Statistical Analysis of Breast Cancer: The Effects of Missing Values on Survival Data

by

Catherine S Thomson

Thesis submitted for the degree of M.Sc.

to the Faculty of Science,

University of Glasgow, 1999

ABSTRACT

Worldwide, breast cancer is the most common cancer in women. In Scotland, there are currently over 3,000 women diagnosed with the disease each year and the incidence continues to rise. Despite some major advances in the treatment of breast cancer, with the discovery of tamoxifen and the on-going development of cytotoxic drugs, only 60% of women are still alive after five years, many of whom have a relapse at some later stage.

Against that background, the main aims of this thesis are to interpret the findings of a survival analysis of cases of breast cancer in Scotland; to investigate whether the method of including extra categories for unknown values in factors in that analysis is appropriate; and to check whether the assumption of proportional hazards is valid. Chapter 1 provides a general introduction, whilst Chapter 2 examines the burden that breast cancer places on the National Health Service in Scotland and throughout the world. The risk factors for getting the disease and the different strategies available for treatment of the cancer are also presented.

To identify how women with breast cancer in Scotland were managed, Chapter 3 outlines some background to a national retrospective audit of all cases of invasive breast cancer in the years 1987 and 1993. Analyses of a subgroup of the 1987 cohort constitute the majority of this thesis.

Chapter 4 examines the associations among the variables included in the survival analysis. The patterns among the missing values in four of the prognostic factors are also investigated, using log-linear modelling. The method employed in analysing this cohort of women was to create extra categories to represent unknown values in each of the factors. Other techniques available for handling missing values in models are discussed, along with a summary of the methods used in other relevant studies of breast cancer survival.

Chapter 5 presents a survival analysis of the cohort, including a discussion of the findings in relation to other relevant studies. Model checking is performed on the best fit model to assess the adequacy of the fit of it and to validate the assumption of proportional hazards. The remainder of the chapter focuses on a comparison of the results from fitting Cox models using the additional categories and the complete cases methods. This investigates whether different interpretations would be concluded from these models.

In Chapter 6, simulated datasets are generated using exponential distributions to investigate whether the proportional hazards assumption is valid when additional categories are used to extend two factors at two levels to three levels in an exponential regression model. The extent of any biases for the parameter estimates is examined.

Chapter 7 provides a summary of the key conclusions and highlights areas of future research.

ACKNOWLEDGEMENTS

I would like to thank the following people:

- Professor Ian Ford for his support and advice;
- clinical colleagues: Dr David Brewster, Professor Robin Leake and Dr Chris Twelves for helping to improve my knowledge of breast cancer;
- all those work colleagues at ISD, Scotland and the Robertson Centre for Biostatistics who have given me their support and encouragement through the two years of study;

and finally,

- all of my friends and family, especially my Mum, Dad, Gillian and Paul, for their support and patience, but mainly for enduring my intermittent moans!

TABLE OF CONTENTS

SECTION A: BACKGROUND

CHAPTER 1 INTRODUCTION

CHAPTER 2 BREAST CANCER

CHAPTER 3 SURVEY OF BREAST CANCER IN SCOTTISH WOMEN IN 1987

SECTION B: ANALYSES

CHAPTER 4 DETERMINATION OF MISSING VALUES AND CHARACTERISATION OF VARIABLES

CHAPTER 5 SURVIVAL ANALYSES

CHAPTER 5 Continued SURVIVAL ANALYSES

CHAPTER 6 INVESTIGATIONS OF BIAS IN MODELS WITH ADDITIONAL CATEGORIES FOR MISSING VALUES

CHAPTER 6 Continued INVESTIGATIONS OF BIAS IN MODELS WITH ADDITIONAL CATEGORIES FOR MISSING VALUES

SECTION C: FINAL THOUGHTS

CHAPTER 7 DISCUSSION AND CONCLUSIONS

TABLE OF FIGURES

TABLE OF TABLES

TABLE OF APPENDICES

SECTION A :

BACKGROUND

CHAPTER 1 INTRODUCTION

1.1 SURVEY OF BREAST CANCER IN SCOTTISH WOMEN IN 1987 AND IN 1993

In 1996, the Scottish Breast Cancer Focus Group (SBCFG), the Scottish Cancer Trials Breast Group and the Scottish Cancer Therapy Network (SCTN) produced a report for the Chief Scientist and Clinical Resource and Audit Group in Scotland, entitled 'Scottish Breast Cancer Audit 1987 & 1993' (SBCFG et al, 1996), known here as the 'Audit Report'. This detailed the preliminary results of a national population-based study of all women diagnosed with invasive breast cancer in Scotland in these two years.

