and possibly weight loss, gender and operation should be predictive, in keeping with Clinical knowledge.

The use of imputed datasets to replace missing FEV1 and serum albumen values, increased development sample size and consequently, event to predictor ratio. The result was a general reduction in coefficient size with a modest improvement in performance, perhaps reflecting 'overfitting' in the original smaller sample.

Validation methods

I used a traditional split sample method for validation, using 50% for each of modelling and validation. The samples were generated randomly and in such a way to ensure that mortality was approximately equal in both samples. However, in these relatively small samples, it is possible for any predictor to split unevenly and affect the model performance. This is particularly so if the predictor distribution is skewed and a 'high risk' low prevalence value may not appear at all in one of the samples (Steyerberg, 2009o). This was observed to a degree with the distribution of RCRI scores, where the validation sample had almost all the high risk scores. Splitting relatively small samples also leads to increased variance and possibly less stable models and validation performance (Steyerberg, 2009o). Sample sizes containing up to 100 events may be necessary for reliable validation (Vergouwe *et al.*, 2005; Peek *et al.*, 2007), a figure far larger than was possible in this sample. A possible approach could be to use bootstrap methods which use the whole dataset. Distributions for parameters such as regression coefficients are derived by generating bootstrap samples, running the analysis on each sample, and obtaining the parameter of interest from each one. The samples are the same size as the original sample and each member of the sample is randomly selected for the bootstrap, but is replaced and could therefore appear more than once. These methods are described by Steyerberg (Steyerberg, 2009k) and were used in developing the Rotterdam model (Steyerberg *et al.*, 2006), although this did not prevent disappointing performance on new datasets during development or by other investigators (Zingg *et al.*, 2009).

4.5 Key findings

1. The models developed from the NOGCU database discriminated and classified 'high' and 'low' risk poorly. They could not be of practical value for this function. The best model was the Rotterdam model, but this was developed on a different dataset and required recoding in the NOGCU dataset.

2. All the models predicted mortality over a small range up to a maximum of about 20%. They all overestimated mortality to some degree, with the best performing being a 'Pragmatic' model with age, gender and operation, which

was developed on a complete dataset and predicted fairly reliably up to about 15% mortality.

3. Age was the most important predictor of outcome supported in the data analysis and also by evidence from the systematic review. Gender and operation should also be included in prediction models. Other predictors which should be considered include cardiorespiratory comorbidity and measures of loss of body mass and activity, but these require much larger studies to elucidate their role.

4. The models based on data driven methods, larger numbers of predictors and smaller datasets were the worst performers on validation. Multiple imputation based models reduced the sizes of odds ratios derived from the original data and improved calibration in validation.

5. These models are hampered by the 'low' prevalence of perioperative mortality and some of the high risk predictors. They do not capture enough of the variance in this data to make effective prediction models, and stronger predictors are needed.

6. This study was based on twenty years worth of data from a 'high volume' surgical centre. The sample sizes from single 'high volume' centres cannot form the basis for Clinical prediction models.

Chapter 5: General Discussion

5.1 A summary of the work in this thesis

In this project I have set out to develop a clinical prediction model for perioperative mortality after oesophagectomy, which could be applicable to patients cared for in the Northern Oesophagogastric Cancer Unit (NOGCU). I used clinical information, which has been recorded in the associated clinical database since 1990, and has not previously been studied for this purpose. Clinical prediction models have been developed on other datasets (Law *et al.*, 1994; Zhang *et al.*, 1994; Bartels *et al.*, 1998; Liu *et al.*, 2000; Bailey *et al.*, 2003; McCulloch *et al.*, 2003; Tekkis *et al.*, 2004; Sanz *et al.*, 2006; Steyerberg *et al.*, 2006; Ra *et al.*, 2008) so my first aim was to investigate whether any of these were 'transportable' to our practise. My next aim was to identify candidate predictors, which can be collected routinely preoperatively, and to use these to develop and validate a clinical prediction model from the NOGCU dataset.

I carried out a systematic review to identify and investigate published clinical prediction models and studies of potential candidate predictors of perioperative mortality. At the time of writing this thesis no relevant systematic review had been reported. I searched for published clinical prediction models and investigations of candidate predictors, which had been studied in 'high volume' surgical centres since 1990. Ten clinical prediction models were identified (Law *et al.*, 1994; Zhang *et al.*, 1994; Bartels *et al.*, 1998; Liu *et al.*, 2000; Bailey *et al.*, 2003; McCulloch *et al.*, 2003; Tekkis *et al.*, 2004; Sanz *et al.*, 2006; Ra *et al.*, 2008; Steyerberg, 2009a) but only the POSSUM model (Tekkis *et al.*, 2004), Bartel's 'Munich' model (Bartels *et al.*, 1998), the 'Rotterdam' model (Steyerberg *et al.*, 2006) and the 'Philadelphia model '(Ra *et al.*, 2008) have been externally validated. They generally overestimated mortality in the higher ranges, and discrimination was moderate at best, particularly on external validation samples. There was considerable variation in predictor definition, and the potential for bias was infrequently managed or reported. The only predictors which were broadly comparable to those available in the NOGCU database were from Steyerberg's 'Rotterdam' model (Steyerberg *et al.*, 2006). Two models had been incorporated into clinical practise (Zhang *et al.*, 1994; Bartels *et al.*, 1998) and reported to reduce operative mortality, but none were subjected to formal clinical impact studies.

Age was, by far, the most investigated candidate predictor. Most studies categorised patients as above or below 70 years and most concluded that there was no association with perioperative mortality. However the samples were usually too small to detect important differences and when combined in a data synthesis the risk of mortality was greater in those over 70 years (pooled odds ratio 1.91). Interpretation of this odds ratio remains circumspect as the incidence of extreme old age and potential confounders (e.g. cardiovascular disease) between groups and studies was not usually obvious. Of the other study designs about half found an association, and there was a suggestion that the effect of age was greater in octogenarians (Moskovitz *et al.*, 2006).

The interpretation of the importance of other candidate predictors from studies in the systematic review was hampered by considerable heterogeneity in predictor definition and no consensus on optimal predictor form. Differences in the reporting of results, in the sample case-mix and in the prevalence of potential bias added further difficulty to predictor selection. The evidence from the systematic review that other candidate predictors should be included was mixed, but supported by some studies for cardiac comorbidity (Mangano, 1990; Eagle *et al.*, 1997), respiratory disease (Bartels *et al.*, 1998; Abunasra *et al.*, 2005; Alexiou *et al.*, 2005; Alexiou *et al.*, 2006), nutritional and immunity based measures (Law *et al.*, 1973; Fekete and Belghiti, 1988; Windsor and Hill, 1988), other comorbidities (Griffin *et al.*, 2002; Bailey *et al.*, 2003) as well as surgical tumour factors (Abunasra *et al.*, 2005; Gockel *et al.*, 2005). I concluded that all the above candidate predictors should be considered, but this was not going to be a simple estimation problem and would require data exploration and predictor selection.

In the next stage of this project I prepared a subset of data from the NOGCU clinical database with which to develop a clinical prediction model. I extracted and cleaned data from fields, which I had identified as likely to represent candidate predictors for the model. Some fields (e.g. age, surgical details, dates, and outcome variables) were nearly complete and were verifiable from other fields within this and other databases. Mortality and survivor status were validated within the database against items such as stated date of death or outpatient follow up dates, and by the database manager against the Northern and Yorkshire Cancer registry and Information Service (Northern and Yorkshire Cancer Registry and Information Service)

Other potential predictors were difficult to use because they were represented by several different fields, often in free text form (e.g. cardiovascular and respiratory morbidity), or had significant quantities of missing or unreliable data (e.g. exercise testing variables, weight loss). Recoding and cleaning of this data involved considerable work and added the risk of information bias (Delgado-Rodriguez and Llorca, 2004). For instance, cardiac comorbidity was represented in seven different fields, some categorical, and some free text. I decided to recode these into a single predictor, the previously validated Revised Cardiac Risk Score (Lee *et al.*, 1999), but this clearly incurred considerable risk of interpretation bias, both from the original data entry and my recoding. This process was repeated for several other candidate predictors, for example respiratory comorbidity and surgical and tumour details.

Some fields had considerable amounts of missing data, for example some exercise testing variables (up to 40% missing) and weight loss (30%). Although I explored these variables, I did not feel confident in their validity and they were excluded from final models. Both spirometry and serum albumen were included in models and both had some missing data (<15%). I used multiple imputation methods to replace this data in order to study models, which optimised the available data.

I took a traditional approach to developing the clinical prediction model and split the data into two equal random samples, with similar mortality rates. One would be for modelling and one for internal validation. Firstly, I used logistic regression to develop a full 'Clinical' model containing predictors whose selections were supported by published information and clinical knowledge. This was to reduce the chance of overfitting, optimism and bias in coefficient estimation which might be caused by statistically significant chance associations from 'data driven' methods (Steyerberg, 2009k). For the full clinical model I selected age, RCRI, FEV1 and serum albumen. Surgical procedure was also included because this project focussed on oesophagectomy and the database also contained data on total gastrectomies and a heterogeneous mix of other lower risk procedures. I selected serum albumen because it was the most frequent predictive nutritional marker in the review, and FEV1 because it is central to the diagnosis of chronic obstructive pulmonary disease and appeared complete and reliable.

I also used univariate statistics and stepwise logistic regression in a 'data driven' analysis to compare with, and perhaps corroborate the 'Clinical' model. I screened a larger number of predictors, including the various different descriptors of comorbidities entered into the database, for example, the various free text and categorical entries for respiratory disease. Statistically significant predictors from univariate analysis included age, surgical procedure, RCRI, white cell count, weight loss and gender. The stepwise regression resulted in a strongly predictive final model containing age, surgical procedure, gender and weight loss. These findings go some way to justifying the selection of predictors for the 'Clinical' model.

Weight loss and gender had also emerged from 'data driven' methods as candidate predictors and both have rationale. However, I ultimately excluded weight loss from the final models, because of the extent of missing data. The missing weight loss data also resulted in a gross imbalance in gender specific mortality between missing and non-missing data, resulting in a spuriously large gender effect. Unsurprisingly, the model with weight loss performed badly on validation, and the effect of gender reduced considerably when weight loss was omitted. However, gender remained statistically significant. There were no strong differences between genders for other predictors in the data and there is some evidence suggesting that it should be considered (Gabriel *et al.*, 2009; Hayashida *et al.*, 2012; LaPar *et al.*, 2012) so I included it in a subsequent 'Clinical' model.

I attempted to increase the sample size, on which the 'Clinical' model was developed by using multiple imputation samples to replace missing serum albumen and FEV1 values. This resulted in an almost complete dataset. Resulting regression coefficients were generally reduced in magnitude, perhaps reflecting some overfitting in the relatively small original sample. I also used the imputation datasets to run a 'Clinical' model including gender, which also had a considerably reduced coefficient compared to the 'data driven' model. Of potential importance, the statistically significant univariate effect of the Revised Cardiac Risk Index was lost in the 'clinical' model containing age, as the two were moderately well correlated (r=0.5). The association between cardiac morbidity and age is not unexpected and shows the complex relationship between the two for a prediction model.

On internal validation the models gave predictions over a fairly narrow range not far from the sample mean mortality, with the value for the highest decile mean being between 10 and 15%. Calibration was best in the models based on most complete datasets and fewest, most reliable predictors, i.e. age, gender and operation. This finding was also reflected in the 'Clinical' model based on imputation to maximise dataset size, when compared with the model based on the original data. Overestimation of mortality rates occurred in models containing RCRI, FEV1 and serum albumen, although general shrinkage improved this at the expense of reducing maximum predictions to nearer the overall mean. Discrimination was, at best, moderate in all models with maximum area under curve values of 0.65 for ROCs. Unsurprisingly the 'data driven' models, which contained weight loss, fared even worse. The 'Rotterdam' model (Steyerberg *et al.*, 2006) was poorly calibrated and failed to discriminate in this dataset; this was not unexpected as the data required some recoding to adapt for the Rotterdam scoring system.

5.2 The potential applications of clinical prediction models

The aim of a clinical prediction model should be to improve clinical care. This could include providing patients and their clinical teams with estimated individual 'risk specific' mortality rates to guide treatment choice or identifying poorly performing provider centres. It could also include classifying patients into high or low risk groups for allocation to research interventions or further diagnostic risk stratification. I discuss some of these issues in the following paragraphs.

5.2.1 Guiding choice of treatment

In July 2010, and subsequently in 2012 the government published their vision for the NHS, which included putting patient choice of both treatment and provider at the centre of healthcare (Secretary of State for Health, 2010; Department of Health, 2012). They embraced the phrase 'nothing about me without me' as a central part of 'shared decision making', a concept which ensures patients can take an active role in clinical decisions. The phrase 'nothing about me without me' was adopted as a guiding principle at the Salzburg Seminar of 1998, an event which was 'founded in 1947 in the spirit of post war reconciliation to provide a forum to challenge and debate a variety of issues and beliefs'(Delbanco *et al.*, 2001). With the backdrop of great changes within healthcare systems worldwide, the 1998 seminar challenged a wide range of healthcare and patient representatives to plan an ideal and utopian healthcare from scratch. Central to their ideas was the use of computers to provide the information, which could enhance patient choice (Delbanco *et al.*, 2001).

Treatment of oesophageal cancer may include options, which incur considerable risk and lack of certainty about benefit both for survival and quality of life. The current development of less invasive therapies, e.g. endoscopic resection, photodynamic therapy or thermal ablation (Allum *et al.*, 2011) may magnify the importance of information about risk in treatment choice. Decision aids, which present such information, may be helpful in enabling patients to make informed decisions. There is some evidence that decision aids not only improve knowledge of treatment options, and influence expectations of risks and benefits, but may also increase selection of more conservative treatments in place of more major surgery (O'Connor *et al.*, 1999; Stacey *et al.*, 2011). This has been reported when the treatment options include potentially curative major surgery, for instance breast and prostate cancer (Auvinen *et al.*, 2004; Armstrong *et al.*, 2005; Waljee *et al.*, 2007) and coronary vessel revascularisation (Morgan *et al.*, 2000). Similar findings have been reported where there is a choice between non-surgical treatments with serious risks (Brundage *et al.*, 2001).

For oesophageal cancer there might be a choice between high risk curative surgery with an impaired quality of life, a prolonged recovery period but potential curative outcome, and an alternative less invasive palliative procedure with a better quality of life but shorter survival. If perioperative or medium term mortality were considered to be an important factor in decision making, one might think the relative size and certainty of the mortality estimates would be important, although I can find no studies to confirm this. The models from the NOGCU using only age, operation and gender, looked to be reasonably reliable for mortalities up to 15%. However from my clinical experience I believe it unlikely that estimates of this magnitude compared with an overall mean of 5% would affect an individual’s decision to have surgery, where the alternative may be certain non-survival. To provide useful information for clinical decision making there is a need to develop and validate models which can provide a greater range of mortality predictions with reasonable confidence.

5.2.2 Provider profiling

Although it has been reported that choice of treatment may be more important to patients than choice of location or provider (Coulter, 2010), selecting a provider may be a reality for regionally provided specialist services. Knowing that mortality for surgery in a particular patient is 10% as opposed to 5% may not necessarily affect their choice of treatment, but could be central to appraising and choosing a provider. Perioperative mortality represents only one aspect of the quality of care (Lilford *et al.*, 2004) in high risk surgery such as oesophagectomy but it is always likely to attract attention (Shahian *et al.*, 2001). The government have expressly stated the provision of outcome information as one of their aims (Secretary of State for Health, 2010; Department of Health, 2012), therefore presumably this type of information will become increasingly important. An example of this has been the national clinical audit database of the Society for Cardiothoracic Surgery in Great Britain & Ireland, which records a substantial proportion of all cardiac procedures performed in the UK and Ireland. Individual provider performance is published at http://heartsurgery.cqc.org.uk (Care Quality Commission Society for Cardiothoracic Surgery in Great Britain & Ireland, 2010), and the Society have developed a methodology for identifying and managing apparent divergences of observed from expected mortality rates in providers. The statistical methods of provider profiling have also been illustrated on a dataset from the Scottish Audit of Gastro-Oesophageal Cancer Services (Collins *et al.*, 2011). However, there are considerable methodological problems with provider profiling (Shahian, Normand et al. 2001) which include the effect of varying case mix between centres, when patient prognostic predictors become potential confounders (Steyerberg 2009). These can then bias the outcome measure for the clinical centre of interest (Julious and Mullee 2000). Risk adjustment scoring (the modified EuroSCORE) is central to the appropriate identification of outlying performance by the Cardiothoracic Society, to allow for variation in case-mix and patient risk factors (Roques *et al.*, 1999). It would seem reasonable that the same should apply when comparing outcomes in oesophageal cancer surgery.

5.2.3 Controlling for prognostic predictors in research and diagnosis

An imbalance of known and unknown predictors between treatment groups may bias the results of interventional treatment studies. Adjusting for this imbalance by using clinical prediction models to select patients for trials, or to balance risk strata in treatment arms may reduce this bias, especially in small sample studies (Assmann *et al.*, 2000; Steyerberg, 2009c). Covariate adjustment may reduce bias, improve precision and increase the statistical power to detect a treatment effect (Hernández *et al.*, 2004; Steyerberg, 2009c). A potential area of application of this principle for oesophagectomy could be in trials to assess perioperative 'goal directed therapy'. This is a particularly topical and debated therapy based on the idea that achieving specific targets for organ oxygen delivery should improve outcome. It might be expected that any tangible benefit would be more likely or greater in 'higher risk' patients, and therefore risk stratification could be an important aspect of such a study.

Another topical and debated subject is the use of cardiopulmonary exercise testing (CPX) to stratify risk for major surgery (Older *et al.*, 1999; Forshaw *et al.*, 2008). This is based on measuring the capacity to increase oxygen delivery in response to an increased oxygen demand induced by exercise, and is assumed to partially mirror the physiological stress caused by major surgery. This seems a rational idea as it provides an individualised response to a physiological stress; however it is labour and cost intensive as it requires a sophisticated bicycle ergometer and its clinical impact is unclear. As with any other prediction model it seems reasonable that it should incorporate known important predictors and demonstrate that it can add value to models which include simple, reliable and relatively inexpensive data, such as age.

CPX has mainly been utilised as a classifier with a cut-off value to denote high and low risk groups. If the subject cannot maintain aerobic respiration above about 11 ml/Kg/minute of oxygen consumption, they would be considered to be at high risk.

The new post-test probability depends on the outcome prevalence and on test sensitivity and specificity (Sackett *et al.*, 1991). Reported sensitivity of CPX has been low, e.g. in Older's study of patients undergoing major surgery, the sensitivity was 60% and specificity was 70% (Older *et al.*, 1999).

I have used the data from this study to illustrate its potential utility in the plot on the following page (Figure 13).The plot was generated from the CEBM Statistics Calculator (Center for Evidence Based Medicine, 2012). The x axis represents the pre-test mortality and the y axis represents the new prediction given a positive CPX test indicating high risk. (Older *et al.*, 1999). A positive CPX test with a pre-test mortality of 5% (typical overall mortality for many centres) results in a post-test probability of around 10%. This level of information is unlikely to be practically useful. A higher pre-test probability of mortality nearer 15 or 20% results in a post-test probability of over 30%, a potentially much more useful estimate. This illustrates the difficulty of predicting mortality using tests with low sensitivity. It also illustrates the potential importance of identifying higher risk patients for such stratification testing, and for generating practically useful information.

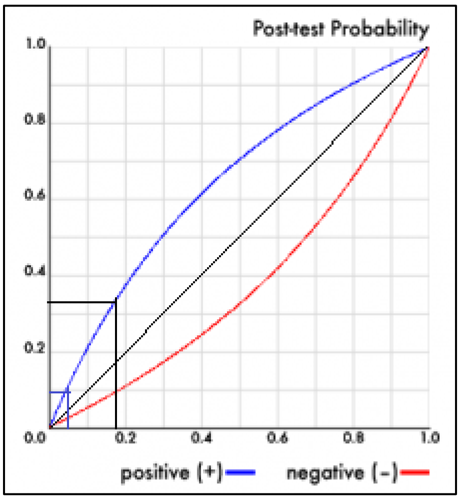


Figure 13 The effect of baseline prevalence on post-test probability given a test sensitivity of 60% (e.g. CPX testing). The x axis is pre-test prevalence and y axis is the post-test probability. See text for explanation. This plot is adapted from the CEBM Statistical calculator (Older et al., 1999).

5.3 What is the current status of the prediction model developed from the NOGCU database?

I could find no research to suggest what level of perioperative mortality might sway a patient's choice away from surgery but I speculated that it might need to be at least 20%, given the alternative outcome of unlikely survival without surgery. These models cannot currently provide estimates beyond about 15% for predicted mortality, so this model is probably not useful in this role. Similarly, if the model is used to classify patients as high or low risk for perioperative mortality using a specified cut-off, sensitivity was poor and the false positive rate high. The high specificity meant that low risk predictions would be mainly correct but little or no better than knowing the overall mortality alone. I originally set out with the intention of developing a clinical prediction model that could provide individual patients with enough information about the fatal risks of surgery to help them balance risks and benefits and an informed choice about treatment. The Rotterdam model, which I adapted to our dataset, appeared to provide the best discrimination but was scored differently and derived on different populations. It was probably giving a better overall estimation of ill health as it represented a collection of comorbidities. Although its performance was not practically useful, further development of its application should perhaps be considered. Clearly none of these clinical prediction models are currently useful for guiding clinical decision making.

However, the simpler models seem well enough calibrated over a wide enough range to potentially adjust for provider profiling. For instance, it could be of considerable public interest if a provider was reporting a mortality rate of 10%, which is about twice the national average. A simple model with reliable predictors such as age and gender could possibly predict this level of mortality reliably, and should arguably be part of any system comparing centres or operators. Similarly, it seems reasonable that new risk stratification techniques, such as cardiopulmonary exercise testing (Older *et al.*, 1999; Forshaw *et al.*, 2008) should incorporate such information into their predictions, firstly to demonstrate in validation and impact studies that they can add useful information, and secondly to improve predictions.

5.4 What are the difficulties with clinical prediction models of perioperative mortality for oesophagectomy?

5.4.1 Sample data

Missing data reduces sample size and can cause bias and spurious chance associations between predictors and outcome (Steyerberg, 2009f). In the systematic review, the reporting of missing data was frequently poor and in this modelling study some predictors had considerable amounts of missing data or unreliable data. This necessitated the exclusion of some potentially important predictors (e.g. exercise tolerance). The inclusion of a predictor with considerable missing data (weight loss) also led to spurious and overoptimistic associations between both itself and mortality, and gender and mortality. In future prospective studies, data validity and the management of missing data should be central to study design.

5.4.2 Selection of predictors

Selection of predictors should be based on clinical knowledge and previously published evidence rather than, data driven methods, such as univariate associations with outcome or stepwise regression methods. This is to reduce random associations, overfitting and overoptimism in a model (Steyerberg *et al.*, 2001b). A key area of difficulty in this study was that, despite the large number of published studies, it was unclear which candidate predictors should be included, because of the wide variety of definitions used by investigators. It is clear from publications and our model that 'age' should be included in any prediction model of perioperative mortality after oesophagectomy. However, even this was studied in a large variety of categorical forms. Ideally age should be included as a continuous variable to avoid loss of information (Steyerberg *et al.*, 2001b), particularly the distribution of very old age.

Definitions and effects of cardiac morbidity were inconsistent across the studies in the systematic review of this thesis. I therefore used the validated Revised Cardiac Risk Index (Lee *et al.*, 1999), which was associated with mortality in this study, but was correlated with age (r=~0.5) and its effect was lost in the multivariate model. In sensitivity testing, altering the RCRI from zero to the maximum six in the predictor equation produced quite plausible results, for instance increasing predicted all cause mortality from about 5% to 17% in elderly patients. Most items in the RCRI are risk factors for, or indicators of past disease and therefore not especially strong predictors. The strongly predicting items (heart failure, unstable coronary syndromes, high total RCRI (Fleisher *et al.*, 2007)) have low prevalence and therefore because of the relatively small sample size we were struggling for predictive power. For instance, in the modelling sample the mortality for an RCRI of 5 in patients over 74 was 50%, and in the whole dataset for patients with unstable coronary syndromes 25%. Of course the cell count for these categories was tiny (2 and 4 respectively), and therefore these could be random findings. This is also true of larger databases such as ICNARC (Park *et al.*, 2009), where the incidence of preoperative severe heart disease was not statistically associated with a plausible 11.1% mortality; the incidence of this predictor was only 0.3%.

The addition of exercise or activity capacity to RCRI adds predictive strength (Fleisher *et al.*, 2007). A simple method of assessing this is to score a patient's best self reported activity capacity as multiples of resting metabolic energy use (or 'MET's), for instance climbing a flight of stairs might be 4 METS (Fleisher *et al.*, 2007). Less subjective measures include cardiopulmonary exercise testing, which included exercise induced electrocardiographic evidence of myocardial ischaemia in original studies (Older *et al.*, 1999). However at the NOGCU, we have only recently started recording METS for risk assessment, and cardiopulmonary testing has not been available.

The high respiratory morbidity and associated mortality which occurs after oesophagectomy, would suggest that preoperative respiratory comorbidity might be an important predictor. This was not the case in our study, in which I used chronic obstructive airways disease as the main predictor, defined from free text data entry and spirometry. The association between respiratory comorbidity and outcome was mixed in the primary studies in the systematic review, again hampered by a wide range of definitions.

Weight loss was also strongly predictive of perioperative mortality in our study. This might be expected from clinical knowledge, and a variety of other measures of nutritional and immune status were associated with mortality in studies from the systematic review. However, the modelling sample was plagued by missing and unreliable data for this predictor. The missing data produced an implausibly large random effect for gender, because of an imbalance in 'gender specific' mortality between missing and non-missing data. Unsurprisingly, although the statistical model which included weight loss was strongly predictive with an impressive area under the ROC, it performed very poorly on validation. I rejected models with weight loss, based on this dataset, but nutritional based predictors clearly have strong rationale and should be investigated further.

There is a need to agree and standardise a format for candidate predictors so that large scale studies can be carried out to definitively characterise their role, if any. Standardisation would also allow potential 'pooling' of data from different studies to optimise the use of available information, and to facilitate comparison between different models (Collins and Moons, 2012).

5.4.3 Predictor strength

For an effective prediction model strong individual predictors are required. The power of a predictor is related to its prevalence and correlation with outcome (Steyerberg, 2009l). The prevalence of some of the potentially important strongly associated predictors in this study was low, for instance high risk RCRI scores. The measure of association in logistic regression is the odds ratio, which tends to range from about 1 to 3 in medical prediction studies (Pepe, 2005; Steyerberg, 2009m); point estimates of odds ratios for predictors in our study were all less than three. The addition of other validated predictors could increase the range of point estimates for predicted mortality and enhance the information available to make treatment choices. Hence there is a need to clarify the potential impact of the predictors which we have studied here, as well as identifying and investigating additional other predictors, for example cardiopulmonary exercise testing.(Older *et al.*, 1999).

The other aspect of prediction is to classify patients into groups of survivors or non-survivors for the purposes of allocating diagnostic, treatment or research interventions. An overall summary of the classification capability can be summarised in the area under ROCs for comparisons, but for clinical application it is more useful to know the true and false positive and negative rates, as their importance is context sensitive. For instance, in this study the high false positive rate would exclude many people from surgery if used for that purpose, but could be acceptable if used to allocate patients to further risk stratification. I used a very simple method to examine classification, but statistical devices specifically designed to assess the net practical benefit of clinical prediction models are available (Vickers and Elkin, 2006; Collins and Moons, 2012).

In general, a good classifier requires a very large degree of association to be effective, for instance an odds ratio of more than 30 (Pepe, 2005). In medical prediction studies including this one, odds ratios for predictors are rarely over 3. This limits the level of achievable false positive and negative rates, which is reflected in the low ROC area, and results in poor classification (Figure 14).

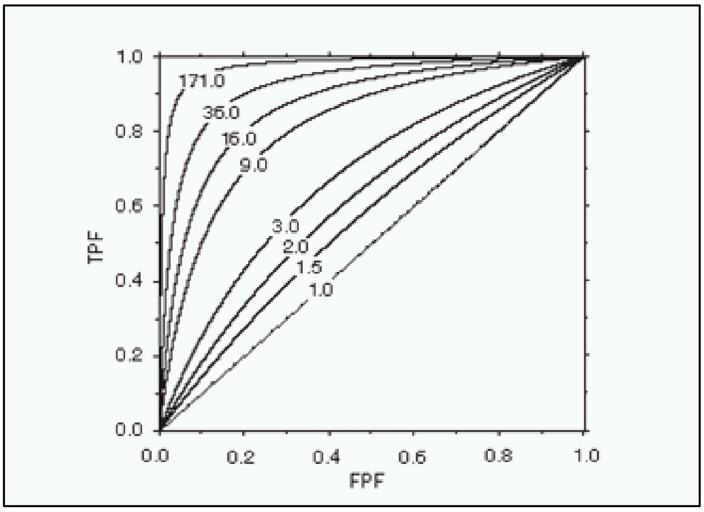


Figure 14 Correspondence between false positive and true positive fractions for different odds ratios (Pepe, 2005). Permission granted by John Wiley and Sons May 09, 2012.

5.4.4 Study methods and potential biases

Several potential biases or their inadequate reporting were identified in the systematic review and in our own modelling exercise. Of particular note were potential case selection bias, missing data, data validation and potential misinformation bias. This was an inevitable consequence of modelling data from a "convenience" sample (Harrell, 2001b). These are problems that should be considered prospectively in future model development.

5.4.5 Sample size

The important determinant of sample size in prediction models is the number of outcome events (Steyerberg, 2009n). Small sample sizes plagued many of the studies reported in the systematic review; my study had only about 40 events in each of the development and validation datasets. This is probably also compounded by the proportion of fatalities caused by surgical technical failure, which is not likely to be associated with comorbidity predictors. For instance, in an earlier audit I estimated that at least 10% of mortality was due to technical failure (Warnell, 2009(unpublished data)). This is similar to other reported data (Law *et al.*, 1994; Griffin *et al.*, 2002;

Law *et al.*, 2004; Abunasra *et al.*, 2005). The resulting small effective sample size probably contributes to model 'overoptimism' and lack of precision. For instance, if only age were in the model, the predicted mortality for an 80 year old was a plausible 9.4% but the 95% confidence interval was from 1.7% to 38%. In contrast the EuroSCORE (Roques *et al.*, 1999), which collected data from 19030 cardiac surgery patients with a similar overall mortality (about 910 events) to oesophagectomy, a matching predicted mortality for an 80 year old would be 7.1% with a 95% confidence interval of 6.6% to 7.6%; a much more informative range. These problems of sample size demonstrate the difficulty in generating useful prediction models from a single clinical centre. After all the NOGCU is a 'high' volume surgical centre, which has been collecting data for its well resourced clinical database since 1991.

5.5 Some potential solutions

The ideal solution would be to develop a clinical prediction model from scratch, in a prospective large sample investigation. This was done in the euroSCORE study for cardiac surgery, when the data from 19030 cardiothoracic procedures was collected from 128 European centres in a three month period in 1999 (Roques *et al.*, 1999). The euroSCORE has been validated and studied many times since then, and an updated and recalibrated version is currently used by The Society for Cardiothoracic Surgery in Great Britain & Ireland to risk-adjust for provider performance audit (The Society for Cardiothoracic Surgery in Great Britain & Ireland, 2011). A prospective study of this magnitude and resource can deliver risk estimates with practically useful precision.

For oesophagectomy, a project of similar scale to the euroSCORE project would take considerably longer, because each clinical unit is likely to do fewer oesophagectomies in three months than the approximately 120 cases submitted from each cardiothoracic centre. For instance, the National Oesophago-Gastric Cancer Audit 2010 (Cromwell *et al.*, 2010), which collected data from a similar number of patients and clinical centres took nearly two years. Clearly it would take considerable time and resources to complete a study of this magnitude and we should think about alternative ways to maximise benefit from limited resources and information.

A potential solution would be to incorporate currently available information into the development of our model (Steyerberg, 2009g) and then to prospectively validate, adjust or recalibrate it as necessary (Steyerberg *et al.*, 2004). Just as I have used clinical knowledge and published studies to select predictors, investigators have combined multiple external data sources in cancer survival studies to estimate important predictor effect sizes (Look *et al.*, 2002), and several statistical methods have been described to achieve this (Steyerberg *et al.*, 2000; Steyerberg, 2009g). One option is the synthesis of aggregate data summaries from primary studies, however in our systematic review the definition of predictors and the reporting of results varied considerably (e.g. the various different categorisations of age), making this a difficult and possibly unachievable task. A preferable option might be an 'individual patient data' systematic review (Steyerberg *et al.*, 2000; Riley *et al.*, 2010), incorporating data from individual patients of previously published primary studies and databases. As well as primary studies there are large databases, which could help to clarify the importance of predictors such as age, comorbidities and surgical details. These include the ICNARC database (Park *et al.*, 2009), and the National Oesophago-Gastric Cancer audits (Cromwell *et al.*, 2010). Hospital Episode Statistics, the administrative database of the NHS in England (The Health and Social Care Information Centre) is another potential source of much information, which has been used to generate hospital mortality prediction models (Aylin *et al.*, 2007) and was central to identifying high mortality rates in paediatric cardiac surgery at Bristol in the early 1990s (Aylin *et al.*, 1999). However, administrative databases have been reported to lack scope of information, data quality and the ability to adjust for case-mix and comorbidity (Mohammed and Andrew, 2007; Mohammed *et al.*, 2009). Large clinical multi-institutional databases focussed on a particular disease or group of procedures may provide the volume, scope and data quality suited to generate and validate clinical prediction models (Westaby *et al.*, 2007).

Clearly any such study would need prospective planning and would take more resources than a study using currently available publications of primary studies, but it could make maximum use of available information and be more reliable (Riley *et al.*, 2010). The results from aggregated data could then be used to inform a prospective multicentre project to validate and subsequently study the clinical impact (Wallace *et al.*, 2011) of a prediction model.

5.6 Conclusions

1. There is increasing momentum for the publication of information about surgical procedures to both aid treatment decisions by patients and to highlight variations in performance amongst providers. Clinical prediction models are central to adjusting for individual patient risk factors.

2. The clinical prediction models developed from the NOGCU clinical database, along with other published models, did not explain enough variation in the data to effectively discriminate between survivors and non-survivors after oesophagectomy, and therefore are unlikely to be useful as an aid to clinical decision making.

3. A simple model incorporating age, gender and operation calibrated well enough to a maximum prediction of about 15% to risk adjust for provider profiling or research.

4. Age is the most important predictor of perioperative mortality after oesophagectomy. There was some weak evidence to suggest that gender, cardiac morbidity and weight loss may add value. Despite the number of published studies, we do not know which other predictors should be included in a clinical prediction model of perioperative mortality after oesophagectomy.

5. In this study I encountered problems with missing data, undefined predictors and free text data entry. These potential sources of bias were poorly reported or not addressed in many of the studies in the systematic

review. Consideration should be given to addressing these issues prospectively in future studies.

6. Sample size was fairly small in this study. This probably contributed to overfitting in some models and limited scope for data exploration. This study sample is from a database with nearly twenty years of data from a 'high volume' centre. Single centres are unlikely to provide enough data to carry out research on clinical prediction for oesophagectomies.

7. Given the time and resources it would take to develop a clinical prediction model in a prospective study, consideration could be given to pooling individual patient data from a range of sources, studies or databases. High quality prospective validation and clinical impact studies could then be carried out. These are likely to require large scale studies which could be facilitated within a multicentre clinical database for upper gastrointestinal surgery.

Appendix A. Terms used for search strategy in electronic databases

Search concept	Concept definition	Search term	notes
Population	Adults		NOT children
	Oesophagus	exp esophagus	tw/mp (o)esophagus; (o)esophageal
	Cancer,Neoplasm, Tumour, Carcinoma	exp neoplasms (inc stomach and oesophageal); exp carcinoma (inc squamous); exp adenocarcinoma; carcinoma, squamous cell	tw/mp cancer; carcinoma; tumour; neoplasm
	Surgery, Oesophagectomy	exp surgical procedures, operative; exp esophagectomy;	Surgery, operative treatment, resection
Study design	Cohort, prospective, retrospective, case control	Exp epidemiologic study characteristics (inc case-control, cohort)	mp prospective, retrospective, observational, cohort, trial, randomised
	Randomised controlled trial	exp evaluation studies (inc clinical trials, reproducibility of results)	
	Database, clinical or administrative	Databases, factual; databases	mp (clinical) database
Clinical outcome	Mortality, including “all cause” mortality, 30 day postoperative mortality, “in hospital” mortality	exp mortality	Includes fatal outcome, hospital mortality, survival
	Morbidity	Not used because scope note definition: “The proportion of patients with a particular disease during a given year per given unit of population”.	

Search concept	Concept definition	Search term	notes
	Postoperative complications	Postoperative complications	Not exp; irrelevant subheadings
	Hospitalisation/length of stay/critical care/cardiac complications/respiratory complications	exp hospitalization (inc length of stay);exp critical care; heart disease/cardiac output, low/heart failure, congestive/myocardial ischemia/arrhythmia;exp respiratory tract diseases	mp 1. critical, intensive, care, therapy, length of stay, hospitalisat on;mp 2. complications, perioperative, postoperative, cardiovascular, cardiac, coronary, myocardial, heart, respiratory, chest
Risk assessment	Prediction, Assessment, Evaluation studies, Estimation, Stratification(maps to risk assessment, prognosis, statistics, exercise test), risk, score, index, Incidence(maps from epidemiological methods, morbidity)	exp prognosis, exp epidemiologic methods	
	cardiac risk assessment, cardiopulmonary exercise testing, physical functional capacity, stair climbing capacity, anaerobic threshold, cardiopulmonary exercise, cardiac risk stratification, electrocardiographic exercise stress testing, clinical database, regression, Bayes methods, computational intelligence, POSSUM, apache, severity if illness index, severity of illness index, karnovsky performance status all map to exp health status indicators, sickness impact profile, tool, instrumen	exp heart function tests, exp respiratory function tests, exp physical endurance, exp physical fitness, exp oxygen consumption, exp health status indicators, exp epidemiologic methods, exp computing methodologies	

Appendix B. Search filters

'Filter source'	Filter details
Guidelines for prognostic tests from NHS Centre for Reviews and Dissemination, York Universitywww.york.ac.uk/inst/crd/index_guidance.htm	Best single terms from 'effective MEDLINE searching strategies for studies of prognosis': exp epidemiologic studies Complex search with the highest sensitivity: incidence.sh. OR exp mortality OR follow-up studies.sh. OR prognos:.tw. OR predict: .tw. OR course:.tw. (exp denotes exploding the succeeding indexing term,":" truncation symbol in Ovid, sh denotes subject heading search, tw textword search)
PubMed Research Methodology Filters http://www.ncbi.nlm.nih.gov/sites/entrez?db=pubmed Accessed 2007	"prognosis" sensitive/broad search 90%/80% (incidence[MeSH:no exp] OR mortality[MeSH Terms] OR follow up studies[MeSH:no exp] OR prognos*[Text Word] OR predict*[Text Word] OR course*[Text Word]) "clinical prediction guides" sensitive/broad 96%/79% (predict*[tiab] OR predictive value of tests[mh] OR scor*[tiab] OR observ*[tiab] OR observer variation[mh])

Appendix C. Search strategy for electronic databases

1. exp Esophagus/

2. esophag$8.mp.

3. oesophag$8.mp.

4. 1 or 2 or 3

5. exp neoplasms/

6. exp carcinoma/

7. exp adenocarcinoma/

8. exp carcinoma, squamous cell/

9. (cancer$1 or carcinoma or tumo?r$1 or neoplasm$1).mp.

10. 5 or 6 or 7 or 8 or 9

11. 4 and 10

12. exp esophagectomy/

13. (esophagectomy or oesophagectomy).mp.

14. exp surgical procedures, operative/

15. (surg$4 or (surg$4 adj treatment) or (Ivor adj Lewis) or (surg$4 adj resection) or operat$4 or (operat$4 adj treatment) or (operat$4 adj resection)).mp.

16. 13 or 14 or 15

17. (11 and 16) or 12 or 13

18. exp epidemiologic study characteristics/

19. databases/

20. databases, factual/

21. (prospective or retrospective or observational or cohort or (clinical adj trial) or random$).mp.

22. clinical trial.mp. or clinical trial.pt. or random:.mp. or tu.xs.

23. exp evaluation studies/

24. "validation studies [publication type]"/

25. or/18-24

26. exp mortality/

27. postoperative complications/

28. exp hospitalization/

29. exp prognosis/

30. exp critical care/

31. (((critical or intensive) adj (care or therapy)) or (length adj stay) or hospitalis$ or hospitaliz$).mp.

32. exp respiratory tract diseases/

33. Heart Diseases/

34. cardiac output, low.mp. [mp=ti, ot, ab, nm, hw, sh, tn, dm, mf]

35. heart failure, congestive.mp.

36. myocardial ischemia.mp.

37. arrrythmia.mp.

38. (mortality or death or fatal$).mp.

39. (complications adj (post?operative or perioperative or cardiovascular or cardiac or coronary or myocardial or heart or respiratory or chest)).mp.

40. or/26-39

41. exp heart function tests/

42. exp respiratory function tests/

43. exp physical endurance/

44. exp physical fitness/

45. exp oxygen consumption/

46. exp health status indicators/

47. exp epidemiologic methods/

48. exp prognosis/

49. exp computing methodologies/

50. exp diagnostic errors

51. (reproducib$ or reliab$ or evaluat$ or predict$ or accuracy or precision or calibration or diagnostic or specificity or sensitivity or performance).mp.

52. ((health adj (status or indicator)) or database$ or comput$ or bayes$ or regression or (artificial adj intelligence) or (neural adj network) or (severity adj illness) or apache or karnovsky or (anaerobic adj threshold) or exercise or possum or o-possum or p-possum or (stress adj test) or (function$ adj3 capacity) or cardiac or respiratory or function$ or cpx or electrocardiograph$ or cardiograph$ or ecg or ekg or cardiopulmonary).mp. [mp=ti, ot, ab, nm, hw, sh, tn, dm, mf]

53. (risk$ or assess$ or estimat$ or stratif$ or evaluat$ or scor$ or index or predict$ or prognos$ or course).mp. [mp=ti, ot, ab, nm, hw, sh, tn, dm, mf]

54. or/41-51

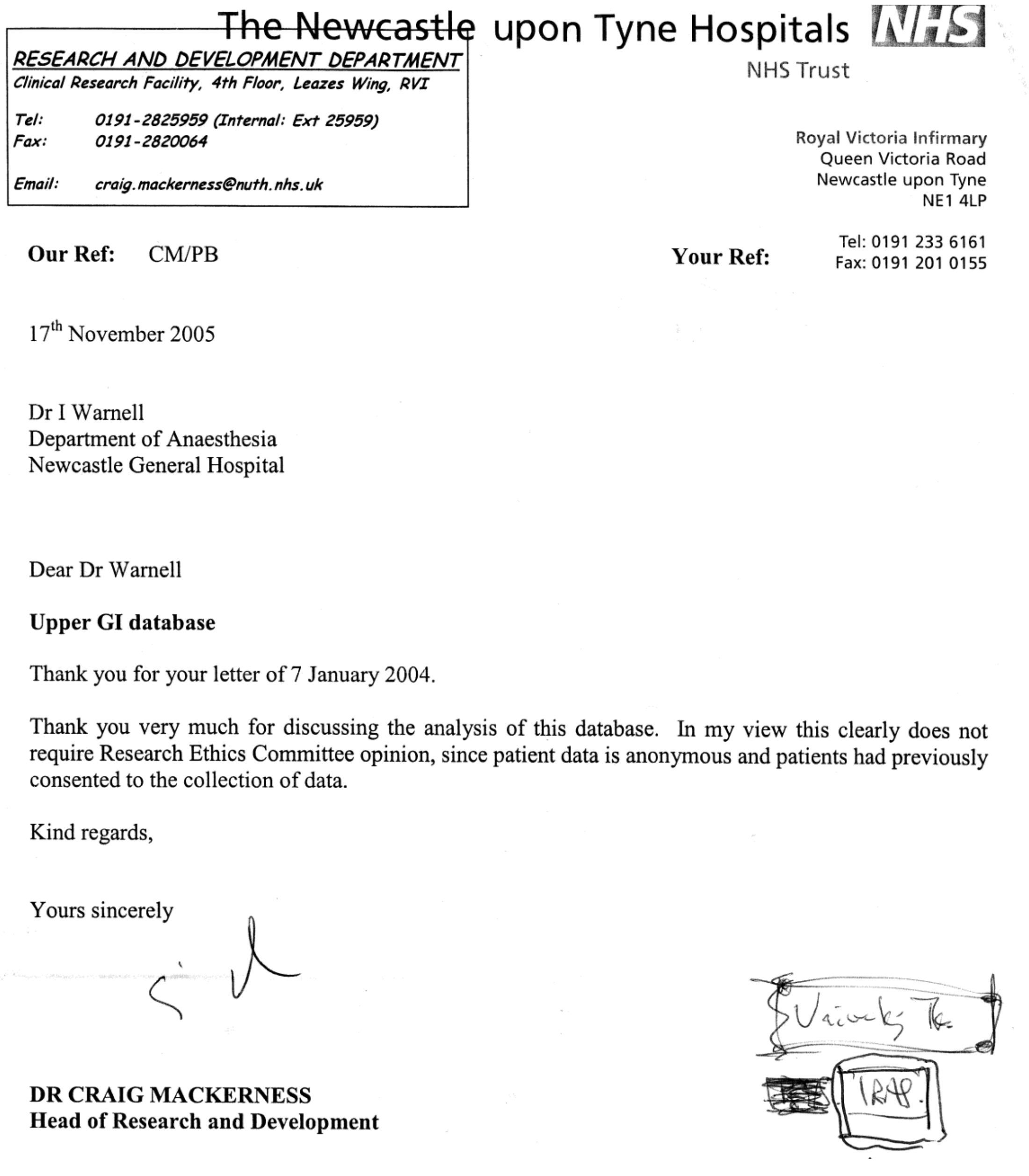
55. 17 and 25 and (40 or 54)

56. limit 55 to (english language and yr="1990-2009")

57. limit 56 to humans

Appendix D. Ethics and data protection

Facsimiles of the ethics and data protection documents are shown below



~~The Newcastle~~ upon Tyne Hospitals NHS
NHS Trust

RESEARCH AND DEVELOPMENT DEPARTMENT
Clinical Research Facility, 4th Floor, Leazes Wing, RVI

Tel: 0191-2825959 (Internal: Ext 25959)
Fax: 0191-2820064

Email: craig.mackerness@nuth.nhs.uk

Royal Victoria Infirmary
Queen Victoria Road
Newcastle upon Tyne
NE1 4LP

Tel: 0191 233 6161
Fax: 0191 201 0155

Our Ref: CM/PB **Your Ref:**

17th November 2005

Dr I Warnell
Department of Anaesthesia
Newcastle General Hospital

Dear Dr Warnell

Upper GI database

Thank you for your letter of 7 January 2004.