The main aims of the audit were:

- to identify how women diagnosed with invasive breast cancer in Scotland in 1987 and in 1993 were managed;
- to investigate whether there had been any changes in the patterns of care between the two study years, during which time a national breast screening programme had been introduced;
- to examine how many women were managed according to best practice;
- to identify the factors affecting various outcome measures for the 1987 cohort only.

The cohorts were based on all women registered with the national Scottish Cancer Registry (SCR) for these two years of diagnosis. However, rather than limit the analysis to data collected and held on the national register, specially trained Data Managers from the SCTN sought all of the case notes relating to the breast cancer for these women. They re-examined the contents of them and collected a large amount of supplementary data to augment that from the SCR records. This additional information relates to referral patterns; initial staging of the tumour at the clinic; the surgical procedures performed; other forms of treatment given; pathology details, including extra staging information; and follow-up and outcome details.

This dataset provides two national 'snap-shots' of the management of breast cancer in Scotland. The Audit Report gives analyses of the quality of the data collected; the effect of the breast screening programme; referral patterns, including surgical case load and time for referral between presentation and diagnosis; pathological information collected; the management of women who did not undergo any surgery; the management of women who had surgery, with and without radiotherapy; the use of systemic adjuvant treatment; survival of those women undergoing surgery in 1987; and finally entry into clinical trials for breast cancer.

Subsequent to the Audit Report, to date, three peer-reviewed papers (Twelves et al, 1998a; Twelves et al, 1998b; Dewar et al, 1999) and a letter to the BMJ (Twelves et al, 1999) based on the data collected in the Audit have been published. These relate to the survival from breast cancer of women undergoing surgery in the 1987 cohort; factors affecting clinical trial entry for breast cancer in Scotland for both cohorts; the increase in workload of oncologists due to increased use of radiotherapy and adjuvant systemic therapy between the two years; and factors which determined whether a woman moved out of her Health Board of residence for her surgery respectively.

1.2 AIMS OF THE MSC

The idea of the project for this thesis came from my heavy involvement in some of the analyses of the Breast Cancer Audit data, especially the survival analysis of the 1987 cohort; both for the preliminary report (SBCFG et al, 1996) and for the publication by Twelves et al (1998a).

One of the problems encountered during analysis of the retrospective Audit data was the large extent of unknown information for some variables. This was due to looking back at case notes, rather than collecting the information prospectively as women are diagnosed with breast cancer. For example, for some women in the 1987 cohort, the case notes were eight years old when they were examined. Whilst some of the case notes simply could not be found, some of the information was not available because it just had not been recorded. How to deal with the missing values in the survival analysis was an issue. For both publications, the decision was taken to include additional categories for the unknowns in each of the factors, so as to avoid throwing away a large number of cases and losing information about other variables for these women.

The first aim of this thesis is, therefore, to examine whether using these additional categories gives different results from those that are obtained when only those women with complete information are retained in the analysis. This will discuss whether different implications, in terms of political and organisational structures, could be drawn from the results.

From this initial aim, others follow naturally. These include looking at simple frequency distributions of the factors to determine the full extent of the missingness; investigating whether there are any patterns or associations between the missing values; and researching other possible techniques for handling missing data to try to identify any which could be applied to the Breast Cancer Audit data.

One final aim is to examine whether it is likely that including the additional categories for the unknowns in the survival analysis of the Audit data violates the assumption of proportional hazards imposed by fitting a Cox regression model. This will be researched using various randomly generated simulated datasets constructed from known theoretical distributions.

CHAPTER 2 BREAST CANCER

This chapter provides some background information about breast cancer - how common it is, what factors increase the chances a woman will get the disease, how it is usually treated and the survival chances for women who have breast cancer.

2.1 DESCRIPTIVE EPIDEMIOLOGY OF BREAST CANCER

Breast cancer is the most common cancer in women worldwide (Parkin et al,1993). In Scotland in 1995 there were 3,156 new cases registered with the national Scottish Cancer Registry, representing 26% of all malignant neoplasms in females (Figure 2.1). Current estimates suggest that 1 in 12 women in the UK will get breast cancer during their lifetime (Evans et al, 1994).