Thank you very much for discussing the analysis of this database. In my view this clearly does not require Research Ethics Committee opinion, since patient data is anonymous and patients had previously consented to the collection of data.

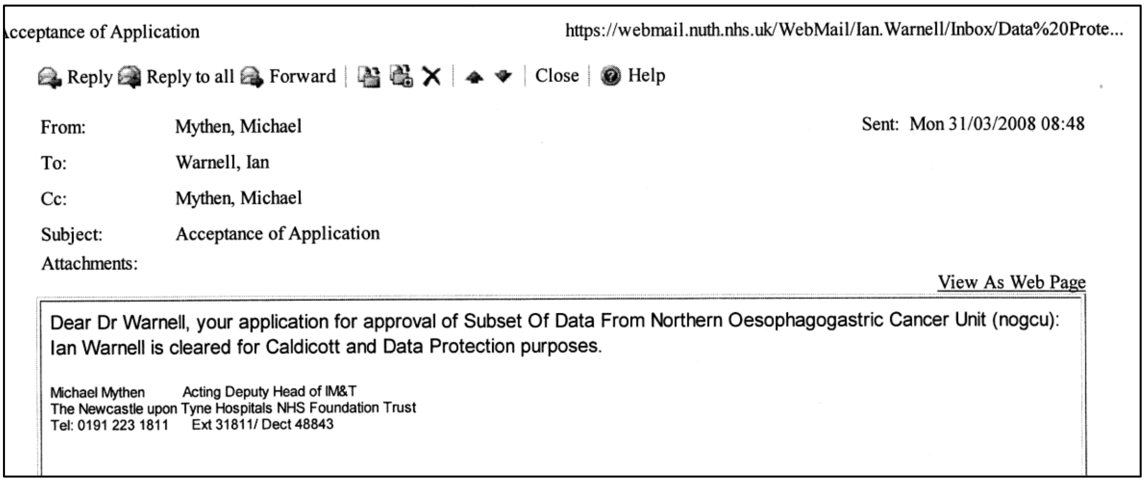
Kind regards,

Yours sincerely

Various Tks

IRAP.

DR CRAIG MACKERNESS
Head of Research and Development

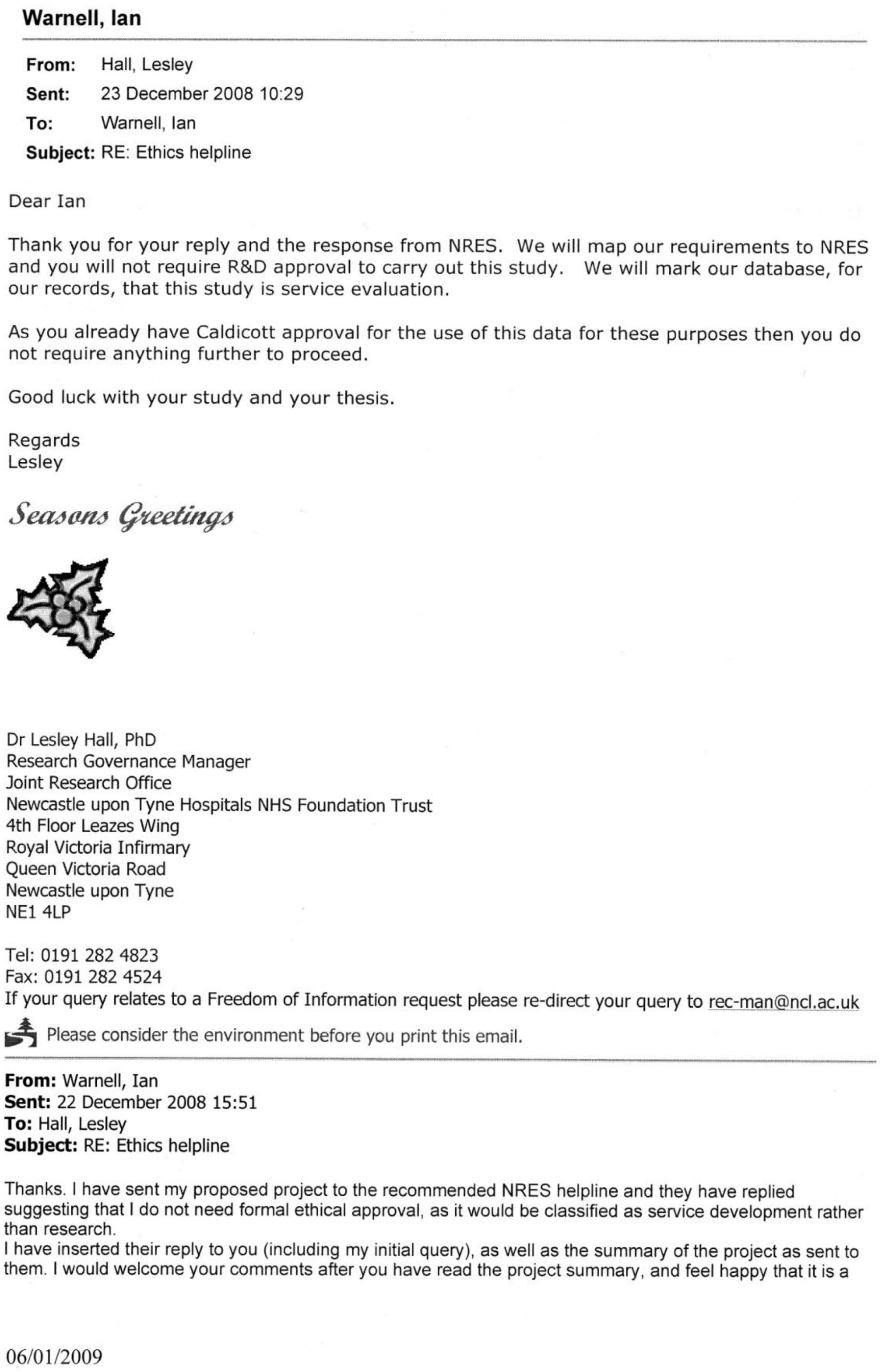


cceptance of Application https://webmail.nuth.nhs.uk/WebMail/Ian.Warnell/Inbox/Data%20Prote...

From: Mythen, Michael
To: Warnell, Ian
Cc: Mythen, Michael
Subject: Acceptance of Application
Attachments:
Sent: Mon 31/03/2008 08:48

Dear Dr Warnell, your application for approval of Subset Of Data From Northern Oesophagogastric Cancer Unit (nogcu): Ian Warnell is cleared for Caldicott and Data Protection purposes.

Michael Mythen Acting Deputy Head of IM&T
The Newcastle upon Tyne Hospitals NHS Foundation Trust
Tel: 0191 223 1811 Ext 31811/ Dect 48843



Warnell, Ian

From: Hall, Lesley
Sent: 23 December 2008 10:29
To: Warnell, Ian
Subject: RE: Ethics helpline

Dear Ian

Thank you for your reply and the response from NRES. We will map our requirements to NRES and you will not require R&D approval to carry out this study. We will mark our database, for our records, that this study is service evaluation.

As you already have Caldicott approval for the use of this data for these purposes then you do not require anything further to proceed.

Good luck with your study and your thesis.

Regards
Lesley

Seasons Greetings

Dr Lesley Hall, PhD
Research Governance Manager
Joint Research Office
Newcastle upon Tyne Hospitals NHS Foundation Trust
4th Floor Leazes Wing
Royal Victoria Infirmary
Queen Victoria Road
Newcastle upon Tyne
NE1 4LP

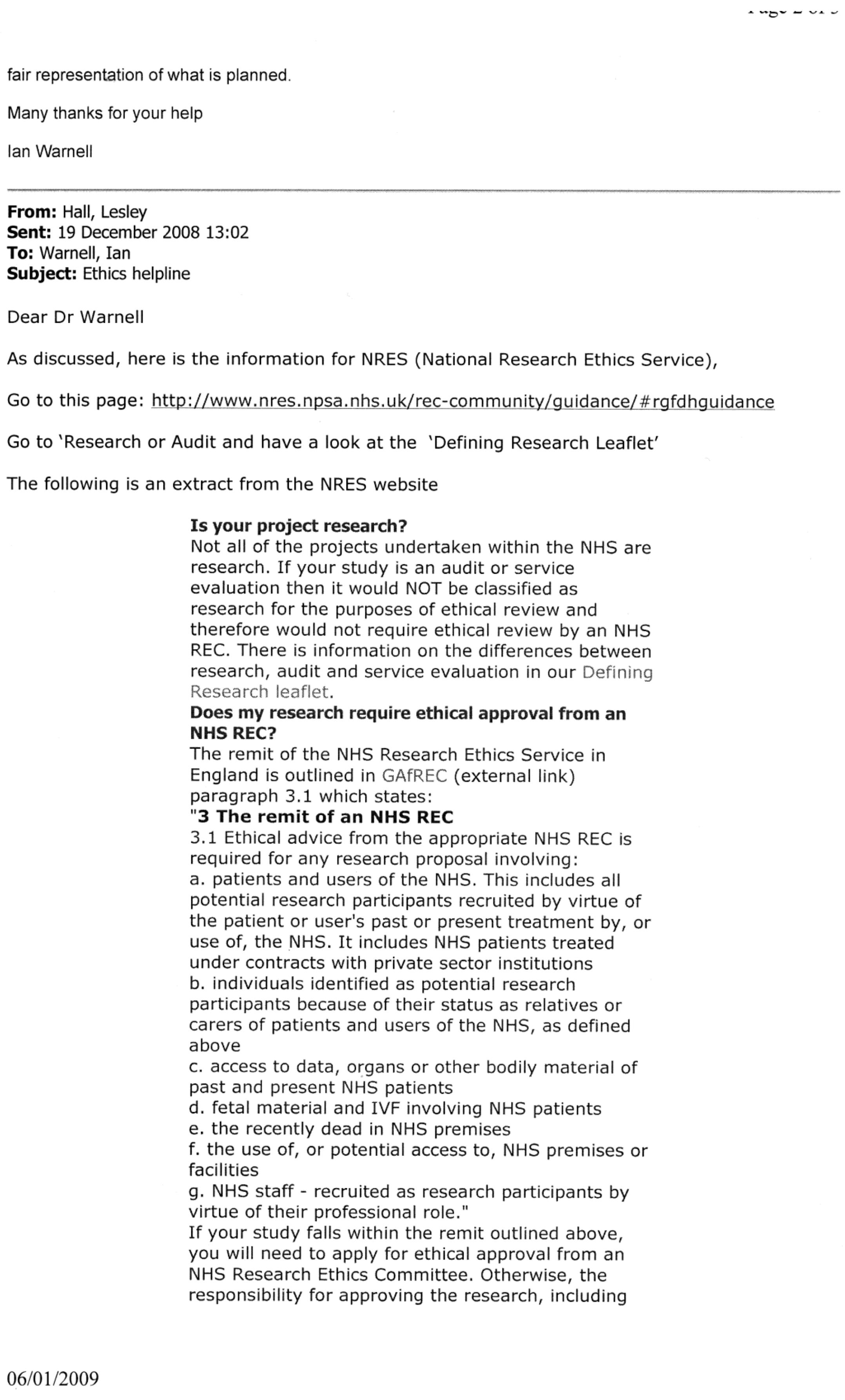
Tel: 0191 282 4823
Fax: 0191 282 4524
If your query relates to a Freedom of Information request please re-direct your query to rec-man@ncl.ac.uk

Please consider the environment before you print this email.

From: Warnell, Ian
Sent: 22 December 2008 15:51
To: Hall, Lesley
Subject: RE: Ethics helpline

Thanks. I have sent my proposed project to the recommended NRES helpline and they have replied suggesting that I do not need formal ethical approval, as it would be classified as service development rather than research.
I have inserted their reply to you (including my initial query), as well as the summary of the project as sent to them. I would welcome your comments after you have read the project summary, and feel happy that it is a

06/01/2009



fair representation of what is planned.

Many thanks for your help

Ian Warnell

From: Hall, Lesley
Sent: 19 December 2008 13:02
To: Warnell, Ian
Subject: Ethics helpline

Dear Dr Warnell

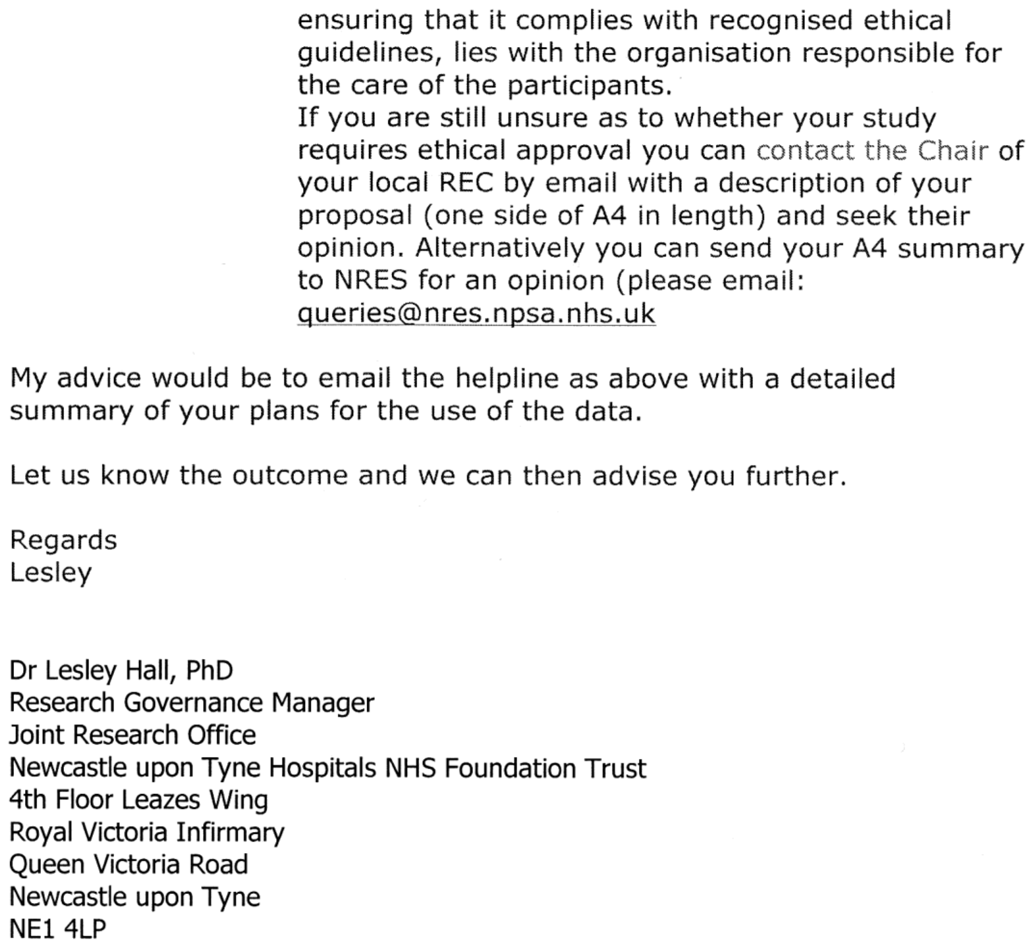
As discussed, here is the information for NRES (National Research Ethics Service),

Go to this page: http://www.nres.npsa.nhs.uk/rec-community/guidance/#rgfdhguidance

Go to 'Research or Audit and have a look at the 'Defining Research Leaflet'

The following is an extract from the NRES website

> **Is your project research?**
> Not all of the projects undertaken within the NHS are research. If your study is an audit or service evaluation then it would NOT be classified as research for the purposes of ethical review and therefore would not require ethical review by an NHS REC. There is information on the differences between research, audit and service evaluation in our Defining Research leaflet.
> **Does my research require ethical approval from an NHS REC?**
> The remit of the NHS Research Ethics Service in England is outlined in GAfREC (external link) paragraph 3.1 which states:
> "**3 The remit of an NHS REC**
> 3.1 Ethical advice from the appropriate NHS REC is required for any research proposal involving:
> a. patients and users of the NHS. This includes all potential research participants recruited by virtue of the patient or user's past or present treatment by, or use of, the NHS. It includes NHS patients treated under contracts with private sector institutions
> b. individuals identified as potential research participants because of their status as relatives or carers of patients and users of the NHS, as defined above
> c. access to data, organs or other bodily material of past and present NHS patients
> d. fetal material and IVF involving NHS patients
> e. the recently dead in NHS premises
> f. the use of, or potential access to, NHS premises or facilities
> g. NHS staff - recruited as research participants by virtue of their professional role."
> If your study falls within the remit outlined above, you will need to apply for ethical approval from an NHS Research Ethics Committee. Otherwise, the responsibility for approving the research, including



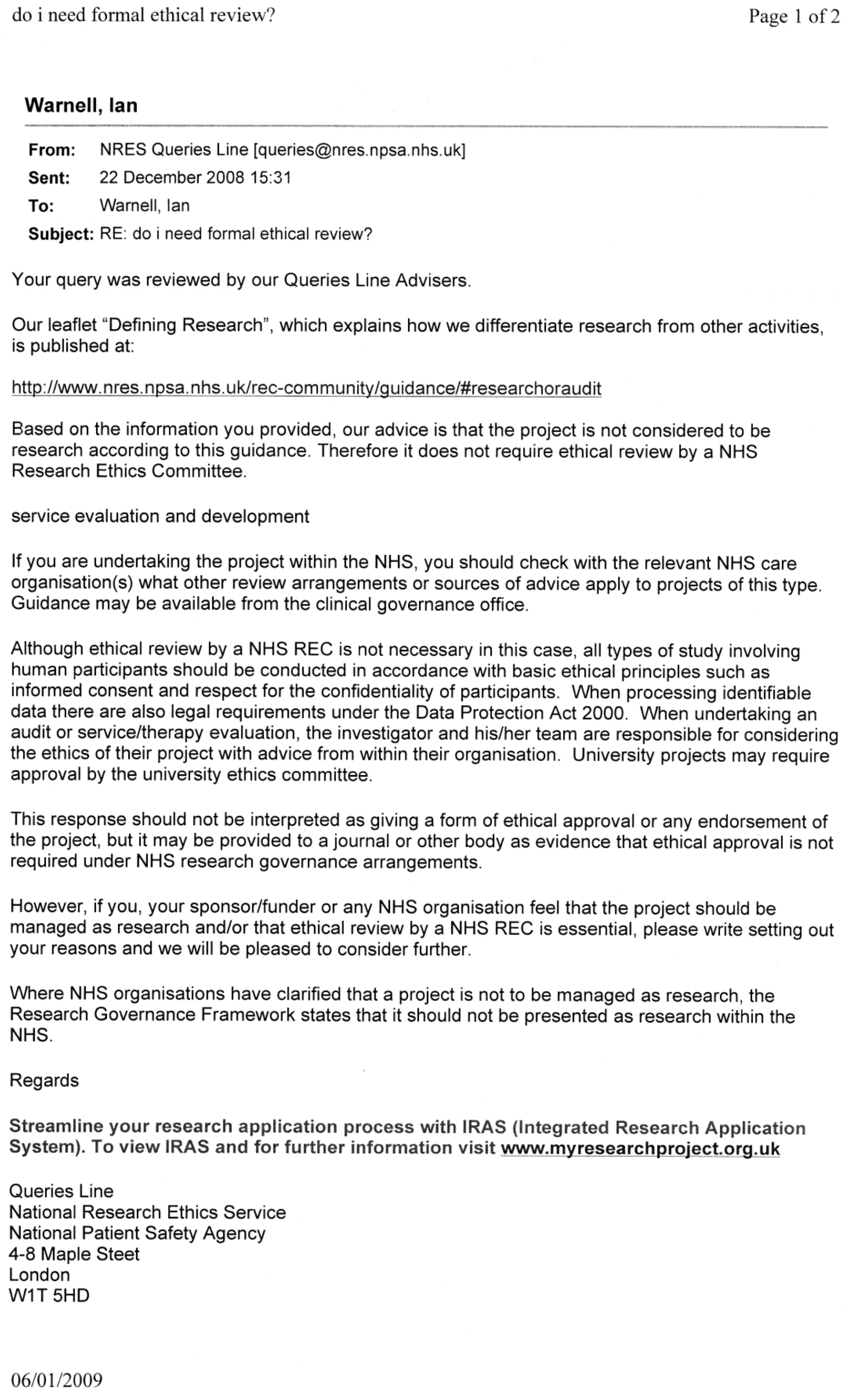
> ensuring that it complies with recognised ethical guidelines, lies with the organisation responsible for the care of the participants.
> If you are still unsure as to whether your study requires ethical approval you can contact the Chair of your local REC by email with a description of your proposal (one side of A4 in length) and seek their opinion. Alternatively you can send your A4 summary to NRES for an opinion (please email: queries@nres.npsa.nhs.uk

My advice would be to email the helpline as above with a detailed summary of your plans for the use of the data.

Let us know the outcome and we can then advise you further.

Regards
Lesley

Dr Lesley Hall, PhD
Research Governance Manager
Joint Research Office
Newcastle upon Tyne Hospitals NHS Foundation Trust
4th Floor Leazes Wing
Royal Victoria Infirmary
Queen Victoria Road
Newcastle upon Tyne
NE1 4LP



do i need formal ethical review? Page 1 of 2

Warnell, Ian

From: NRES Queries Line [queries@nres.npsa.nhs.uk]
Sent: 22 December 2008 15:31
To: Warnell, Ian
Subject: RE: do i need formal ethical review?

Your query was reviewed by our Queries Line Advisers.

Our leaflet "Defining Research", which explains how we differentiate research from other activities, is published at:

http://www.nres.npsa.nhs.uk/rec-community/guidance/#researchoraudit

Based on the information you provided, our advice is that the project is not considered to be research according to this guidance. Therefore it does not require ethical review by a NHS Research Ethics Committee.

service evaluation and development

If you are undertaking the project within the NHS, you should check with the relevant NHS care organisation(s) what other review arrangements or sources of advice apply to projects of this type. Guidance may be available from the clinical governance office.

Although ethical review by a NHS REC is not necessary in this case, all types of study involving human participants should be conducted in accordance with basic ethical principles such as informed consent and respect for the confidentiality of participants. When processing identifiable data there are also legal requirements under the Data Protection Act 2000. When undertaking an audit or service/therapy evaluation, the investigator and his/her team are responsible for considering the ethics of their project with advice from within their organisation. University projects may require approval by the university ethics committee.

This response should not be interpreted as giving a form of ethical approval or any endorsement of the project, but it may be provided to a journal or other body as evidence that ethical approval is not required under NHS research governance arrangements.

However, if you, your sponsor/funder or any NHS organisation feel that the project should be managed as research and/or that ethical review by a NHS REC is essential, please write setting out your reasons and we will be pleased to consider further.

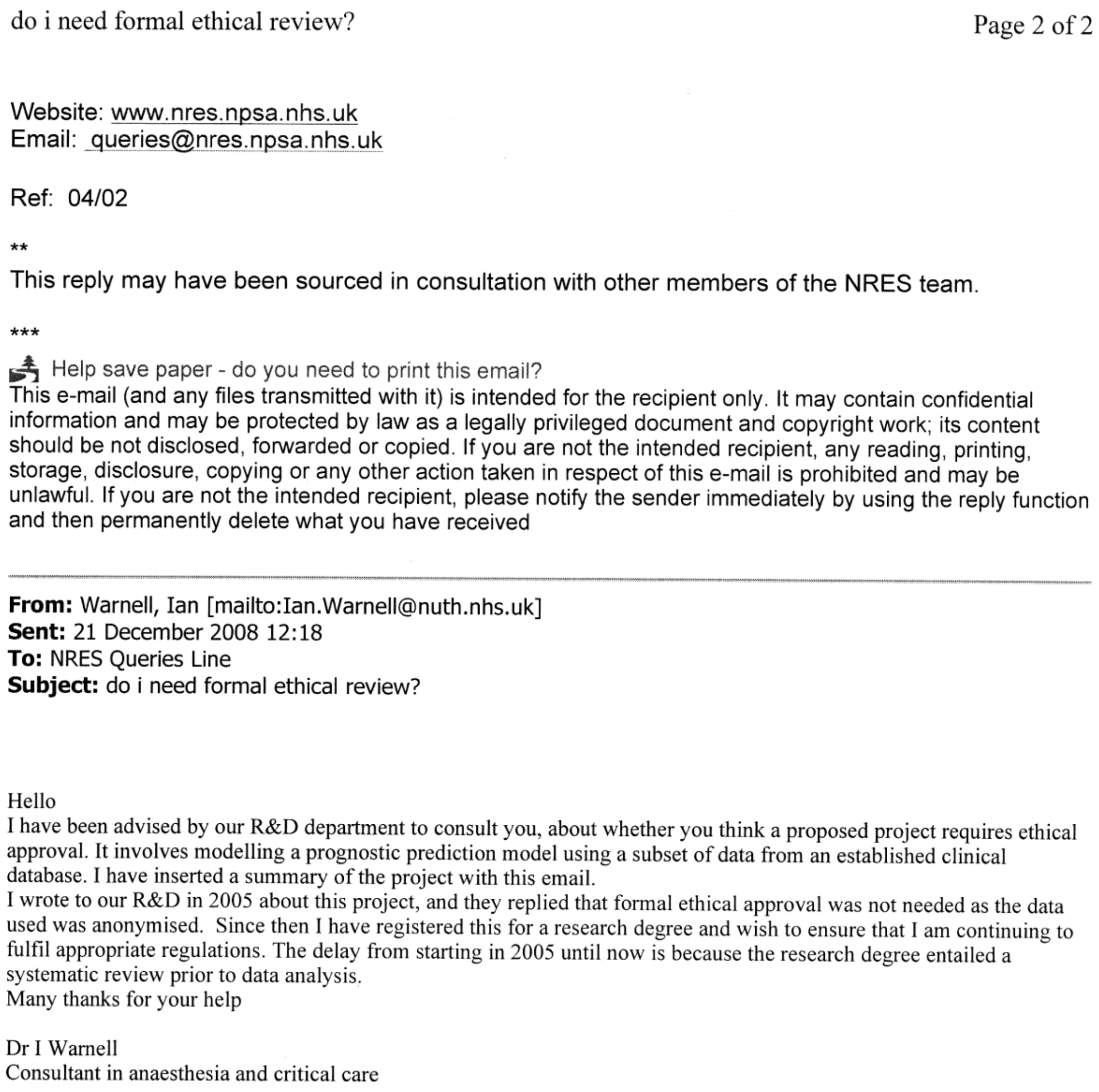
Where NHS organisations have clarified that a project is not to be managed as research, the Research Governance Framework states that it should not be presented as research within the NHS.

Regards

Streamline your research application process with IRAS (Integrated Research Application System). To view IRAS and for further information visit www.myresearchproject.org.uk

Queries Line
National Research Ethics Service
National Patient Safety Agency
4-8 Maple Steet
London
W1T 5HD

06/01/2009



do i need formal ethical review? Page 2 of 2

Website: www.nres.npsa.nhs.uk
Email: queries@nres.npsa.nhs.uk

Ref: 04/02

**

This reply may have been sourced in consultation with other members of the NRES team.

***

Help save paper - do you need to print this email?
This e-mail (and any files transmitted with it) is intended for the recipient only. It may contain confidential information and may be protected by law as a legally privileged document and copyright work; its content should be not disclosed, forwarded or copied. If you are not the intended recipient, any reading, printing, storage, disclosure, copying or any other action taken in respect of this e-mail is prohibited and may be unlawful. If you are not the intended recipient, please notify the sender immediately by using the reply function and then permanently delete what you have received

From: Warnell, Ian [mailto:Ian.Warnell@nuth.nhs.uk]
Sent: 21 December 2008 12:18
To: NRES Queries Line
Subject: do i need formal ethical review?

Hello
I have been advised by our R&D department to consult you, about whether you think a proposed project requires ethical approval. It involves modelling a prognostic prediction model using a subset of data from an established clinical database. I have inserted a summary of the project with this email.
I wrote to our R&D in 2005 about this project, and they replied that formal ethical approval was not needed as the data used was anonymised. Since then I have registered this for a research degree and wish to ensure that I am continuing to fulfil appropriate regulations. The delay from starting in 2005 until now is because the research degree entailed a systematic review prior to data analysis.
Many thanks for your help

Dr I Warnell
Consultant in anaesthesia and critical care

Appendix E. Appendix-Risk of bias in individual primary studies

	Bias item	Surgical exclusions described	Consecutive cases	Sample characteristics described	Data collection	Data validation	Missing values stated or deducible	Missing values handled appropriately	Cases lost to follow up do not differ from main sample	Procedures to ensure follow up	Follow up rate reported & acceptable	Cases lost to follow up do not differ from sample	Prognostic predictors defined	Valid prognostic predictor	Continuous variables handled appropriately	Adequate number of predictor values	Appropriate handling of missing predictors	Important confounders recorded	Important confounders accounted for in study	Description of appropriate statistical model	Sufficient data to assess analysis
Han-Geurts		N	U	M	P	N	U	U	U	U	P	U	M	M	P	P	U	N	N	P	U
Sabel		U	M	M	U	N	U	U	U	U	P	U	M	M	P	M	U	P	P	P	P
Tsai		N	P	M	R	N	U	U	U	U	P	U	M	M	P	U	U	P	P	M	P
Rahamim		N	M	M	P	P	U	U	U	U	P	U	M	M	P	U	U	N	N	M	M
Moskovitz		N	M	P	P	P	U	U	U	U	P	U	M	M	M	U	U	P	P	M	M
Finlayson		N	U	P	U	U	U	U	U	U	U	U	M	M	P	U	U	P	P	M	P
Law 2004		N	M	M	P	N	U	U	U	U	U	U	P	P	M	U	U	M	P	M	P
Fang		U	M	M	U	N	U	U	U	U	U	U	M	M	P	U	U	M	M	M	U
Alexiou		N	M	P	U	N	U	U	U	U	P	U	P	P	M	U	U	M	M	P	N
Kinugasa		N	M	M	U	N	U	U	U	U	P	U	M	M	P	U	U	M	M	P	M
Ferguson		N	M	P	R	N	U	P	U	U	P	U	P	M	M	M	M	M	M	M	M
Rentz		N	U	P	P	M	U	U	U	U	P	U	P	P	P	U	U	P	P	M	P
Ruol		N	P	M	U	N	U	U	U	U	P	U	M	M	P	U	U	P	P	M	P
Bailey		N	U	P	P	M	U	U	U	U	P	U	P	M	M	U	U	U	P	M	M
Alexiou		M	M	M	U	N	U	U	U	U	P	U	P	P	P	U	U	P	P	M	P
Law 1994		M	U	M	P	N	U	U	U	U	U	U	M	M	M	U	U	M	M	M	M

Appendix E (Risk of bias items for included studies)

	Bias item	Surgical exclusions described	Consecutive cases	Sample characteristics described	Data collection	Data validation	Missing values stated or deducible	Missing values handled appropriately	Cases lost to follow up do not differ from main sample	Procedures to ensure follow up	Follow up rate reported & acceptable	Cases lost to follow up do not differ from sample	Prognostic predictors defined	Valid prognostic predictor	Continuous variables handled appropriately	Adequate number of predictor values	Appropriate handling of missing predictors	Important confounders recorded	Important confounders accounted for in study	Description of appropriate statistical model	Sufficient data to assess analysis
Adam		N	M	M	U	P	U	U	U	M	P	U	P	P	U	U	U	P	P	M	U
Sauvanet		N	U	P	R	N	P	U	U	U	P	U	M	M	P	U	U	P	P	M	M
Liu		N	N	P	U	N	U	U	U	U	P	U	M	M	P	U	U	P	P	P	U
Kuwano		N	M	P	U	N	U	U	U	U	P	U	M	M	M	U	U	P	P	U	U
Schroder		N	U	M	P	N	U	U	U	U	P	U	M	P	P	U	U	M	M	M	M
Griffin		N	M	M	P	N	U	U	U	U	M	U	P	M	M	U	U	P	P	P	M
Abunasra		P	P	M	U	N	U	U	U	U	P	U	P	M	P	U	U	P	P	M	M
Ruol		N	P	M	U	N	U	U	U	U	P	U	M	P	P	U	U	M	M	P	M
Jougon		M	M	M	R	N	U	U	U	M	P	U	P	M	P	U	U	P	P	P	U
Tekkis		N	U	M	U	M	M	U	U	U	P	U	M	M	P	U	U	M	M	M	M
Thomas		N	U	M	U	N	U	U	U	U	U	U	M	M	P	U	U	M	P	M	M
Ellis Jr		N	M	M	R	N	U	U	U	U	P	U	M	M	P	M	U	N	N	M	M
Poon		P	M	M	P	N	U	U	U	U	P	U	M	M	P	U	U	M	P	M	M
Atkins		N	M	M	R	N	U	U	U	U	P	U	M	M	M	U	U	P	M	M	M
Forshaw		M	M	M	U	N	U	U	U	U	P	U	M	M	M	M	U	N	N	M	M
Gockel		N	U	M	P	N	U	U	U	U	P	U	M	M	M	U	U	M	M	M	N
Griffin		N	U	M	U	N	U	U	U	U	U	U	P	U	M	U	U	M	P	M	U
Murray		N	N	N	P	N	U	U	U	U	M	U	M	M	M	U	U	N	N	U	N

Appendix E (Risk of bias items for included studies)

	Bias item	Surgical exclusions described	Consecutive cases	Sample characteristics described	Data collection	Data validation	Missing values stated or deducible	Missing values handled appropriately	Cases lost to follow up do not differ from main sample	Procedures to ensure follow up	Follow up rate reported & acceptable	Cases lost to follow up do not differ from sample	Prognostic predictors defined	Valid prognostic predictor	Continuous variables handled appropriately	Adequate number of predictor values	Appropriate handling of missing predictors	Important confounders recorded	Important confounders accounted for in study	Description of appropriate statistical model	Sufficient data to assess analysis
Bartels		N	M	P	U	N	U	U	U	U	U	U	P	P	P	U	U	M	P	P	N
Lai		N	U	P	R	P	P	M	U	U	U	U	M	M	P	U	U	P	P	M	M
Nagabhushan		N	U	M	U	N	P	M	U	U	P	U	M	M	P	U	U	M	M	M	M
Lagarde		N	M	M	U	N	U	P	U	U	P	U	M	M	P	U	U	M	M	M	M
Saito		N	U	P	U	N	U	U	U	U	U	U	M	U	M	U	U	U	U	M	M
Sanz		M	M	M	P	N	U	U	U	U	U	U	P	P	P	U	U	M	M	M	U
Steyerberg		P	U	P	R	U	P	M	U	U	U	U	P	M	P	M	M	P	P	M	M
Zafirellis		N	M	M	R	N	U	P	U	U	P	U	M	M	P	M	U	M	M	M	M
Zhang		N	P	P	U	N	U	U	U	U	U	U	P	M	P	U	U	P	P	M	M
Whooley		N	M	P	U	N	U	U	U	U	U	U	M	M	M	U	U	M	U	M	N
Johansson		N	U	M	U	N	U	U	U	U	P	U	M	M	P	M	U	N	N	M	P
Healy		N	M	M	P	N	U	U	U	U	U	U	M	M	M	U	U	P	P	M	M
Ra		P	U	P	U	N	U	U	U	U	U	U	M	M	P	U	U	P	P	M	M
Leigh		N	U	N	U	N	P	P	U	U	U	U	M	M	P	M	U	N	N	M	M
Alibakhshi		P	M	M	R	N	U	U	U	U	P	U	M	M	P	U	U	P	P	M	M
Braiteh		P	U	M	R	N	U	U	U	U	U	U	M	M	P	U	U	P	P	M	M
Park		N	U	N	P	P	U	U	U	U	U	U	M	M	P	U	U	na	na	P	M

	Bias item	Surgical exclusions described	Consecutive cases	Sample characteristics described	Data collection	Data validation	Missing values stated or deducible	Missing values handled appropriately	Cases lost to follow up do not differ from main sample	Procedures to ensure follow up	Follow up rate reported & acceptable	Cases lost to follow up do not differ from sample	Prognostic predictors defined	Valid prognostic predictor	Continuous variables handled appropriately	Adequate number of predictor values	Appropriate handling of missing predictors	Important confounders recorded	Important confounders accounted for in study	Description of appropriate statistical model	Sufficient data to assess analysis
Tagagawa		N	M	M	R	U	U	U	U	P	P	U	M	M	P	U	U	P	P	M	M
Takeno		P	U	M	U	N	U	U	U	U	U	U	M	M	M	U	U	P	P	P	M
Zingg		N	M	M	R	N	M	na	na	U	U	U	M	M	P	U	na	M	P	M	P

Appendix F. Data handling procedures

Some specific data handling procedures which were written for data cleaning procedures in Visual Basic for Applications are listed below. Comments are preceded by the symbol # or ‘.

Appendix F i. Check merged and moved fields are aligned correctly.

```
Sub CheckIndexVarAlignment1576()
#Select first column to check; select second column; insert column to
left of selected field; check against another index (e.g. unique key); if
same 0, if different 1, find 1. Target fields moved with ‘index’ &/ ‘date
of birth’ fields to check alignment in new spreadsheet
Dim i As Integer 'looping through rows
Dim j As Integer 'variable for second index variable
Dim k As Integer 'variable for first index variable
'input the two index fields to be checked
k = InputBox("which number column is the first index field in-input as
column number?")
j = InputBox("which number column is the second index field?")
 'insert col and move others to right to make space; note index now in
k+1
ActiveSheet.Columns(k).Insert 'Shift:=xlToLeft
    For i = 2 To 1576
        If Cells(i, k + 1).Value = Cells(i, j).Value Then
        Cells(i, k).Value = 0
        Else: Cells(i, k).Value = 1
        End If
        Next i
End Sub
'complete check by using Excel FIND to find ‘1’s', which indicate non-
aligned fields
```

Appendix F ii. Clean spreadsheet cells of invisible characters

```
Sub clean_textcells()
‘Removes spaces and unprinted symbols from text; input column for
original data and column for new data; loops from 2 to 1576 rows
Dim i As Integer 'looping through rows
Dim j As Integer 'variable for new data
Dim k As Integer 'variable for old data
k = InputBox("which number column is the old data in?")
j = k + 1
ActiveSheet.Columns(k).Insert 'Shift:=xlToLeft
For i = 2 To 1576
Cells(i, k).Value =
WorksheetFunction.clean(WorksheetFunction.Trim(Cells(i, j).Value))
Next i
Cells(1, k) = Cells(1, j).Value & "Cln" ‘cln signifies cleaned data field
ActiveSheet.Columns(j).Delete
```

Appendix F iii. Conversion macros

```
#Create 'Age at operation' field
# 'Operation date'  minus 'DOB'
=DATEDIF(start_period, end_period, "y")
#"y" is code for years
# Create 'Gender' logical field
#Conversion from text to male=1, female=2, blank =""
=IF("male",1, IF("female",2,""))
#Convert height field to metres
'Conversion from centimeters to metres; non empty cell with value greater
than 3 (only likely be centimeters) divided by 100; less than 3 can only
be metres
=IF(O2="","", IF(O2>3,O2/100,O2))
#Body surface area (BSA) (m^2)
Dubois formula for BSA = (W ^0.425 x H ^0.725) x 0.007184, where H is height
(cm) and W is weight (Kg)
=IF(D2="", "", IF(C2="","",POWER(D2*100,0.725)*POWER(C2,0.425)*0.007184))
Test for missing data, when leave field 'BSA' empty, otherwise use
function above to calculate BSA
#Body Mass Index BMI
BMI = wt (Kg) / Ht(m^2)
=IF(C1576="", "", IF(D1576="","",PRODUCT(C1576,1/POWER(D1576,2))))
```

Appendix F iv. Extract fatalities with details from database

```
Sub MortalityCauses()
Option Explicit
#declarations
Dim i As Integer
Dim j As Integer
Dim k As Integer
Dim rngMort As Range
Dim rngCopy As Range
#set values
Set rngCopy = Worksheets("OutcomeData").Range("A2:Q1576")
Set rngMort = Worksheets("Mortality").Range("A2:Q120")
j =1
'activate OutcomeData sheet
Sheets("OutcomeData").Activate
#loop through each record in OutcomeData worksheet in mortality outcome
field & copy details into "Mortality" worksheet
For i = 2 To 1576
k = i - 1
    If Cells(i, 2) = "Yes " Then
    rngCopy.Rows(k).Select
    Selection.Copy
    Sheets("Mortality").Activate
    rngMort.Cells(j, 1).Select
    ActiveSheet.Paste
    #Alternative code:
    'ActiveSheet.Paste Destination:=Worksheets("Mortality").Cells(j, 1)
    #stops moving border
    Application.CutCopyMode = False
    j = j + 1
    Worksheets("OutcomeData").Activate
    End If
Next i
End Sub
```

Appendix F v. Checks that all 30 day mortality has matching 'in hospital' mortality

```
Sub tallyinhosp_30d_mort()
#checks that all 30 day mortality has corresponding ‘in hospital’
mortality
#declarations
Dim i As Integer
Dim strHosp As String
Dim strThirtyDay As String
Dim strTally
#loops through all 30 day mortality
For i = 2 To 1576
  strThirtyDay = Cells(i, 5)
  strHosp = Cells(i, 4)
  If strThirtyDay = "Yes" And strHosp = "No" Then
  strTally = 1    'if 30d mort and not inhosp
  Else
  strTally = 0 'all other combinations inc no inhosp mort
  End If
Cells(i, 8).Value = strTally
Next i
End Sub
```

Appendix F vi. Checks that all classified as survivor has attended a matching follow up outpatient appointment

```
Sub InhospMort_Discharge()
# checks that survivors in mortality field had a reported outpatient
follow up date
Dim strHosp As String
Dim strTally
For i = 2 To 1576
If strHosp = "No" And IsEmpty(Cells(i, 3).Value) Then
strTally = 1
Else: strTally = 0
End If
Cells(i, 8).Value = strTally
Next i
End Sub
```

Appendix F vii. Recodes the comorbidity variable "OTHER".

```
Option Explicit
'Excel code to code the variable "OTHER";
'this is a freetext field with preoperative morbidity. cells are searched
for terms which are mapped to a comorbidity in a lookup table
Private Sub CommandButton1_Click()
Dim other As String
Dim otherpivot As String
Dim j As Integer
Dim i As Integer
Dim rng As Range
Set rng = Worksheets("Sheet2").Range("C1:V618")
For i = 2 To 1576
other = Worksheets("NOGCDAT").Cells(i, 20).Value
      For j = 5 To 618
      otherpivot = Worksheets("Sheet2").Cells(j, 1).Value
If other = otherpivot Then
Worksheets("NOGCDAT").Range(Cells(i, 21), Cells(i, 40)) =
rng.Cells.Rows(j).Value
      End If
      Exit For
Next j
End Sub
```

Appendix F viii. Extracting and categorizing free text description of smoker status

```
Private Sub smokecode_Click()
Dim smoke As String
Dim code As Integer
Dim i As Integer
For i = 2 To 1576
        smoke = Cells(i, 49).Value
        Select Case smoke
            Case "Current"
                code = 1
            Case "Ex-smoker (> 1 year)"
                code = 2
            Case "Never"
                code = 3
            Case "Unknown"
                code = 4
            Case ""
                code = 5
          End Select
          Cells(i, 50).Value = code
Next i
End Sub
```

Appendix F ix. Recoding the free text 'OPERATION' field into surgical categories

Free text was initially recoded into categories of 'thoracic oesophagectomy, 'transhiatal oesophagectomy', 'other thoracic procedure', total gastrectomy, 'other'. Cell contents were mapped to surgical categories in a lookup table. The surgical procedure was subsequently collapsed to 'thoracic oesophagectomy', 'total gastrectomy' and 'other'. The free text field 'operation' was recoded into 3 surgical categories from lookup table shown on next page.

```
Sub Surgical_classn1()
'reclassifies surgical recorded according to lookup table
Dim i As Integer 'loops through spreadsheet to 1576
Dim j As Integer 'column num for recorded operations data
Dim k As Integer 'column num for new surgical classification
Dim l As Integer 'top row of surgical lookup table
Dim m As Integer 'lowest row of lookup table
Dim n As Integer
Dim x As String  'string variable
Dim p As Integer
Dim q As Integer
j = InputBox("which number column is your main data in?")
k = InputBox("which number column is your new surgical classification
going in?")
l = InputBox("which is the top row num of your surgical lookup table?")
m = InputBox("which is the lowest row num of your surgical lookup
table?")
p = InputBox("which col num of the lookup table is the reported
operation?")
q = InputBox("which col num is the new surgical classfn in?")
For i = 2 To 1576
    If IsEmpty(Cells(i, j)) Then
    Cells (i, k).Value = ""
    Else: x = Cells(i, j).Value
        For n = l To m
            If Cells(n, p).Value = x Then
            Cells(i, k).Value = Cells(n, q).Value
            Exit For
            End If
        Next n
    End If

Next i
End Sub
```

Appendix F x.

Original Freetext description of surgical procedure	New Surgical_Classfn1
Completion Gastrectomy (stump gastrectomy)	Other
Completion Gastrectomy (stump gastrectomy) and Feeding Jejunostomy	Other
Completion Gastrectomy (stump gastrectomy) and Other specify	Other
Completion Gastrectomy (stump gastrectomy), Feeding Jejunostomy and Other specify	Other
Extended total gastrectomy	Total
Extended total gastrectomy and Feeding Jejunostomy	Total
Feeding Jejunostomy	Other
Feeding Jejunostomy and Other specify	Other
Laparotomy and Thoracotomy Only	Thoracot Other
Laparotomy and Thoracotomy Only & Feeding Jejunostomy	Thoracot Other
Laparotomy Only	Other
Laparotomy Only & Other specify	Other
Laparotomy Only and Feeding Jejunostomy	Other
Laparotomy Only and Other specify	Other
Laparotomy Only, Feeding Jejunostomy and Other specify	Other
Left Thoraco-Abdominal Oesophagectomy	Thoracic Oesph
Left Thoraco-Abdominal Oesophagectomy & Feeding Jejunostomy	Thoracic Oesph
Left Thoraco-Abdominal Oesophagectomy and Other specify	Thoracic Oesph
McKeown 3 Stage Sub Total Oesophagectomy	Thoracic Oesph
McKeown 3 Stage Sub Total Oesophagectomy & Feeding Jejunostomy	Thoracic Oesph
McKeown 3 Stage Sub Total Oesophagectomy, Feeding Jejunostomy & Other specify	Thoracic Oesph
Other specify	Other
Partial Gastrectomy	Other
Right 2 Phase Sub-Total Oesophagectomy	Thoracic Oesph
Right 2 Phase Sub-Total Oesophagectomy & Feeding Jejunostomy	Thoracic Oesph
Right 2 Phase Sub-Total Oesophagectomy & Other specify	Thoracic Oesph
Right 2 Phase Sub-Total Oesophagectomy and Feeding Jejunostomy	Thoracic Oesph
Right 2 Phase Sub-Total Oesophagectomy and Other specify	Thoracic Oesph
Right 2 Phase Sub-Total Oesophagectomy, Feeding Jejunostomy & Other specify	Thoracic Oesph
Right 2 Phase Sub-Total Oesophagectomy, Feeding Jejunostomy and Other specify	Thoracic Oesph
Sub-Total Gastrectomy	Other
Sub-Total Gastrectomy & Other specify	Other
Sub-Total Gastrectomy & Wedge/localised resection	Other
Sub-Total Gastrectomy and Feeding Jejunostomy	Other
Sub-Total Gastrectomy and Other specify	Other
Sub-Total Gastrectomy, Feeding Jejunostomy and Other specify	Other
Total Gastrectomy	Total
Total Gastrectomy & Feeding Jejunostomy	Total
Total Gastrectomy & Other specify	Total
Total Gastrectomy and Feeding Jejunostomy	Total
Total Gastrectomy and Other specify	Total

Total Gastrectomy, Feeding Jejunostomy and Other specify	Total
Total Gastrectomy, Right 2 Phase Sub-Total Oesophagectomy and Other specify	Thoracic Oesph
Total Gastrectomy, Right 2 Phase Sub-Total Oesophagectomy, Feeding Jejunostomy and Other specify	Thoracic Oesph
Trans-hiatal Oesophagectomy	THOesoph
Trans-hiatal Oesophagectomy & Feeding Jejunostomy	THOesoph
Wedge/localised resection	Other
Wedge/localised resection & Other specify	Other
Wedge/localised resection and Other specify	Other

Appendix F xi. Recoding of histology from free text field to new categories.

```
Sub HistNew1_classn1()
'reclassifies histology recorded in original database into new field
according to lookup table derived from pivot table_
'on original data. Original freetext descriptions and new categories are
shown after the VBA code
Dim i As Integer 'loops through spreadsheet to 1576
Dim j As Integer 'column num for recorded histology
Dim k As Integer 'column num for new histology classification
Dim l As Integer 'top row of histology lookup table
Dim m As Integer 'bottom row of lookup table
Dim n As Integer
Dim x As String  'string variable
Dim p As Integer
Dim q As Integer
If IsEmpty(Cells(i, j)) Then
Cells(i, k).Value = ""
Else: x = Cells(i, j).Value
        For n = l To m
            If Cells(n, p).Value = x Then
            Cells(i, k).Value = Cells(n, q).Value
            Exit For
            End If
        Next n
End If
Next i
End Sub
```

Table 36 Lookup table to convert histology free text to new histology code (Abreviations: ACA, adenocarcinoma; SCC, squamous cell carcinoma, HGC, high grade dyplasia)

Old Histology term	New Histology category
ACA	ACA
ACA & Barrett's	ACA
ACA & HGD	ACA
ACA & Intramucosal cancer	ACA
ACA & Other specify	ACA
ACA and HGD	ACA
ACA and Leioyoma	ACA
ACA and Other	ACA
ACA and SCC	ACA
ACA and Small Cell	ACA
ACA, Barrett's & HGD	ACA
ACA, Intramucosal cancer & Lymphoma	ACA
ACA, SCC and Other	Other
ACA, SCC and Small Cell	Other
Adenoid-cystic	Other
Adenosquamous	Other
Barrett's & HGD	Benign
Benign	Benign
Benign and Leioyoma	Benign

Old Histology term	New Histology category
Benign and Other	Benign
Benign, HGD & Leiomyoma	Benign
Carcinoid	Other
Carcinoid & Other specify	Other
Carcinoma	Other
Dysplasia	Benign
Dysplasia, HGD & Other specify	Benign
EGC	Other
HGD	Benign
HGD & Intramucosal cancer	Other
HGD and Other	Other
Intramucosal cancer	Other
Leiomyoma	Benign
Leioyoma	Benign
Leioyoma and Other	Other
Lymphoma	Other
Melanoma	Other
Neuroendocrine	Other
Neuroendocrine & Other specify	Other
No tumour	Benign
Normal/Benign	Benign
Normal/Benign & Leiomyoma	Benign
Other	Other
Other specify	Other
SCC	SCC
SCC	SCC
Small Cell	Other
Undifferentiated	Other

Appendix F xii. Generation of RCRI terms from free text fields of cardiac morbidity

The contents of the cardiac comorbidity free text fields ‘Cardiac’, ‘Comorbid’, ‘Details’, ‘ECG’, ‘ECGDETAILS’ were searched using the Excel Pivot-Table function. Cardiac terms were extracted and mapped to terms suitable for generating the RCRI (Fleisher *et al.*, 2007; Poldermans *et al.*, 2009) in an Excel lookup table. The contents of all records were searched and scored 0 or 1 for each RCRI item. Any version of the RCRI (total score, 2 or 4 level) could be then calculated. The RCRI terms were: PVD(peripheral vascular disease), IHD(ischaemic heart disease), valve(acquired valve disease), CHD(congenital heart disease), VentOther(ventricular diagnosis excluding IHD or heart failure), HF(heart failure), cholesterol(any hyperlipidaemia),SOB (shortness of breath), CVD (cerebrovascular disease), unstable coronary(unstable coronary syndromes). The lookup tables for mapping the terms from ‘cardiac’ fields are in Table 37, and from the electrocardiographic fields are in Table 38. The conversion code follows on the next page.

Extraction and remapping of cardiac comorbidity to new Revised Cardiac Risk Index terms.

```
Sub CardiacExtraction()
'extracts txt from field 'Cardiac' & 'Comorbid' & 'Details' and codes it
to new cardiac categorical fields;
'These are stored in 'Worksheet.CardiacClassification. The cardiac txt
and categories were found by using the Excel pivot table function
Dim i As Integer 'integer counter for loop through 'pulmonary'
Dim k As Integer 'loop counter for cardiac term lookup table
Dim j As Integer
Dim q As Integer 'input
Dim l As Integer
Dim m As Integer
Dim p As Integer
Dim category As Integer 'column number for each new cardiac category
Dim emptycellBo As Boolean
Dim test As Boolean
Dim r1 As Integer 'input
Dim s As Integer
Dim T As Integer
Dim x As Integer 'counter for each of main data cols
'set variables for this macro without using input macro
s = 2
T = 1576
r1 = 7
l = 2
m = 73
p = 30
q = 32
For i = s To T
      For j =3 To 6
            emptycellBo = IsEmpty(Cells(i, j))  'is the cell of the main
data col empty
            If emptycellBo = False Then
                  'loop through each row in the new cardiac category
lookup table
                  For k = l To m
                  test =
WorksheetFunction.IsNumber(Application.Search(Cells(k, p), Cells(i, j)))
                        'does the cell contain the cardiac text as
written in the cardiac lookup
                        If test = True Then
                              'colCat = Cells(k, q).Value
                              category = r1 + (Cells(k, q).Value)
                              ‘move to appropriate column, one column for
each cardiac category
                              Cells(i, category) = 1'code 1 if cardiac
category present
                        End If
                  Next k
             End If
      Next j
Next i
End Sub
```

Table 37 Lookup table to convert text terms to RCRI terms

Words extracted from database	New cardiac terms	Words extracted from database	New cardiac terms
aaa	PVD	ischaem	IHD
aneurysm	PVD	L vent.Impairment	VentOther
PVD	PVD	LVF	HF
AF	arrythmia	LVH	VentOther
claudication	PVD	mitral	valve
angiop	IHD	ngina	IHD
aortic	valve	None	None
fem	PVD	Pacemaker	arrythmia
arrythmia	arrythmia	palpitations	arrythmia
ASD	CHD	paroxysmal ventricular fibrillation	arrythmia
atrial ectopics	arrythmia	ovale	CHD
atrial fibrilation	arrythmia	pvd	PVD
atrial fibrillation	arrythmia	raised blood pressure	hypertension
BP	hypertension	RBBB	arrythmia
bradycardia	arrythmia	ablation	arrythmia
bypass.	Unknown	sob	SOB
CABG	IHD	stent	IHD
cardiomyopathy	VentOther	svt	arrythmia
carditis	VentOther	TIA	CVD
carotid artery stenosis	CVD	tia	CVD
claudication	PVD	triple	IHD
cva	CVD	unstable	unstable coronary
cvs	Unknown	Valve	valve
dvt	VTE	Wolf	arrythmia
enlarged	VentOther	Peripheral Vascular Disease	PVD
failure	HF	Endarterectomy	CVD
block	arrythmia	Arteriopath	PVD
HT	hypertension	Stroke	CVD
hypercholest	cholesterol	hyperlipid	cholesterol
hypertension	hypertension	brain haemorrhage	CVD
hypotension	other	cerebro	CVD
IHD	IHD	ETT	Unknown
infarct	IHD	artery disease	Unknown
irregular	arrythmia	PM	Unknown
RFV	Unknown	PMH	Unknown
bypass	Unknown	mx	IHD

Table 38 Lookup table to convert terms from ECG field to RCRI terms

Terms used to search fields ECG, ECGDETAILS	New ECG (RCRI) categories	Terms used to search fields ECG, ECGDETAILS	New ECG (RCRI) categories
Bradycardia	Arrythmia	wave abnormality	ST T wave abnormality
hypertrophy	LVH	atrial fibrillation	Arrythmia
lvh	LVH	block	Conduction
AF	arrythmia	white	Arrythmia
lbbb	Conduction	junctional	Arrythmia
ischaem	ST T wave abnormality	T wave	ST T wave abnormality
rbbb	Conduction	atrial fibrilation	Arrythmia
ST T	ST T wave abnormality	tachycardia	Arrythmia
wpw	arrythmia	pace	Conduction
		prem	Arrythmia

ECG terms LVH(left ventricular hypertrophy), ST/T wave abnormality were mapped to 'VentOther' in cardiac morbidity. Key: PVD(peripheral vascular disease), IHD(ischaemic heart disease), valve(acquired valve disease), CHD(congenital heart disease), VentOther(ventricular diagnosis excluding IHD or heart failure), HF(heart failure), cholesterol(any hyperlipidaemia),SOB (shortness of breath), CVD (cerebrovascular disease), unstable coronary(unstable coronary syndromes).

Appendix F xiii. Generation of respiratory comorbidity categories from free text fields

The code for extracting respiratory terms form the database fields are given below. The Lookup table of old and new terms are shown in *Table 39*.

```
Sub PULMCoding()
'extracts pulmonary disease txt from field 'PULM' and codes it to new
pulmonary categorical fields;
'COPD(chronic obstructive pulmonary disease); CHRONIC LUNG DISEASE;
ASTHMA; ACUTE PULMONARY DISEASE EVENT. The pulmonary freetext entries and
categories were summarised using the Excel PivotTable & listed in 'Pulm
terms from count' field.
Dim i As Integer 'integer counter for loop through 'pulmonary'
Dim k As Integer 'loop counter for respiratory lookup table
Dim j As Integer
Dim q As Integer 'input
Dim l As Integer
Dim m As Integer
Dim p As Integer
Dim category As Integer 'column number for each new drug category
Dim emptycellBo As Boolean
Dim test As Boolean
Dim r1 As Integer 'input
Dim s As Integer
Dim t As Integer
'loop through the original field 'pulm'
For i = s To t
      emptycellBo = IsEmpty(Cells(i, j))'is the cell of the main data col
empty
            If emptycellBo = False Then
                  'loop through each row in the new pulmonary category
list
                  For k = l To m
                  test =
WorksheetFunction.IsNumber(Application.Search(Cells(k, p), Cells(i, j)))
                        'does the cell contain the resp text in the 'Pulm
terms from count' field
                        If test = True Then
                        'colCat = Cells(k, q).Value
                        category = r1 + Cells(k, q).Value
                        'move to appropriate column one column for each
resp category
                        Cells(i, category) = 1             'code 1 if
respiratory category present
                        End If
                  Next k
            End If
Next i
End Sub
```

Table 39 Lookup table of old database terms and new classification for respiratory comorbidity

Pulmonary free text and categorical terms extracted from the database	New pulmonary classification
asbestos	Chronic lung disease
asthma	Asthma
copd	COPD
coad	COPD
emphysema	COPD
fibrosis	Chronic lung disease
bronchiectasis	Chronic lung disease
lobectomy	Chronic lung disease
embolism	acute
PE	acute
lung cancer	Chronic lung disease
embolism	acute
chronic bronchitis	COPD
pneumothorax	acute
pneumonia	acute
collapsed	acute
farmer	Chronic lung disease
TB	Chronic lung disease
pigeon	Chronic lung disease
sarcoid	Chronic lung disease

Appendix F xiv. Sample code to extract general comorbidities from database. This code extracts terms for diabetes

```
Code to create categorical diabetes field.
Sub ExtractDiabetes()
'extracts txt from field 'Comorbid', 'Other' & 'Details' and codes it
to new diabetes categorical fields;
'These are stored in 'Worksheet.Diabetes The diabetes txt and
categories were found by using the
'Excel pivot table function
Dim i As Integer 'integer counter for loop through 'pulmonary'
Dim k As Integer 'loop counter for diabetes term in search term lookup
table
Dim j As Integer,Dim q As Integer 'input,Dim l As Integer,Dim m As
Integer,Dim p As Integer
Dim category As Integer 'column number for each new cardiac category
Dim emptycellBo As Boolean,Dim test As Boolean,Dim r1 As Integer
'input
Dim s As Integer,Dim t As Integer,Dim x As Integer 'counter for each
of main data cols
'set variables for this macro without using input
s = 2,t = 1576,r1 = 7,l = 2,m = 4,p = 8,'q = 32
'loop through all records of the fields 'comorbid', 'other' and
'details'
For i = s To t
      For j =2 To 4
      emptycellBo = IsEmpty(Cells(i, j))  'is the cell of the data col
empty
            If emptycellBo = False Then
                  'loop through each row in the diabetic search term
lookup table
                  For k = l To m
                  test =
WorksheetFunction.IsNumber(Application.Search(Cells(k, p), Cells(i,
j)))
                        'does the cell contain the diabetes text as
written in the diabetic search term lookup
                        If test = True Then
                        'OPTIONS FOR POPULATING NEW COLS
                        'colCat = Cells(k, q).Value
                        'category = r1 + (Cells(k, q).Value)
                        'IF DIABETIC TERM PRESENT COL IS 1, OTHERWISE
0
                        'Cells(i, category) = 1              'code 1 if
cardiac category present
                        Cells(i, 10) = 1                   ' code 1 if
diabetes category
                        End If
                  Next k
            End If
      Next j
Next i
End Sub
```

Appendix G. Code for prediction modelling and validation

The random development and validation samples were generated in Excel. Graphics plots were made in software packages based on R , an open source statistical software (Crawley, 2007; R Development Core Team, 2011). Code is in the text boxes in Courier New font and explanatory notes are preceded by the symbol '#'. The code for setting up data in R is in sections following.

Appendix G i. Creating two randomly split data samples using Excel

Two approximately equal sized randomly selected samples of data were created using VBA for Excel. The following code allocates a random number to each record using the RAND function, a pseudorandom number generator which satisfies stringent tests for producing random numbers in samples of this size (**http://i.cs.hku.hk/~diehard**, **http://support.microsoft.com/kb/828795**). It separates the data into subsets of survivors and non-survivors. The subsets were sorted in increasing random number size (0-1), and 50% of each set merged to give two approximately equal random samples, each containing approximately equal numbers of non-survivors. The following allocates random numbers to each record

```
Random sample generator
# RandomSample() was applied to the full dataset of  unique ID field
(col 1), and mortality field, “yes”/”no” (col2); the random number was
generated in col 3.
Public Sub RandomSample()
Dim myRange As Range
Set myRange = Worksheets("Randomisation").Range("C2:C1576")
myRange.Formula = "=RAND()"
myRange.Font.Bold = True
End Sub
```

The following code separates mortality status after random number has been allocated

```
#creates two new triple column sets of survivors and non-survivors
from three cols (id, mortality status, random number)
Sub SeparateYesNoMortRandom()
Dim j As Integer
Dim i As Integer
Dim k As Integer
j =2
k =2
 For i =2 To 1576
  If Cells(i, 2).Value = "Yes" Then
  Cells(j, 6).Value = Cells(i, 1).Value
  Cells(j, 7).Value = Cells(i, 2).Value
  Cells(j, 8).Value = Cells(i, 3).Value
  j =j +1
  ElseIf Cells(i, 2).Value = "No" Then
  Cells(k, 9).Value = Cells(i, 1).Value
  Cells(k, 10).Value = Cells(i, 2).Value
  Cells(k, 11).Value = Cells(i, 3).Value
  k=k +1
  End If
Next i
End Sub
```

The following code populates the random samples with required fields

```
Sub ExtractSampleField()
' ExtractSampleField Macro
' creates samples based on the random samples 1 or 2. Extracts required field
from main dataset
Dim x As Integer 'row of target in sample field "Field"
Dim i As Integer 'loop counter
Dim rge As Range 'ID index col number for main dataset
Dim sampleTop As Integer 'top data row position in spreadsheet of random
‘sample index field
Dim sampleBottom As Integer 'bottom data row in spreadsheet random sample
‘index field
Dim sampleIndex As Integer 'random sample index column num
Dim sampleContent As Variant 'contents of sample index cell
Dim Field As Range '
Dim FieldCol As Integer
Dim sampleNewFld As Integer
On Error Resume Next
Application.DisplayAlerts = False
Set rge = Application.InputBox(Prompt:="Select field which contains IDindex
‘for main dataset", _
Title:="Select Main index column", Type:=8)
Set Field = Application.InputBox(Prompt:="Select array of columns which
contain main data index and fields, which contain the data to be extracted", _
Title:="Select Main index column and attached fields", Type:=8)
'DATA INPUT
sampleIndex = InputBox("What column num is random sample index ID in?")
sampleTop = InputBox("What is the top row num in worksheet of the random
sample index ID?")
sampleBottom = InputBox("What is the bottom row num in worksheet of the random
sample index ID?")
sampleNewFld = InputBox("which is the new sample data column")
FieldCol = InputBox("Which field (relative to the array)of the main data array
contains the extractable data")
On Error GoTo 0
Application.DisplayAlerts = True
'loop through each sample cell
For i = sampleTop To sampleBottom
sampleContent = Cells(i, sampleIndex).Value
x = Application.Match(sampleContent, rge, 0)
Cells(i, sampleNewFld).Value = Application.Index(Field, x, FieldCol)
Next i
End Sub
```

The following procedures are in R.

Appendix G ii. Data setup (O'Day, 2011)

```
rm(list=ls()) #remove any old variables
link <- choose.files()#use this to find datafile path and copy into
script
#read data
my_data <- read.table(link)
skip =0, #skip records
sep = ",", #records separated by ","
dec=".",  #decimal place symbol
row.names = NULL,
header = T,#include first row as variables names
colClasses = c( "numeric","numeric","numeric","numeric"),#specify data
cols
comment.char = "#",
na.strings = c(""))
```

Appendix G iii. Distribution & mortality plots for each predictor

This splits age into 10 groups with preselected cut points to give approximately equal numbers of cases in each level. The 'binom.exact' code from 'epitools'(Aragon, 2010) created exact 95% confidence intervals for mortality proportions in each age decile. The following code generates data.

```
b<-c(20,53,58,61,64,67,69,72,74,77,99)
my_data$bin<-cut(my_data$age, breaks=b, labels = NULL,
include.lowest = FALSE, right = TRUE, dig.lab = 3,
ordered_result = FALSE, )
##Creates tables of frequencies of mortality by the age groups and
puts in array 2X10
bot<-array(table(my_data$mortnum,my_data$bin),c(2,10))
x<-bot[2,]
notx<-bot[1,]
n<-x+notx
agetable<-as.data.frame(binom.exact(x, n, conf.level = 0.95))
agetable$binlevels<-levels(my_data$bin)
agetable$binlevels##adds binlevel vector to dataframe
```

The following code generates the plot in ggplot2 (Wickham, 2009)

```
library(ggplot2)
age_gpl<-ggplot(agetable,aes( agetable$binlevels,
agetable$proportion))
age_gpl<-age_gpl+geom_point(size=4, shape=18)+
geom_linerange(aes(x=agetable$binlevels, ymax=agetable$upper,
ymin=agetable$lower,),size=1) +
xlab("Equal groups of ascending age with exact 95% CI")+
ylab("Mortality rate%")+
###sets markers
opts(
#panel.grid.major = theme_blank(), #removes grids
panel.grid.minor=theme_blank(),
title="Mortality by age group",
plot.title=theme_text(size=sizetitle,hjust=0.5,face="bold"))
#panel.background = theme_rect(fill="grey95",colour=NA) )
age_dist<- ggplot(my_data,aes(age))
 age_dist<-
age_dist+geom_histogram(binwidth=3,fill="grey60")+labs(x="Age at
surgery in years", y="Number of patients")+
opts(title="Age distribution",plot.title=theme_text(size=sizetitle,
hjust=0.5,face="bold"))
```

Appendix G iv. Code to plot histograms of predicted mortality in validation samples

```
#set variables
y<-my_data$mortnum
xlabel_title="Predicted mortality"
model<-my_data$pred_clin
plot_title<-"Prespecified clinical model"
#Generate histograms of mortality predictions in survivors and non-
survivors
pred_dist<- ggplot(my_data,aes(x=model, group=mortnum, fill=factor(mortnum)))
+scale_fill_manual(values = c("grey60", "black"),
          name="Key",
          breaks=c("0", "1"),
          labels=c("survivors", "deaths"))
pred_dist<-pred_dist+geom_histogram(binwidth=0.01, , position="dodge") +
labs(x=xlabel_title, y="Number of patients")+

 opts(
          title=plot_title,
          #panel.grid.major = theme_blank(), #######removes grids
           panel.grid.minor=theme_blank(),
          legend.key=theme_blank(),
          legend.name=theme_blank(),
          legend.background=theme_blank(),
          #legend.text=theme_blank(),
          legend.title=  theme_text(hjust=0),
          legend.justification=c(1,0),
          legend.position=c(0.7,0.7),
          panel.background = theme_rect(fill="grey95",colour=NA),
          plot.title=theme_text(size=15,colour="black",hjust=0.5),
          axis.text.x = theme_text(),
          axis.title.x=theme_text(size=15,face="bold"))
#Save output to "png" file
png(filename = "hist_clin.png", width = 760, height = 760,
   units = "px", pointsize = 12, bg = "white", res = 150,
   restoreConsole = TRUE)
dev.off()
```

Appendix G v. The function 'val.prob.ci'

This generates a set of calibration plots and summary validation statistics, which are written to a table and can be inserted into an Excel spreadsheet (Vergouwe and Steyerberg, 2009). This is a modification of the val.prob function but with 95% confidence intervals for observed events in the validation sample.

```
#load libraries
library(ggplot2)
library(epitools)
library(rms)
library(reshape2)
library(grid)
library(PredictABEL)
library(digest)
library(proto)
#val.prob.ci
clin_model_df<-as.data.frame(val.prob.ci(model, y, pl = T, smooth = T,
xlim = c(0, 0.35), ylim = c(-0.05, 0.3), legendloc =  c(0.15 , 0.05),
statloc = c(0.01,0.25), dostats=c(12,13,2,15,3),roundstats=2,
logistic.cal = T, xlab=xlabel_title,g=10,  emax.lim=c(0,1), d0lab="0",
d1lab="1", cex = 0.75, mkh = 0.02, connect.group =
      F, connect.smooth = T,  cex.d01=0.8, dist.label=0.04,
line.bins=-.01, dist.label2=.03, cutoff, cex.lab=1, las=1,
length.seg=1.5))
write.table(clin_model_df,"clipboard",sep="\t",col.names=NA)
```

Appendix G vi. The 'plotROC' function (Kundu et al., 2011)

This function plots ROC operator curves from predicted outcome and observed binary outcomes. Sample code for comparing two models is shown in the box below.

```
#shows ROC plots for 2 models
cOutcome<-my_data$mortnum
predrisk<-
cbind(my_data$clin_corr,my_data$clin__imp_corr,my_data$clin_int_corr)
labels<-c("clinical","clinical_imp","clinical & age*rcri")
plotitle<-c("'Clinical' model variations")
fileplot<-c("ROC_clinical")
plotROC(data=my_data, cOutcome=2, predrisk=predrisk,
labels=labels,plottitle=plotitle,fileplot=fileplot,plottype="png")
```

References

Abbas, S.M. and Hill, A.G. (2008) 'Systematic review of the literature for the use of oesophageal Doppler monitor for fluid replacement in major abdominal surgery', *Anaesthesia*, 63(1), pp. 44-51.

Abunasra, H., Lewis, S., Beggs, L., Duffy, J., Beggs, D. and Morgan, E. (2005) 'Predictors of operative death after oesophagectomy for carcinoma', *British Journal of Surgery*, 92(8), pp. 1029-1033.

Adam, D.J., Craig, S.R., Sang, C.T., Walker, W.S. and Cameron, E.W. (1996) 'Oesophagogastrectomy for carcinoma in patients under 50 years of age', *Journal of the Royal College of Surgeons of Edinburgh*, 41(6), pp. 371-3.

Al-Sarira, A.A., David, G., Willmott, S., Slavin, J.P., Deakin, M. and Corless, D.J. (2007) 'Oesophagectomy practice and outcomes in England', *British Journal of Surgery*, 94(5), pp. 585-591.

Alexiou, C., Beggs, D., Salama, F.D., Brackenbury, E.T. and Morgan, W.E. (1998) 'Surgery for esophageal cancer in elderly patients: the view from Nottingham', *Journal of Thoracic & Cardiovascular Surgery*, 116(4), pp. 545-53.

Alexiou, C., Khan, O., Onyeaka, P., Beggs, L., Morgan, E. and Beggs, D. (2005) 'Oesophagectomy for squamous cell carcinoma: Lessons from a decade of consecutive resections', *Interactive Cardiovascular and Thoracic Surgery*, 4(3), pp. 180-183.

Alexiou, C., Khan, O.A., Black, E., Field, M.L., Onyeaka, P., Beggs, L., Duffy, J.P. and Beggs, D.F. (2006) 'Survival after esophageal resection for carcinoma: the importance of the histologic cell type', *Annals of Thoracic Surgery*, 82(3), pp. 1073-7.

Alibakhshi, A., Aminian, A., Mirsharifi, R., Jahangiri, Y., Dashti, H. and Karimian, F. (2009) 'The effect of age on the outcome of esophageal cancer surgery', *Annals of Thoracic Medicine*, 4 (2), pp. 71-74.

Allareddy, V., Allareddy, V. and Konety, B.R. (2007) 'Specificity of procedure volume and in-hospital mortality association', *Annals of Surgery*, 246(1), pp. 135-9.

Allum, W.H., Blazeby, J.M., Griffin, S.M., Cunningham, D., Jankowski, J.A. and Wong, R. (2011) 'Guidelines for the management of oesophageal and gastric cancer', *Gut*, 60(11), pp. 1449-1472.

Altman, D.G. (2001) 'Systematic reviews in health care: systematic reviews of evaluations of prognostic variables', *British Medical Journal*, 323(7306), pp. 224 - 228.

Altman, D.G. and Lyman, G.H. (1998) 'Methodological challenges in the evaluation of prognostic factors in breast cancer', *Breast Cancer Research and Treatment*, 52, pp. 289-303.

Altman, D.G. and Riley, R.D. (2005) 'Primer: an evidence-based approach to prognostic markers', *Nature Reviews: Clinical Oncology*, pp. 466-471.

Altman, D.G. and Royston, P. (2000) 'What do we mean by validating a prognostic model?', *Statistics in Medicine*, 19(4), pp. 453-473.

Altman, D.G., Vergouwe, Y., Royston, P. and Moons, K.G.M. (2009) 'Prognosis and prognostic research: validating a prognostic model', *British Medical Journal*, 338, p. 1432.

Aragon, T. (2010) *epitools: Epidemiology Tools* (Version 0.5-6) [Computer program]. Available at: http://CRAN.R-project.org/package=epitools.

Armstrong, K., Weber, B., Ubel, P.A., Peters, N., Holmes, J. and Schwartz, J.S. (2005) 'Individualized Survival Curves Improve Satisfaction With Cancer Risk Management Decisions in Women With BRCA1/2 Mutations', *Journal of Clinical Oncology*, 23(36), pp. 9319-9328.

Assmann, S.F., Pocock, S.J., Enos, L.E. and Kasten, L.E. (2000) 'Subgroup analysis and other (mis)uses of baseline data in clinical trials', *The Lancet*, 355(9209), pp. 1064-1069.

Atkins, B.Z., Shah, A.S., Hutcheson, K.A., Mangum, J.H., Pappas, T.N., Harpole, D.H., Jr. and D'Amico, T.A. (2004) 'Reducing hospital morbidity and mortality following esophagectomy', *Annals of Thoracic Surgery*, 78(4), pp. 1170-6.

Auvinen, A., Hakama, M., Ala-Opas, M., Vornanen, T., Leppilahti, M., Salminen, P. and Tammela, T.L.J. (2004) 'A randomized trial of choice of treatment in prostate cancer: the effect of intervention on the treatment chosen', *British Journal of Urology International*, 93(1), pp. 52-56.

Avery, K.N.L., Metcalfe, C., Barham, C.P., Alderson, D., Falk, S.J. and Blazeby, J.M. (2007) 'Quality of life during potentially curative treatment for locally advanced oesophageal cancer', *British Journal of Surgery*, 94(11), pp. 1369-1376.

Aylin, P., Alves, B., Cook, A., Bennett, J., Bottle, A., Best, N., Catena, B. and Elliott, P. (1999) *Analysis of Hospital Episode Statistics for the Bristol Royal Infirmary Inquiry*. 1999, C.C. [Online]. Available at: http://www.bristol-inquiry.org.uk/Documents/hes_(Aylin).pdf.

Aylin, P., Bottle, A. and Majeed, M. (2007) 'Use of administrative data or clinical databases as predictors of risk of death in hospital: comparison of models', *British Medical Journal*, 334(7602), p. 1044.

Baba, Y., Haga, Y., Hiyoshi, Y., Imamura, Y., Nagai, Y., Yoshida, N., Hayashi, N., Toyama, E., Miyanari, N. and Baba, H. (2008) 'Estimation of Physiologic Ability and Surgical Stress (E-PASS system) in patients with esophageal squamous cell carcinoma undergoing resection', *Esophagus*, 5(2), pp. 81-86.

Bachmann, M.O., Alderson, D., Edwards, D., Wotton, S., Bedford, C., Peters, T.J. and Harvey, I.M. (2002) 'Cohort study in South and West England of the influence of specialization on the management and outcome of patients with oesophageal and gastric cancers', *British Journal of Surgery*, 89(7), pp. 914-22.

Bailey, S.H., Bull, D.A., Harpole, D.H., Rentz, J.J., Neumayer, L.A., Pappas, T.N., Daley, J., Henderson, W.G., Krasnicka, B. and Khuri, S.F. (2003) 'Outcomes after esophagectomy: a ten-year prospective cohort', *Annals of Thoracic Surgery*, 75(1), pp. 217-22; discussion 222.

Bartels, H., Stein, H.J. and Siewert, J.R. (1998) 'Preoperative risk analysis and postoperative mortality of oesophagectomy for resectable oesophageal cancer', *British Journal of Surgery*, 85(6), pp. 840-4.

Bartels, H., Stein, H.J. and Siewert, J.R. (2000) 'Risk analysis in esophageal surgery', *Recent Results in Cancer Research*, 155, pp. 89-96.

Bennett-Guerrero, E., Welsby, I., Dunn, T.J., Young, L.R., Wahl, T.A., Diers, T.L., Phillips-Bute, B.G., Newman, M.F. and Mythen, M.G. (1999) 'The Use of a Postoperative Morbidity Survey to Evaluate Patients with Prolonged Hospitalization After Routine, Moderate-Risk, Elective Surgery', *Anesthesia & Analgesia*, 89(2), p. 514.

Birkmeyer, J.D., Siewers, A.E., Finlayson, E.V. and et al (2002) 'Hospital volume and and operative mortality in the United States', *New England Journal of Medicine*, 346, pp. 1128-1137.

Birkmeyer, J.D., Stukel, T.A., Siewers, A.E. and et al (2003) 'Surgeon volume and operative mortality in the United States', *New England Journal of Medicine*, 349, pp. 2117-2127.

Blazeby, J.M., Conroy, T., Hammerlid, E., Fayers, P., Sezer, O., Koller, M., Arraras, J., Bottomley, A., Vickery, C.W., Etienne, P.L. and Alderson, D. (2003) 'Clinical and psychometric validation of an EORTC questionnaire module, the EORTC QLQ-OES18, to assess quality of life in patients with oesophageal cancer', *European Journal of Cancer*, 39(10), pp. 1384-1394.

Blazeby, J.M., Farndon, J.R., Donovan, J. and Alderson, D. (2000) 'A prospective longitudinal study examining the quality of life of patients with esophageal carcinoma', *Cancer*, 88(8), pp. 1781-1787.

Blazeby, J.M., Kavadas, V., Vickery, C.W., Greenwood, R., Berrisford, R.G. and Alderson, D. (2005a) 'A Prospective Comparison of Quality of Life Measures for Patients with Esophageal Cancer', *Quality of Life Research*, 14(2), pp. 387-393.

Blazeby, J.M., Metcalfe, C., Nicklin, J., Barham, C.P., Donovan, J. and Alderson, D. (2005b) 'Association between quality of life scores and short-term outcome after surgery for cancer of the oesophagus or gastric cardia', *British Journal of Surgery*, 92(12), pp. 1502-1507.

Boersma, E., Kertai, M.D., Schouten, O., Bax, J.J., Noordzij, P., Steyerberg, E.W., Schinkel, A.F.L., van Santen, M., Simoons, M.L., Thomson, I.R., Klein, J., van Urk, H. and Poldermans, D. (2005) 'Perioperative cardiovascular mortality in noncardiac surgery: Validation of the Lee cardiac risk index', *The American Journal of Medicine*, 118(10), pp. 1134-1141.

Bonavina, L., Incarbone, R., Saino, G., Clesi, P. and Peracchia, A. (2003) 'Clinical outcome and survival after esophagectomy for carcinoma in elderly patients', *Diseases of the Esophagus*, 16(2), pp. 90-3.

Bouwmeester, W., Zuithoff, N.P.A., Mallett, S., Geerlings, M.I., Vergouwe, Y., Steyerberg, E.W., Altman, D.G. and Moons, K.G.M. (2012) 'Reporting and Methods in Clinical Prediction Research: A Systematic Review', *PLoS Medicine*, 9(5), p. e1001221.

Braiteh, F., Correa, A.M., Hofstetter, W.L., Rice, D.C., Vaporciyan, A.A., Walsh, G.L., Roth, J.A., Mehran, R.J., Swisher, S.G. and Ajani, J.A. (2009) 'Association of age and survival in patients with gastroesophageal cancer undergoing surgery with or without preoperative therapy', *Cancer*, 115 (19), pp. 4450-4458.

Brundage, M.D., Feldman-Stewart, D., Cosby, R., Gregg, R., Dixon, P., Youssef, Y., Davies, D. and Mackillop, W.J. (2001) 'Phase I Study of a Decision Aid for Patients With Locally Advanced Non–Small-Cell Lung Cancer', *Journal of Clinical Oncology*, 19(5), pp. 1326-1335.

Butler, C.W., Snyder, M., Wood, D.E., Curtis, J.R., Albert, R.K. and Benditt, J.O. (2001) 'Underestimation of Mortality Following Lung Volume Reduction Surgery Resulting From Incomplete Follow-up*', *Chest*, 119(4), pp. 1056-1060.

Cancer Research UK *Oesophageal cancer - UK incidence statistics*. Available at: http://info.cancerresearchuk.org/cancerstats/types/oesophagus/incidence/ (Accessed: 17/01/2012).

Cant, S. (2005) 'Briefing about data protection and medical research', *Data Protection and Medical Research*. Available at: http://www.parliament.uk/documents/post/postpn235.pdf.

Care Quality Commission Society for Cardiothoracic Surgery in Great Britain & Ireland (2010) *Heart surgery in the United Kingdom*. Available at: http: // heartsurgery.cqc.org.uk. (Accessed: 15/12/2011).

Cariati, A., Casano, A., Campagna, A., Cariati, E. and Pescio, G. (2002) 'Prognostic factors influencing morbidity and mortality in esophageal carcinoma', *Revista do Hospital das Clinicas; Faculdade de Medicina Da Universidade de Sao Paulo*, 57(5), pp. 201-4.

Carpenter, J., Bartlett, J. and Kenward, M. *www.missingdata.org.uk* Available at: www.missingdata.org.uk (Accessed: October 2011).

Center for Evidence Based Medicine (2012) 'generates sensitivity/probability plots from data ', *CEBM Statistics Calculator* Available at: http://www.cebm.utoronto.ca/practise/ca/statscal (Accessed: June 2012).

Centre for Reviews and Dissemination, York University (2009) *CRD's Guidance for undertaking reviews in health care: Chapter 2.3, Prognostic Tests*. Available at: http://www.york.ac.uk/inst/crd/SysRev/!SSL!/WebHelp/SysRev3.htm.

Chamogeorgakis, T., Anagnostopoulos, C.E., Connery, C.P., Ashton, R.C., Dosios, T., Kostopanagiotou, G., Rokkas, C.K. and Toumpoulis, I.K. (2007)

'Independent predictors for early and midterm mortality after thoracic surgery', *Thoracic and Cardiovascular Surgeon*, 55(6), pp. 380-384.

Charlson, M., Pompei, P. and MacKenzie, C. (1987) 'A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. ', *Journal of Chronic Diseases*, 40, pp. 373-383.

Charoenpan, P., Vathesatokit, P., Kiatboonsri, S., Saenghiranvattana, S., Aursudkit, B. and Vongvivat, K. (1993) 'Preoperative pulmonary evaluation: which parameters determine the surgical outcome', *Journal of the Medical Association of Thailand*, 76(8), pp. 429-35.

Clark, T.G. and Altman, D.G. (2003) 'Developing a prognostic model in the presence of missing data: an ovarian cancer case study', *Journal of Clinical Epidemiology*, 56(1), pp. 28-37.

Cleveland, W.S. (1979) 'Robust Locally Weighted Regression and Smoothing Scatterplots', *Journal of the American Statistical Association*, 74(368), pp. 829-836.

Cochrane Information Management System (2011) *Revman* (Version 5.1) [Computer program]. Cochrane Information Management System,. Available at: http://ims.cochrane.org/news/revman-51-released.

Cochrane Prognosis Methods Group (2011) *Cochrane Prognosis Methods Group*. Available at: http://prognosismethods.cochrane.org.

Collins, G.S., Jibawi, A. and McCulloch, P. (2011) 'Control chart methods for monitoring surgical performance: A case study from gastro-oesophageal surgery', *European Journal of Surgical Oncology (EJSO)*, 37(6), pp. 473-480.

Collins, G.S. and Moons, K.G.M. (2012) 'Comparing risk prediction models', *BMJ*, 344, p. 8.

Commissioner, T.I. (2002) *The Use and Disclosure of Health Data. Guidance on the Application of the Data Protection Act 1998*. Available at:

http://www.ico.gov.uk/upload/documents/library/data_protection/practical_application/health_data_-_use_and_disclosure001.pdf.

Copas, J.B. (1983) 'Regression, Prediction and Shrinkage', *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(3), pp. 311-354.

Copeland, P., Jones, D. and Walters, M. (1991) 'POSSUM:a scoring system for surgical audit', *British Journal of Surgery*, 78(3), pp. 355-360.

Coulter, A. (2010) 'Do patients want a choice and does it work?', *British Medical Journal*, 341, p. 4989.

Crawley, M.J. (2007) *The R Book*. Chichester, England: John Wiley & Sons, Ltd.

Cromwell, D., Palser, T., van der Meulen, J., Hardwick, R., Riley, S., Greenaway, K. and Dean, S. (2010) *National Oesophago-gastric Cancer Audit 2010* (17/07/2011). Centre, T.N.I. [Online]. Available at: www.augis.org/pdf/NHS-IC-OGC-Audit-2010-interactive.pdf.

Deans, C. and Patterson-Brown, S. (2009) 'Staging of oesophageal and gastric cancer', in Griffin, S.M. and Raimes, S.A. (eds.) *Oesophagogastric Surgery*. Fourth edn. London: Saunders Elsevier, pp. 41-68.

Deeks, J.J. (2001) 'Systematic reviews in health care: systematic reviews of evaluations of diagnostic and screening tests', *British Medical Journal*, 323, pp. 157-162.

Delbanco, T., Berwick, D.M., Boufford, J.I., Edgman, L., Ollenschläger, G., Plamping, D. and Rockefeller, R.G. (2001) 'Healthcare in a land called PeoplePower: nothing about me without me', *Health Expectations*, 4(3), pp. 144-150.

Delgado-Rodriguez, M. and Llorca, J. (2004) 'Bias', *J Epidemiol Community Health*, 58(8), pp. 635-641.

Department of Health (2012) *Liberating the NHS: No decision about me, without me. Further consultation on proposals to secure shared decision-making.*HM Government, Department of Health.

Di Martino, N., Izzo, G., Cosenza, A., Cerullo, G., Torelli, F., Brillantino, A. and del Genio, A. (2005) 'Adenocarcinoma of gastric cardia in the elderly: Surgical problems and prognostic factors', *World Journal of Gastroenterology*, 11(33), pp. 5123-5128.

Dimick, J.B., Goodney, P.P., Orringer, M.B. and Birkmeyer, J.D. (2005a) 'Specialty training and mortality after esophageal cancer resection', *Annals of Thoracic Surgery*, 80(1), pp. 282-286.

Dimick, J.B., Wainess, R.M., Upchurch, G.R., Jr., Iannettoni, M.D. and Orringer, M.B. (2005b) 'National trends in outcomes for esophageal resection', *Annals of Thoracic Surgery*, 79(1), pp. 212-6.

Dimick, J.B., Welch, H.G. and Birkmeyer, J.D. (2004) 'Surgical mortality as an indicator of hospital quality: the problem with small sample size.[see comment]', *Journal of the American Medical Association*, 292(7), pp. 847-51.

Dreiseitl, S. and Ohno-Machado, L. (2002) 'Logistic regression and artificial neural network classification models: a methodology review', *Journal of Biomedical Informatics*, 35(5-6), pp. 352-359.

Eagle, K.A., Rihal, C.S., Mickel, M.C., Holmes, D.R., Foster, E.D., Gersh, B.J. and Investigators for the Cass Program, U.o.M.H.C., , (1997) 'Cardiac Risk of Noncardiac Surgery : Influence of Coronary Disease and Type of Surgery in 3368 Operations', *Circulation*, 96(6), pp. 1882-1887.

Earlam, R. and Cunha-Melo, J.R. (1980) 'Oesophageal squamous cell carcinoma: I. A critical review of surgery', *British Journal of Surgery*, 67, pp. 381-390.

Ellis Jr, F.H., Williamson, W.A. and Heatley, G.J. (1998) 'Cancer of the esophagus and cardia: does age influence treatment selection and surgical outcomes?', *Journal of the American College of Surgeons*, 187(4), pp. 345-351.

Fang, W., Igaki, H., Tachimori, Y., Sato, H., Daiko, H. and Kato, H. (2001) 'Three-field lymph node dissection for esophageal cancer in elderly patients over 70 years of age.[see comment]', *Annals of Thoracic Surgery*, 72(3), pp. 867-71.

Fekete, K. and Belghiti, J. (1988) 'Nutritional factors and oesophageal resection', in Jamieson, G.E. (ed.) *Surgery of the Oesophagus,* . Edinburgh: Churchill Livingstone, pp. 119-124.

Ferguson, M.K. and Durkin, A.E. (2002) 'Preoperative prediction of the risk of pulmonary complications after esophagectomy for cancer', *Journal of Thoracic & Cardiovascular Surgery*, 123(4), pp. 661-9.

Ferguson, M.K., Martin, T.R., Reeder, L.B. and Olak, J. (1997) 'Mortality after esophagectomy: risk factor analysis', *World Journal of Surgery*, 21(6), pp. 599-603; discussion 603-4.

Field, A. (2000) *Discovering Statistics using SPSS for Windows*. Trowbridge, Great Britain: The Cromwell Press.

Finlayson, E., Fan, Z. and Birkmeyer, J.D. (2007) 'Outcomes in Octogenarians Undergoing High-Risk Cancer Operation: A National Study', *Journal of the American College of Surgeons*, 205(6), pp. 729-734.

Finlayson, E.V.A. and Birkmeyer, J.D. (2001) 'Operative Mortality with Elective Surgery in Older Adults', *Effective Clinical Practice*, 4(4), pp. 172-7.

Fleisher, L.A., Beckman, J.A., Brown, K.A., Calkins, H., Chaikof, E.L., Fleischmann, K.E., Freeman, W.K., Froehlich, J.B., Kasper, E.K., Kersten, J.R., Riegel, B. and Robb, J.F. (2007) 'ACC/AHA 2007 Guidelines on Perioperative Cardiovascular Evaluation and Care for Noncardiac Surgery: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Writing Committee to Revise the 2002 Guidelines on Perioperative Cardiovascular Evaluation for Noncardiac Surgery)', *Circulation*, 116(17), pp. e418-500.

Ford, M.K., Beattie, W.S. and Wijeysundera, D.N. (2010) 'Systematic Review: Prediction of Perioperative Cardiac Complications and Mortality by the Revised Cardiac Risk Index', *Annals of Internal Medicine*, 152(1), pp. 26-35.

Forshaw, M.J., Strauss, D.C., Davies, A.R., Wilson, D., Lams, B., Pearce, A., Botha, A.J. and Mason, R.C. (2008) 'Is Cardiopulmonary Exercise Testing a Useful Test Before Esophagectomy?', *The Annals of Thoracic Surgery*, 85(1), pp. 294-299.

Gabriel, R., Alonso, M., Reviriego, B., Muniz, J., Vega, S., Lopez, I., Novella, B., Suarez, C. and Rodriguez-Salvanes, F. (2009) 'Ten-year fatal and non-fatal myocardial infarction incidence in elderly populations in Spain: the EPICARDIAN cohort study', *BMC Public Health*, 9, p. 360.

Gockel, I., Exner, C. and Junginger, T. (2005) 'Morbidity and mortality after esophagectomy for esophageal carcinoma: A risk analysis', *World Journal of Surgical Oncology*, 3(37).

Golubovi, V. and Golubovi, S. (2002) 'ASA score as prognostic criterion for incidence of postoperative complications after transhiatal esophagectomy', *Collegium Antropologicum*, 26 Suppl, pp. 149-53.

Griffin, S.M. (2009) 'Surgery for cancer of the oesophagus', in Griffin, S.M. and Raimes, S.A. (eds.) *Oesophagogastric Surgery*. Fourth edn. Saunders Elsevier.

Griffin, S.M., Shaw, I.H. and Dresner, S.M. (2002) 'Early complications after Ivor Lewis subtotal esophagectomy with two-field lymphadenectomy: risk factors and management', *Journal of the American College of Surgeons*, 194(3), pp. 285-97.

Grimshaw, J., McAuley, L.M., Bero, L.A., Grilli, R., Oxman, A.D., Ramsay, C., Vale, L. and Zwarenstein, M. (2003) 'Systematic reviews of the effectiveness of quality improvement strategies and programmes', *Quality and Safety in Health Care*, 12(4), pp. 298-303.

Grocott, M.P.W., Browne, J.P., Van der Meulen, J., Matejowsky, C., Mutch, M., Hamilton, M.A., Levett, D.Z.H., Emberton, M., Haddad, F.S. and Mythen, M.G. (2007) 'The Postoperative Morbidity Survey was validated and used to describe morbidity after major surgery', *Journal of Clinical Epidemiology*, 60(9), pp. 919-928.

Gulliford, M.C., Barton, J.R. and Bourne, H.M. (1993) 'Selection for oesophagectomy and postoperative outcome in a defined population', *Quality in Health Care*, 2(1), pp. 17-20.

Gupta, R. and Ihmaidat, H. (2003) 'Nutritional effects of oesophageal, gastric and pancreatic carcinoma', *European Journal of Surgical Oncology*, 29(8), pp. 634-643.

Han-Geurts, I.J.M., Hop, W.C., Tran, T.C.K. and Tilanus, H.W. (2006) 'Nutritional status as a risk factor in esophageal surgery', *Digestive Surgery*, 23(3), pp. 159-63.

Hanley, J.A. and McNeil, B.J. (1982) 'The meaning and use of the area under a receiver operator characteristic (ROC) curve', *Radiology*, 143(1), pp. 29-36.

Harrell, F.E. (2001a) 'Binary Logistic Regression', in *Regression Modeling Strategies With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York, USA: Springer, p. pp. 219.

Harrell, F.E. (2001b) 'Introduction', in *Regression Modeling Strategies With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York, USA: Springer, p. pp. 5.

Harrell, F.E. (2001c) 'Multivariable Modelling Strategies', in *Regression Modeling Strategies With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York, USA: Springer, pp. 60-64.

Harrell, F.E. (2001d) 'Multivariable Modelling Strategies', in *Regression Modeling Strategies With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York, USA: Springer, p. pp. 56.

Harrell, F.E. (2001e) *Regression Modeling Strategies With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York, USA: Springer.

Harrell, F.E.J. (2012) *Validate Predicted Probabilities (val.prob)* (Version 3.5-0) [Computer program]. Available at: http://cran.r-project.org/web/packages/rms/rms.pdf.

Harrell Jr, F.E. (24/03/2011) *Regression modelling strategies* (Version 3.5-0) [Computer program]. Available at: http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/RmS.

Hayashida, K., Morice, M.-C., Chevalier, B., Hovasse, T., Romano, M., Garot, P., Farge, A., Donzeau-Gouge, P., Bouvier, E., Cormier, B. and Lefevre, T. (2012) 'Sex-related differences in clinical presentation and outcome of transcatheter aortic valve implantation for severe aortic stenosis', *Journal of the American College of Cardiology*, 59(6), pp. 566-71.

Hayden, J.A., Cote, P. and Bombardier, C. (2006) 'Evaluation of the Quality of Prognosis Studies in Systematic Reviews', *Annals of Internal Medicine*, 144(6), pp. 427-437.

Healy, L.A., Ryan, A.M., Moore, J., Rowley, S., Ravi, N., Byrne, P.J. and Reynolds, J.V. (2008) 'Health-related quality of life assessment at presentation may predict complications and early relapse in patients with localized cancer of the esophagus', *Diseases of the Esophagus*, 21(6), pp. 522-528.

Hernández, A.V., Steyerberg, E.W. and Habbema, J.D.F. (2004) 'Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements', *Journal of Clinical Epidemiology*, 57(5), pp. 454-460.

Hofstetter, W., Swisher, S.G., Correa, A.M., Hess, K., Putnam, J.B., Jr., Ajani, J.A., Dolormente, M., Francisco, R., Komaki, R.R., Lara, A., Martin, F., Rice, D.C., Sarabia, A.J., Smythe, W.R., Vaporciyan, A.A., Walsh, G.L.

and Roth, J.A. (2002) 'Treatment outcomes of resected esophageal cancer', *Annals of Surgery*, 236(3), pp. 376-84.

Howell, D.C. (3/7/2009) 'Description of methods for handling missing data', *Treatment of missing data*. Available at: http://www.uvm.edu/~dhowell/StatPages/More_Stuff/Missing_Data/Missing.html.

Iverson, A., Lidell, K., Fear, N., Hotopf, M. and Wessely, S. (2006) 'Consent, confidentiality, and the Data Protection Act', *British Medical Journal*, 332, pp. 165-169.

Jamieson, G.G., Mathew, G., Ludemann, R., Wayman, J., Myers, J.C. and Devitt, P.G. (2004) 'Postoperative mortality following oesophagectomy and problems in reporting its rate', *British Journal of Surgery*, 91(8), pp. 943-947.

Jiao, W.-J., Wang, T.-Y., Gong, M., Pan, H., Liu, Y.-B. and Liu, Z.-H. (2006) 'Pulmonary complications in patients with chronic obstructive pulmonary disease following transthoracic esophagectomy', *World Journal of Gastroenterology*, 12(16), pp. 2505-9.

Johansson, J. and Walther, B. (2000) 'Clinical outcome and long-term survival rates after esophagectomy are not determined by age over 70 years', *Journal of Gastrointestinal Surgery*, 4(1), pp. 55-62.

Jougon, J.B., Ballester, M., Duffy, J., Dubrez, J., Delaisement, C., Velly, J.-F. and Couraud, L. (1997) 'Esophagectomy for cancer in the patient aged 70 years and older', *The Annals of Thoracic Surgery*, 63(5), pp. 1423-1427.

Justice, A.C., Covinsky, K.E. and Berlin, J.A. (1999) 'Assessing the Generalizability of Prognostic Information', *Annals of Internal Medicine*, 130(6), pp. 515-524.

Karl, R.C.M.D., Schreiber, R.M.D., Boulware, D.M.S.M.H.S., Baker, S.M.A. and Coppola, D.M.D. (2000) 'Factors Affecting Morbidity, Mortality, and

Survival in Patients Undergoing Ivor Lewis Esophagogastrectomy', *Annals of Surgery*, 231(5), pp. 635-643.

Karnofsky, D. (1984) 'Reporting results of cancer treatment', *Cancer*, 1, pp. 634-635.

Khuri, S.F., Henderson, W.G., DePalma, R.G., Mosca, C., Healey, N.A., Kumbhani, D.J. and and the Participants in the VA National Surgical Quality Improvement Program (2005) 'Determinants of Long-Term Survival After Major Surgery and the Adverse Effect of Postoperative Complications', *Annals of Surgery*, 242(3), pp. 326-343.

Killeen, S.D., O'Sullivan, M.J., Coffey, J.C., Kirwan, W.O. and Redmond, H.P. (2005) 'Provider volume and outcomes for oncological procedures', *British Journal of Surgery*, 92(4), pp. 389-402.

Kinugasa, S., Tachibana, M., Yoshimura, H., Dhar, D.K., Shibakita, M., Ohno, S., Kubota, H., Masunaga, R. and Nagasue, N. (2001) 'Esophageal resection in elderly esophageal carcinoma patients: Improvement in postoperative complications', *Annals of Thoracic Surgery*, 71(2), pp. 414-418.

Kleinbaum, D.G. (1994) *Logistic Regression A self-Learning Text*. New York: Springer.

Kristman, V., Manno, M. and Cote, P. (2004) 'Loss to Follow-Up in Cohort Studies: How Much Is Too Much?', *European Journal of Epidemiology*, 19(8), pp. 751-760.

Kundu, S., Aulchenko, Y.S. and Janssens, A.C.J.W. (2011) *PredictABEL: Assessment of risk prediction models* [Computer program]. Available at: http://CRAN.R-project.org/package=PredictABEL.

Kuwano, H., Sumiyoshi, K., Sonoda, K., Kitamura, K., Tsutsui, S., Toh, Y., Kitamura, M. and Sugimachi, K. (1998) 'Relationship between preoperative assessment of organ function and postoperative morbidity in patients with oesophageal cancer', *European Journal of Surgery*, 164(8), pp. 581-6.

L'Italien, G.J., Paul, S.D., Hendel, R.C., Leppo, J.A., Cohen, M.C., Fleusher, L.A., Brown, K.A., Zarich, S.W., Cambria, R.P., Cutler, B.S. and Eagle, K.A. (1996) 'Development and validation of a bayesian model for perioperative cardiac risk assessment in a cohort of 1,081 vascular surgical candidates', *Journal of the American College of Cardiology*, 27(4), pp. 779-786.

Lagarde, S.M., Maris, A.K.D., de Castro, S., Busch, O.R.C., Obertop, H. and van Lanschot, J.J.B. (2007) 'Evaluation of O-POSSUM in predicting in-hospital mortality after resection for oesophageal cancer', *British Journal of Surgery*, 94(12), pp. 1521-1526.

Lagarde, S.M., Reitsma, J.B., Maris, A.K.D., van Berge Henegouwen, M.I., Busch, O.R.C., Obertop, H., Zwinderman, A.H. and van Lanschot, J.J.B. (2008) 'Preoperative Prediction of the Occurrence and Severity of Complications After Esophagectomy for Cancer With Use of a Nomogram', *Annals of Thoracic Surgery*, 85(6), pp. 1938-1945.

Lai, F., Kwan, T.K., Yuen, W.C., Wai, A. and Shung, Y.C.S.E. (2007) 'Evaluation of various POSSUM models for predicting mortality in patients undergoing elective oesophagectomy for carcinoma', *British Journal of Surgery*, 94(9), pp. 1172-1178.

LaPar, D.J., Mulloy, D.P., Crosby, I.K., Lim, D.S., Kern, J.A., Kron, I.L. and Ailawadi, G. (2012) 'Contemporary outcomes for surgical mitral valve repair: a benchmark for evaluating emerging mitral valve technology', *Journal of Thoracic & Cardiovascular Surgery*, 143(4 Suppl), pp. S12-6.

Law, D., Dudrick, S. and Abdou, N. (1973) ' Immune competence of patients with protein-calorie malnutrition. The effects of nutrition repletion', *Annals of Internal Medicine* 95, pp. 545-550.

Law, S., Wong, K.-H., Kwok, K.-F., Chu, K.-M. and Wong, J. (2004) 'Predictive factors for postoperative pulmonary complications and mortality after esophagectomy for cancer', *Annals of Surgery*, 240(5), pp. 791-800.

Law, S.Y.K., Fok, M. and Wong, J. (1994) 'Risk analysis in resection of squamous cell carcinoma of the esophagus', *World Journal of Surgery*, 18(3), pp. 339-346.

Lee, T.H., Marcantonio, E.R., Mangione, C.M., Thomas, E.J., Polanczyk, C.A., Cook, E.F., Sugarbaker, D.J., Donaldson, M.C., Poss, R., Ho, K.K.L., Ludwig, L.E., Pedan, A. and Goldman, L. (1999) 'Derivation and Prospective Validation of a Simple Index for Prediction of Cardiac Risk of Major Noncardiac Surgery', *Circulation*, 100(10), pp. 1043-1049.

Leigh, Y., Seagroatt, V., Goldacre, M. and McCulloch, P. (2006) 'Impact of socio-economic deprivation on death rates after surgery for upper gastrointestinal tract cancer', *British Journal of Cancer*, 95(7), pp. 940-3.

Lemeshow, S. and Hosmer, D.W.J. (1982) 'A review of goddness of fit statistics for use in the development of logistic regression models', *American Journal Epidemiology*, 115(1), pp. 92-106.

Liedman, B.L., Bennegard, K., Olbe, L.C. and Lundell, L.R. (1995) 'Predictors of postoperative morbidity and mortality after surgery for gastro-oesophageal carcinomas', *European Journal of Surgery*, 161(3), pp. 173-80.

Lilford, R., Mohammed, M.A., Spiegelhalter, D. and Thomson, R. (2004) 'Use and misuse of process and outcome data in managing performance of acute medical care: avoiding institutional stigma', *The Lancet*, 363(9415), pp. 1147-1154.

Little, R.J.A. (1992) 'Regression With Missing X's: A Review', *Journal of the American Statistical Association*, 87(420), pp. 1227-1237.

Liu, J.F., Watson, D.I., Devitt, P.G., Mathew, G., Myburgh, J. and Jamieson, G.G. (2000) 'Risk factor analysis of post-operative mortality in oesophagectomy', *Diseases of the Esophagus*, 13(2), pp. 130-5.

Look, M.P., van Putten, W.L.J., Duffy, M.J., Harbeck, N., Christensen, I.J., Thomssen, C., Kates, R., Spyratos, F., Fernö, M., Eppenberger-Castori, S., Sweep, C.G.J.F., Ulm, K., Peyrat, J.-P., Martin, P.-M., Magdelenat, H., Brünner, N., Duggan, C., Lisboa, B.W., Bendahl, P.-O., Quillien, V., Daver, A., Ricolleau, G., Meijer-van Gelder, M.E., Manders, P., Fiets, W.E., Blankenstein, M.A., Broët, P., Romain, S., Daxenbichler, G., Windbichler, G., Cufer, T., Borstnar, S., Kueng, W., Beex, L.V.A.M., Klijn, J.G.M., O'Higgins, N., Eppenberger, U., Jänicke, F., Schmitt, M. and Foekens, J.A. (2002) 'Pooled Analysis of Prognostic Impact of Urokinase-Type Plasminogen Activator and Its Inhibitor PAI-1 in 8377 Breast Cancer Patients', *Journal of the National Cancer Institute*, 94(2), pp. 116-128.

Lund, O., Kimose, H.H., Aagaard, M.T., Hasenkam, J.M. and Erlandsen, M. (1990) 'Risk stratification and long-term results after surgical treatment of carcinomas of the thoracic esophagus and cardia. A 25-year retrospective study', *Journal of Thoracic & Cardiovascular Surgery*, 99(2), pp. 200-9.

Mangano, D.T. (1990) 'Perioperative cardiac morbidity', *Anesthesiology*, 72, pp. 153-84.

McCulloch, P., Ward, J., Tekkis, P.P., Ascot group of surgeons and British Oesophago-Gastric Cancer Group (2003) 'Mortality and morbidity in gastro-oesophageal cancer surgery: initial results of ASCOT multicentre prospective cohort study', *British Medical Journal*, 327(7425), pp. 1192-7.

Migliore, M., Choong, C.K., Lim, E., Goldsmith, K.A., Ritchie, A. and Wells, F.C. (2007) 'A surgeon's case volume of oesophagectomy for cancer strongly influences the operative mortality rate', *European Journal Cardiothoracic Surgery*, 32(2), pp. 375-380.

Mohammed, A.M. and Andrew, S. (2007) 'The value of administrative databases', *British Medical Journal*, 334(7602), pp. 1014-1015.

Mohammed, M.A., Deeks, J.J., Girling, A., Rudge, G., Carmalt, M., Stevens, A.J. and Lilford, R.J. (2009) 'Evidence of methodological bias in hospital standardised mortality ratios: retrospective database study of English hospitals', *Btitish Medical Journal*, 338.

Moher, D., Liberati, A., Tetzlaff, J. and Altman, D.G. (2009) 'Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement', *Journal of Clinical Epidemiology*, 62(10), pp. 1006-1012.

Mokart, D., Leone, M., Sannini, A., Brun, J.P., Tison, A., Delpero, J.R., Houvenaeghel, G., Blache, J.L. and Martin, C. (2005) 'Predictive perioperative factors for developing severe sepsis after major surgery', *British Journal of Anaesthesia*, 95(6), pp. 776-781.

Moonesinghe, S.R., Mythen, M.G. and Grocott, M.P.W. (2011) 'High risk surgery: epidemiology and outcomes', *Anesthesia and analgesia*, 112(4), pp. 891-901.

Moons, K.G.M., Altman, D.G., Vergouwe, Y. and Royston, P. (2009a) 'Prognosis and prognostic research: application and impact of prognostic models in clinical practice', *British Medical Journal*, 338.

Moons, K.G.M., Royston, P., Vergouwe, Y., Grobbee, D.E. and Altman, D.G. (2009b) 'Prognosis and prognostic research: what, why, and how?', *British Medical Journal*, 338.

Morgan, M., Deber, R., Llewellyn-Thomas, H., Gladstone, P., Cusimano, R., O'Rourke, K., Tomlinson, G. and Detsky, A. (2000) 'Randomized, controlled trial of an interactive videodisc decision aid for patients with ischemic heart disease', *Journal of General Internal Medicine*, 15(10), pp. 685-693.

Morgan, M.A., Lewis, W.G., Hopper, A.N., Escofet, X., Harvard, T.J., Brewster, A.E., Crosby, T.D.L., Roberts, S.A. and Clark, G.W.B. (2007) 'Prognostic significance of body mass indices for patients undergoing esophagectomy for cancer', *Diseases of the Esophagus*, 20(1), pp. 29-35.

Moskovitz, A.H., Rizk, N.P., Venkatraman, E., Bains, M.S., Flores, R.M., Park, B.J.H. and Rusch, V.W. (2006) 'Mortality increases for octogenarians undergoing esophagogastrectomy for esophageal cancer', *Annals of Thoracic Surgery*, 82(6), pp. 2031-6; discussion 2036.

Muller, J.M., Erasmi, H., Stelzner, M., Zieren, U. and Pichlmaier, H. (1990) 'Surgical therapy of oesophageal carcinoma', *British Journal of Surgery*, 77, pp. 845-857.

Murray, P., Whiting, P., Hutchinson, S.P., Ackroyd, R., Stoddard, C.J. and Billings, C. (2007) 'Preoperative shuttle walking testing and outcome after oesophagogastrectomy', *British Journal of Anaesthesia*, 99(6), pp. 809-811.

Nagabhushan, J.S., Srinath, S., Weir, F., Angerson, W.J., Sugden, B.A. and Morran, C.G. (2007) 'Comparison of P-POSSUM and O-POSSUM in predicting mortality after oesophagogastric resections', *Postgraduate Medical Journal*, 83(979), pp. 355-8.

National Cancer Institute (2012) *Surveillance Epidemiology and End Results*. Available at: http://seer.cancer.gov/.

National Clinical Guideline Centre (2010) *Chronic obstructive pulmonary disease: management of chronic obstructive pulmonary disease in adults in primary and seconday care*. London: National Clinical Guideline Centre. [Online]. Available at: http://guidance.nice.org.uk/CG101/Guidance/pdf/English (Accessed: November 2 2011).

Nixon, D.W., Heymsfield, S.B., Cohen, A.E., Kutner, M.H., Ansley, J., Lawson, D.H. and Rudman, D. (1980) 'Protein-calorie undernutrition in hospitalized cancer patients', *The American Journal of Medicine*, 68(5), pp. 683-690.

Northern and Yorkshire Cancer Registry and Information Service (2012) *Northern and Yorkshire Cancer Registry and Information Service*. Available at: http://www.nycris.nhs.uk.

Nozoe, T., Kimura, Y., Ishida, M., Saeki, H., Korenaga, D. and Sugimachi, K. (2002) 'Correlation of pre-operative nutritional condition with post-operative complications in surgical treatment for oesophageal carcinoma', *European Journal of Surgical Oncology*, 28(4), pp. 396-400.

O'Connor, A.M., Rostom, A., Fiset, V., Tetroe, J., Entwistle, V., Llewellyn-Thomas, H., Holmes-Rovner, M., Barry, M. and Jones, J. (1999) 'Decision aids for patients facing health treatment or screening decisions: systematic review', *British Medical Journal*, 319(7212), pp. 731-734.

O'Day, D.K. (2011) *Learn R Toolkit*. Available at: http://processtrends.com/Learn_R_Toolkit.htm.

Ochsner, J.L. and DeBakey, M. (1941) 'Surgical aspects of carcinoma of the esophagus;review of the literature and report of 4 cases', *Journal of Thoracic Surgery*, 10, pp. 401-445.

Oken, M.M., Creech, R., Tormey, D. and et al. (1982) 'Toxicity and response criteria of the Eastern Cooperative Oncology Group', *American Journal of Clinical Oncology*, 5(6), pp. 649-655.

Older, P., Hall, A. and Hader, R. (1999) 'Cardiopulmonary Exercise Testing as a Screening Test for Perioperative Management of Major Surgery in the Elderly', *Chest*, 116(2), pp. 355-362.

Park, D.P., Welch, C.A., Harrison, D.A., Palser, T.R., Cromwell, D.A., Gao, F., Alderson, D., Rowan, K.M. and Perkins, G.D. (2009) 'Outcomes following oesophagectomy in patients with oesophageal cancer: A secondary analysis of the ICNARC Case Mix Programme Database', *Critical Care*, 13 (suppl. 2)(S1).

Peduzzi, P., Concato, J., Kemper, E., Holford, T.R. and Feinstein, A.R. (1996) 'A simulation study of the number of events per variable in logistic regression analysis', *Journal of Clinical Epidemiology*, 49(12), pp. 1373-1379.

Peek, N., Arts, D.G.T., Bosman, R.J., van der Voort, P.H.J. and de Keizer, N.F. (2007) 'External validation of prognostic models for critically ill patients required substantial sample sizes', *Journal of Clinical Epidemiology*, 60(5), pp. 491.e1-491.e13.

Pennefather, S.H. (2007) 'Anaesthesia for oesphagectomy', *Current Opinion in Anaesthesiology*, 20, pp. 15-20.

Pepe, M.S. (2005) 'Evaluating technologies for classification and prediction in medicine', *Statistics in Medicine*, 24(24), pp. 3687-3696.

Pezzullo, J.C. (Updated May 2009) 'Interactive statistical webpage', *Javastat-binomial proportions*. Available at: http://statpages.org/proppowr.html (Accessed: July 2009).

Poldermans, D., Bax, J.J., Boersma, E., De Hert, S., Eeckhout, E., Fowkes, G., Gorenek, B., Hennerici, M.G., Iung, B., Kelm, M., Kjeldsen, K.P., Kristensen, S.D., Lopez-Sendon, J., Pelosi, P., Philippe, F., Pierard, L., Ponikowski, P., Schmid, J.-P., Sellevold, O.F.M., Sicari, R. and Van den Berge, G. (2009) 'Guidelines for pre-operative cardiac risk assessment and perioperative cardiac management in non-cardiac surgery', *European Heart Journal*, 30, p. 2769.

Poon, R.T.P., Law, S.Y.K., Chu, K.M., Branicki, F.J.D.M. and Wong, J. (1998) 'Esophagectomy for Carcinoma of the Esophagus in the Elderly: Results of Current Surgical Management', *Annals of Surgery*, 227(3), pp. 357-364.

Pope, C., Mays, N. and Popay, J. (2007) *Synthesising Qualitative and Quantitative Health Evidence; A guide to methods*. McGraw Hill, Open University Press.

Powell, J., McConkey, C.C., Gillison, E.W. and Spychal, R.T. (2002) 'Continuing rising trend in oesophageal adenocarcinoma', *International Journal of Cancer*, 102(4), pp. 422-427.

Priebe, H.J. (2000) 'The aged cardiovascular risk patient', *British Journal of Anaesthesia*, 85(5), pp. 763-778.

Prytherch, D.R., Whiteley, M.S., Higgins, B., Weaver, P.C., Prout, W.G. and Powell, S.J. (1998) 'POSSUM and Portsmouth POSSUM for predicting mortality', *British Journal of Surgery*, 85(9), pp. 1217-1220.

Qaseem, A., Snow, V., Fitterman, N., Hornbake, E.R., Lawrence, V.A., Smetana, G.W., Weiss, K. and Owens, D.K. (2006) 'Risk Assessment for and Strategies To Reduce Perioperative Pulmonary Complications for Patients

Undergoing Noncardiothoracic Surgery: A Guideline from the American College of Physicians', *Annals of Internal Medicine*, 144(8), pp. 575-580.

Quanjer, P.H., Tammeling, G.J., Cotes, J.E., Pederson, O.F., Peslin, R. and Yernault, J.C. (1993) 'Lung volumes and forced ventilatory flows. Report Working Party Standardization of Lung Function Tests, European Community for Steel and Coal. Official Statement of the European Respiratory Society', *European Respiratory Journal Supplement* 16, pp. 5-40.

R Development Core Team (2011) *R: A Language and Environment for Statistical Computing* (Version R version 2.13.0 (2011-04-13)) [Computer program]. R Foundation for Statistical Computing. Available at: http://www.R-project.org.

Ra, J., Paulson, E.C., Kucharczuk, J., Armstrong, K., Wirtalla, C., Rapaport-Kelz, R. and et al (2008) 'Postoperative mortality after esophagectomy for cancer: Development of a risk prediction model', *Annals of Surgical Oncology*, 15(6), pp. 1577-1584.

Rahamim, J.S., Murphy, G.J., Awan, Y. and Junemann-Ramirez, M. (2003) 'The effect of age on the outcome of surgical treatment for carcinoma of the oesophagus and gastric cardia', *European Journal of Cardio-thoracic Surgery*, 23(5), pp. 805-810.

Rentz, J., Bull, D., Harpole, D., Bailey, S., Neumayer, L., Pappas, T., Krasnicka, B., Henderson, W., Daley, J. and Khuri, S. (2003) 'Transthoracic versus transhiatal esophagectomy: a prospective study of 945 patients', *Journal of Thoracic & Cardiovascular Surgery*, 125(5), pp. 1114-20.

Riley, R.D., Lambert, P. and Abo-Zaid, G. (2010) 'Meta-analysis of individual participant data: rationale, conduct, and reporting', *British Medical Journal*, 340, p. 521.

Roques, F., Nashef, S.A.M., Michel, P., Gauducheau, E., Vincentiis, C.d., Baudet, E., Cortina, J., David, M., Faichney, A., Gabrielle, F., Gams, E., Harjula, A., Jones, M.T., Pinna Pintor, P., Salamon, R. and Thulin, L. (1999)

'Risk factors and outcome in European cardiac surgery: analysis of the EuroSCORE multinational database of 19030 patients', *European Journal of Cardio-thoracic Surgery*, 15, pp. 816-823.

Rothstein, H.R., Sutton, J.A. and Borenstein, M. (eds.) (2005) *Publication Bias in Metanalysis*. John Wiley and Sons, Ltd.

Rouvelas, I., Zeng, W., Lindblad, M., Viklund, P., Ye, W. and Lagergren, J. (2005) 'Survival after surgery for oesophageal cancer: A population-based study', *Lancet Oncology*, 6(11), pp. 864-870.

Royston, P., Moons, K.G.M., Altman, D.G. and Vergouwe, Y. (2009) 'Prognosis and prognostic research: Developing a prognostic model', *British Medical Journal*, 338.

Ruol, A., Portale, G., Castoro, C., Merigliano, S., Cagol, M., Cavallin, F., Chiarion Sileni, V., Corti, L., Rampado, S., Costantini, M. and Ancona, E. (2007(b)) 'Effects of neoadjuvant therapy on perioperative morbidity in elderly patients undergoing esophagectomy for esophageal cancer', *Annals of Surgical Oncology*, 14(11), pp. 3243-3250.

Ruol, A., Portale, G., Zaninotto, G., Cagol, M., Cavallin, F., Castoro, C., Sileni, V.C., Alfieri, R., Rampado, S. and Ancona, E. (2007(a)) 'Results of esophagectomy for esophageal cancer in elderly patients: age has little influence on outcome and survival', *Journal of Thoracic & Cardiovascular Surgery*, 133(5), pp. 1186-92.

Sabel, M.S., Smith, J.L., Nava, H.R., Mollen, K., Douglass, H.O. and Gibbs, J.F. (2002) 'Esophageal resection for carcinoma in patients older than 70 years', *Annals of Surgical Oncology*, 9(2), pp. 210-4.

Sackett, D.L., Haynes, B.R., Guyatt, G.H. and Tugwell, P. (1991) 'The interpretation of diagnostic data', in *Clinical Epidemiology: A Basic Science for Clinical medicine*. 2nd edn. Boston/Toronto/London: Little, Brown and Company, p. pp. 92.

Sackett, D.L., Haynes, B.R., Guyatt, G.H. and Tugwell, P. (2006) 'Determining prognosis and creating clinical prediction rules', in *Clinical Epidemiology: How to do Clinical Practice Research*. Sixth edn. United States: Lippincott Williams and Wilkins.

Saito, T., Shimoda, K., Kinoshita, T., Shigemitsu, Y., Miyahara, M., Kobayashi, M. and Shimaoka, A. (1993) 'Prediction of operative mortality based on impairment of host defense systems in patients with esophageal cancer', *Journal of Surgical Oncology*, 52(1), pp. 1-8.

Saklad, M. (1941) 'Grading of Patients for Surgical Procedures', *Anesthesiology*, 2(3), pp. 281-284.

Sanz, L., Ovejero, V.J., Gonzalez, J.J., Laso, C.A., Azcano, E., Navarrete, F. and Martinez, E. (2006) 'Mortality risk scales in esophagectomy for cancer: Their usefulness in preoperative patient selection', *Hepato-Gastroenterology*, 53(72), pp. 869-873.

Sargent, D.J. (2001) 'Comparison of Artificial Neural Networks with Other Statistical Approaches', *Cancer*, 91(Supplement 8), pp. 1636-1642.

Sauvanet, A., Mariette, C., Thomas, P., Lozac'h, P., Segol, P., Tiret, E., Delpero, J.-R., Collet, D., Leborgne, J., Pradere, B., Bourgeon, A. and Triboulet, J.-P. (2005) 'Mortality and morbidity after resection for adenocarcinoma of the gastroesophageal junction: predictive factors', *Journal of the American College of Surgeons*, 201(2), pp. 253-62.

Schiesser, M., Chen, J.W.C., Maddern, G.J. and Padbury, R.T.A. (2008) 'Perioperative Morbidity Affects Long-Term Survival in Patients Following Liver Resection for Colorectal Metastases ', *Journal of Gastrointestinal Surgery*, 12(6), pp. 1054-60.

Schroder, W., Bollschweiler, E., Kossow, C. and Holscher, A.H. (2006) 'Preoperative risk analysis - A reliable predictor of postoperative outcome after transthoracic esophagectomy?', *Langenbeck's Archives of Surgery*, 391(5), pp. 455-460.

Secretary of State for Health (2010) *Equity and excellence: Liberating the NHS*. United Kingdom: The Stationery Office Limited, United Kingdom. [Online]. Available at: http://www.official-documents.gov.uk/document/cm78/7881/7881.pdf.

Shaheen, N.J., Crosby, M.A., Bozymski, E.M. and Sandler, R.S. (2000) 'Is There Publication Bias in the Reporting of Cancer Risk in Barrett's Esophagus?', *Gastroenterology*, 119(2), pp. 333-338.

Shahian, D.M., Normand, S.-L., Torchiana, D.F., Lewis, S.M., Pastore, J.O., Kuntz, R.E. and Dreyer, P.I. (2001) 'Cardiac surgery report cards: comprehensive review and statistical critique', *Annals of Thoracic Surgery*, 72(6), pp. 2155-2168.

Shaw, I.H. (2008) 'Anaesthetic aspects and case selection for oesophageal and gastric surgery ', in Griffin, S.M. and Raimes, S. (eds.) *Oesophagogastric Surgery: A Companion to Specialist Surgical Practice* 4th edn. W B Saunders Company Ltd.

Shende, M.R., Waxman, J. and Luketich, J.D. (2007) 'Predictive Ability of Preoperative Indices for Esophagectomy', *Thoracic Surgery Clinics*, 17(3), pp. 337-341.

Skipworth, J., Foster, J., Raptis, D. and Hughes, F. (2009) 'The effect of preoperative weight loss and body mass index on postoperative outcome in patients with esophagogastric carcinoma', *Diseases of the esophagus*, 22 (7), pp. 559-563.

Smetana, G.W., Lawrence, V.A. and Cornell, J.E. (2006) 'Preoperative Pulmonary Risk Stratification for Noncardiothoracic Surgery: Systematic Review for the American College of Physicians', *Annals of Internal Medicine*, 144(8), pp. 581-595.

Song, F., Eastwood, A.J., Gilbody, S., Duley, L. and Sutton, A.J. (2000) *Publication and related biases* (10). [Online]. Available at: http://www.hta.ac.uk/1051

Stacey, D., Bennett, C., Barry, M., Col, N., Eden, K., Holmes-Rovner, M., Llewellyn-Thomas, H., Lyddiatt, A., Légaré, F. and Thomson, R. (2011) 'Decision aids for people facing health treatment or screening decisions (Review)', *The Cochrane Library* (10).

Statistical Consulting Group (October, 2011) *FAQ: What are pseudo R-squareds?* Available at: http://www.ats.ucla.edu/stat/mult_pkg/faq/general/psuedo_rsquareds.htm (Accessed: 10/05/2012).

Steyerberg, E.W. (2009a) 'Application of prediction models', in Gail, M., Tsiatis, A., Krickeberg, K. and Sarnet, J. (eds.) *Clinical prediction models. A Practical Approach to development, Validation, and Updating*. New York: Springer.

Steyerberg, E.W. (2009b) 'Applications of prediction models', in Gail, M., Tsiatis, A., Krickeberg, K. and Sarnet, J. (eds.) *Clinical prediction models. A Practical Approach to development, Validation, and Updating*. New York: Springer, p. pp. 11.

Steyerberg, E.W. (2009c) 'Applications of prediction models', in Gail, M., Tsiatis, A., Krickeberg, K. and Sarnet, J. (eds.) *Clinical prediction models. A Practical Approach to development, Validation, and Updating*. New York: Springer, p. pp. 24.

Steyerberg, E.W. (2009d) 'Choosing between alternative statistical models', in Gail, M., Tsiatis, A., Krickeberg, K. and Sarnet, J. (eds.) *Clinical prediction models. A Practical Approach to development, Validation, and Updating*. New York: Springer, p. pp. 105.

Steyerberg, E.W. (2009e) *Clinical prediction models. A Practical Approach to development, Validation, and Updating*. New York: Springer.

Steyerberg, E.W. (2009f) 'Dealing with Missing Values', in Gail, M., Tsiatis, A., Krickeberg, K. and Sarnet, J. (eds.) *Clinical prediction models. A*

Practical Approach to development, Validation, and Updating. New York: Springer, pp. 115-136.

Steyerberg, E.W. (2009g) 'Estimation with external information', in *Clinical Prediction Models. A Practical Approach to Development, Validation, and Updating*. New York: Springer, p. pp. 243.

Steyerberg, E.W. (2009h) 'Evaluation of performance', in Gail, M., Tsiatis, A., Krickeberg, K. and Sarnet, J. (eds.) *Clinical prediction models. A Practical Approach to development, Validation, and Updating*. New York: Springer, p. pp. 274.

Steyerberg, E.W. (2009i) 'Introduction', in Gail, M., Tsiatis, A., Krickeberg, K. and Sarnet, J. (eds.) *Clinical prediction models. A Practical Approach to development, Validation, and Updating*. New York: Springer, p. pp. 5.

Steyerberg, E.W. (2009j) 'Modern estimation methods', in Gail, M., Tsiatis, A., Krickeberg, K. and Sarnet, J. (eds.) *Clinical prediction models. A Practical Approach to development, Validation, and Updating*. New York: Springer, p. pp. 233.

Steyerberg, E.W. (2009k) 'Overfitting and optimism in regression models', in Gail, M., Tsiatis, A., Krickeberg, K. and Sarnet, J. (eds.) *Clinical prediction models. A Practical Approach to development, Validation, and Updating*. New York: Springer, p. pp. 87.

Steyerberg, E.W. (2009l) 'Statistical models for Prediction', in *Clinical Prediction Models. A Practical Approach to Development, Validation, and Updating*. New York: Springer, p. pp. 39.

Steyerberg, E.W. (2009m) 'Study design for prediction models', in Gail, M., Tsiatis, A., Krickeberg, K. and Sarnet, J. (eds.) *Clinical prediction models. A Practical Approach to development, Validation, and Updating*. New York: Springer, p. pp. 39.

Steyerberg, E.W. (2009n) 'Study design for prediction models', in Gail, M., Tsiatis, A., Krickeberg, K. and Sarnet, J. (eds.) *Clinical prediction models. A*

Practical Approach to development, Validation, and Updating. New York: Springer, p. pp. 50.

Steyerberg, E.W. (2009o) 'Validation of prediction models', in Gail, M., Tsiatis, A., Krickeberg, K. and Sarnet, J. (eds.) *Clinical prediction models. A Practical Approach to development, Validation, and Updating*. New York: Springer, p. pp. 301.

Steyerberg, E.W., Borsboom, G.J.J.M., van Houwelingen, H.C., Eijkemans, M.J.C. and Habbema, J.D.F. (2004) 'Validation and updating of predictive logistic regression models: a study on sample size and shrinkage', *Statistics in Medicine*, 23(16), pp. 2567-2586.

Steyerberg, E.W., Eijkemans, M.J.C. and Habbema, J.D.F. (1999) 'Stepwise Selection in Small Data Sets: A Simulation Study of Bias in Logistic Regression Analysis', *Journal of Clinical Epidemiology*, 52(10), pp. 935-942.

Steyerberg, E.W., Eijkemans, M.J.C., Harrell, F.E. and Habbema, J.D.F. (2001a) 'Prognostic Modeling with Logistic Regression Analysis', *Medical Decision Making*, 21(1), pp. 45-56.

Steyerberg, E.W., Eijkemans, M.J.C., Harrell, F.E. and Habbema, J.D.F. (2001b) 'Prognostic Modeling with Logistic Regression Analysis: In Search of a Sensible Strategy in Small Data Sets', *Medical Decision Making*, 21(1), pp. 45-56.

Steyerberg, E.W., Eijkemans, M.J.C., Van Houwelingen, J.C., Lee, K.L. and Habbema, J.D.F. (2000) 'Prognostic models based on literature and individual patient data in logistic regression analysis', *Statistics in Medicine*, 19(2), pp. 141-160.

Steyerberg, E.W., Neville, B.A., Koppert, L.B., Lemmens, V.E.P.P., Tilanus, H.W., Coebergh, J.-W.W., Weeks, J.C. and Earle, C.C. (2006) 'Surgical mortality in patients with esophageal cancer: development and validation of a simple risk score', *Journal of Clinical Oncology*, 24(26), pp. 4277-84.

Steyerberg, E.W., Vickers, A.J., Cook, N.R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M.J. and Kattane, M.W. (2010) 'Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures', *Epidemiology*, 21, pp. 128-138.

Stroup, D.F., Berlin, J.A., Morton, S.C., Olkin, I., Williamson, G.D., Rennie, D., Moher, D., Becker, B.J., Sipe, T.A., Thacker, S.B. and for the Meta-analysis Of Observational Studies in Epidemiology, G. (2000) 'Meta-analysis of Observational Studies in Epidemiology: A Proposal for Reporting', *Journal of the American Medical Association*, 283(15), pp. 2008-2012.

Takagawa, R., Kunisaki, C., Makino, H., Oshima, T., Nagano, Y., Fujii, S., Kosaka, T., Ono, H.A., Akiyama, H. and Shimada, H. (2008) 'Therapeutic management of elderly patients with esophageal cancer', *Esophagus*, 5(3), pp. 133-139.

Takeno, S., Takahashi, Y., Watanabe, S., Ono, K., Kamei, M., Yamashita, S.I. and Kawahara, K. (2008) 'Esophagectomy in patients aged over 80 years with esophageal cancer', *Hepato-Gastroenterology*, 55(82-83), pp. 453-456.

Tekkis, P.P., McCulloch, P., Poloniecki, J.D., Prytherch, D.R., Kessaris, N. and Steger, A.C. (2004) 'Risk-adjusted prediction of operative mortality in oesophagogastric surgery with O-POSSUM', *British Journal of Surgery*, 91(3), pp. 288-295.

The Health and Social Care Information Centre, N. (2012) 'Hospital Episode Statistics'. Available at: http://www.hesonline.nhs.uk/Ease/servlet/ContentServer?siteID=1937&categoryID=456.

The POISE Study Group (2008) 'Effects of extended-release metoprolol succinate in patients undergoing non-cardiac surgery (POISE trial): a randomised controlled trial', *The Lancet*, 371(9627), pp. 1839-1847.

The Society for Cardiothoracic Surgery in Great Britain & Ireland (2011) 'website of cardiothoracic surgery society', *Maintaining patients' trust:*

modern medical professionalism 2011. Available at: http://www.scts.org/_userfiles/resources/634420268996790965_SCTS_Professionalism_FINAL.pdf (Accessed: 10/12/2011).

Thomas, P., Doddoli, C., Neville, P., Pons, J., Lienne, P., Giudicelli, R., Giovannini, M., Seitz, J.F. and Fuentes, P. (1996) 'Esophageal cancer resection in the elderly', *European Journal of Cardiothoracic Surgery*, 10(11), pp. 941-946.

Tsai, C.-H., Hsu, H.-S., Wang, L.-S., Wang, H.-W., Wu, Y.-C., Hsieh, C.-C., Huang, B.-S., Hsu, W.-H. and Huang, M.-H. (2003) 'Surgical results of squamous cell carcinoma of the esophagus in young patients', *Journal of the Chinese Medical Association*, 66(5), pp. 288-93.

Tu, J.V. (1996) 'Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes', *Journal of Clinical Epidemiology*, 49(11), pp. 1225-1231.

Vach, W. (1997) 'Some issues in estimating the effect of prognostic factors from incomplete covariate information', *Statistics in Medicine*, 16, pp. 57-72.

van Buuren, S., Boshuizen, H.C. and Knook, D.L. (1999) 'Multiple imputation of missing blood pressure covariates in survival analysis', *Statistics in Medicine*, 18(6), pp. 681-694.

Vergouwe, Y. and Steyerberg, E.W. (2009) *val.prob.ci* [Computer program]. Available at: http://www.clinicalpredictionmodels.org/.

Vergouwe, Y., Steyerberg, E.W., Eijkemans, M.J.C. and Habbema, J.D.F. (2005) 'Substantial effective sample sizes were required for external validation studies of predictive logistic regression models', *Journal of Clinical Epidemiology*, 58(5), pp. 475-483.

Verhoef, C., Van De Weyer, R., Schaapveld, M., Bastiaannet, E. and Plukker, J.T.M. (2007) 'Better survival in patients with esophageal cancer after surgical treatment in university hospitals: A plea for performance by surgical oncologists', *Annals of Surgical Oncology*, 14(5), pp. 1678-1687.

Vickers, A.J. and Elkin, E.B. (2006) 'Decision Curve Analysis: A Novel Method for Evaluating Prediction Models', *Medical Decision Making*, 26(6), pp. 565-574.

Vizcaino, A.P., Moreno, V., Lambert, R. and Parkin, D.M. (2002) 'Time trends incidence of both major histologic types of esophageal carcinomas in selected countries, 1973–1995', *International Journal of Cancer*, 99(6), pp. 860-868.

Waljee, J.F., Rogers, M.A.M. and Alderman, A.K. (2007) 'Decision Aids and Breast Cancer: Do They Influence Choice for Surgery and Knowledge of Treatment Options?', *Journal of Clinical Oncology*, 25(9), pp. 1067-1073.

Wallace, E., Smith, S., Perera-Salazar, R., Vaucher, P., McCowan, C., Collins, G., Verbakel, J., Lakhanpaul, M. and Fahey, T. (2011) 'Framework for the impact analysis and implementation of Clinical Prediction Rules (CPRs)', *BMC Medical Informatics and Decision Making*, 11(1), p. 62.

Warnell, I. (2009) *Reported causes of perioperative mortality in the NOGCU dataset*. Clinical audit. Northern Oesophago-Gastric Cancer Unit.

Warnell, I. (2012) The function val.prob.ci, May 2012.

Westaby, S., Archer, N., Manning, N., Adwani, S., Grebenik, C., Ormerod, O., Pillai, R. and Wilson, N. (2007) 'Comparison of hospital episode statistics and central cardiac audit database in public reporting of congenital heart surgery mortality', *British Medical Journal*, 335(7623), p. 759.

Whooley, B.P.M.D., Law, S., Murthy, S.C.M.D., Alexandrou, A. and Wong, J. (2001) 'Analysis of Reduced Death and Complication Rates After Esophageal Resection', *Annals of Surgery*, 233(3), pp. 338-344.

Wickham, H. (2009) *ggplot2: elegant graphics for data analysis*. New York: Springer.

Wild, C.P. and Hardie, L.J. (2003) 'Reflux, Barrett's oesophagus and adenocarcinoma: burning questions', *National Review of Cancer*, 3(9), pp. 676-684.

Windsor, J.A. and Hill, G.L. (1988) 'Risk factors for postoperative pneumonia. The importance of protein depletion', *Annals of surgery*, 208(2), p. 209.

Wright, C.D., Kucharczuk, J.C., O'Brien, S.M., Grab, J.D. and Allen, M.S. (2009) 'Predictors of major morbidity and mortality after esophagectomy for esophageal cancer: A Society of Thoracic Surgeons General Thoracic Surgery Database risk adjustment model', *Journal of Thoracic and Cardiovascular Surgery*, 137 (3), pp. 587-596.

Wu, P.C. and Posner, M.C. (2003) 'The role of surgery in the management of oesophageal cancer', *The Lancet Oncology*, 4(8), pp. 481-488.

Zafirellis, K.D., Fountoulakis, A., Dolan, K., Dexter, S.P.L., Martin, I.G. and Sue-Ling, H.M. (2002) 'Evaluation of POSSUM in patients with oesophageal cancer undergoing resection', *British Journal of Surgery*, 89(9), pp. 1150-1155.

Zhang, G.H., Fujita, H., Yamana, H. and Kakegawa, T. (1994) 'A prediction of hospital mortality after surgical treatment for esophageal cancer', *Surgery Today*, 24(2), pp. 122-127.

Zingg, U., Langton, C., Addison, B., Wijnhoven, B.P.L., Forberger, J., Thompson, S.K., Esterman, A.J. and Watson, D.I. (2009) 'Risk Prediction Scores for Postoperative Mortality After Esophagectomy', *Journal of Gastrointestinal Surgery*, 13, pp. 611-618.