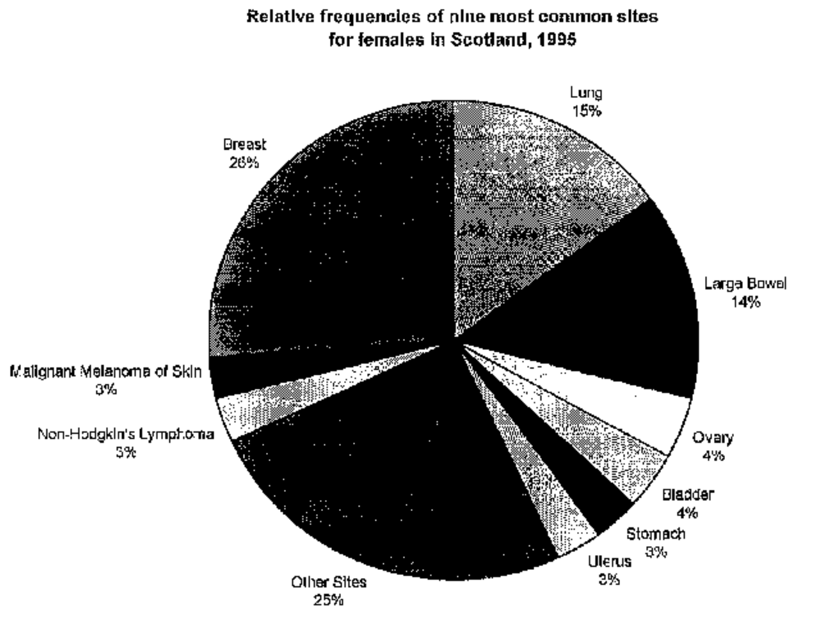


Figure 2.1: Relative frequencies of female cancers in Scotland in 1995.

2.1.1 CANCER REGISTRATION IN SCOTLAND

1947 - 1996

Information about new cases of cancer in Scotland has been collected on a national basis since 1947, although computer records only date back to 1958. The file was tumour-based with each new malignancy as the basis of a new record. The Scottish Cancer Registry, based at the Information & Statistics Division (ISD) of the NHS in Scotland (NHSiS), manages the data centrally.

Data were either collected manually using Scottish Morbidity Record 6 (SMR6) paper forms or electronically from various databases held independently within hospital or pathology departments. The data were sent to ISD via five regional registries. Basic information included name, date of birth and postcode of the patient; date of birth; the hospital where the cancer was registered; and the 'date treatment commenced'. There were also details about the site code of the cancer to 4 digits, based on the ninth revision of the International Classification of Diseases (ICD-9) (World Health Organisation, 1977); whether the patient had had previous tumours; whether the current diagnosis was histologically verified and if it was, the morphology code, based on the International Classification of Diseases for Oncology (ICD-O) (World Health Organisation, 1976).

Derived fields, such as Health Board of residence (based on postcode) and age at diagnosis, were then attached to the record. A regular record linkage using probabilistic matching (Kendrick & Clarke, 1993) between the cancer file and the death records supplied by the General Registers Office (GRO) enabled follow-up of cancer patients from diagnosis to death.

1997 ONWARDS

The SMR6 scheme was replaced by a fully computerised system, known as SOCRATES (Scottish Open Cancer Registration And Tumour Enumeration System), in 1997.

The aim of SOCRATES is to identify possible new cases of cancer from multiple sources of records. These include hospital discharge records (SMR1), GRO death records and pathology and oncology departmental records. SOCRATES links information obtained from the various sources and automatically creates a provisional cancer registration. The information held on SOCRATES is patient-based, rather than tumour-based.

After allowing six months for treatment details to accumulate, a trained Cancer Registration Officer (CRO) scrutinises medical case notes relating to the provisional registration. The registration is then confirmed or deleted if invalid. The CRO supplements the basic details with information about the management of the cancer including whether surgery was performed, chemotherapy or radiotherapy were given, and the initial stage of the tumour is recorded. For breast cancer, pathological information about axillary node status and the size of tumour are also noted. All of the additional data will be useful for clinical audit and will allow extra prognostic factors to be included in survival analyses. These factors were not available from the old SMR6 files.

The computer records from the SMR6 scheme were appended to SOCRATES to enable epidemiological studies of incidence and mortality over time. Therefore, the SOCRATES system has records dating back to 1958, now linked by patients rather than tumours.

The validity of performing epidemiological studies on the Scottish Cancer Registration database is supported by the long collection period, and also more importantly, because the information is widely recognised to be of a high standard in terms of both accuracy and completeness (Brewster et al, 1994; Brewster et al, 1997). In general, high accuracy is reflected by a high percentage of registrations based on tumours having a microscopic verification (%MV). The %MV for breast cancer in Scotland in 1995 was 89.6%.

Completeness of ascertainment describes the proportion of cases which are registered out of those which should have been registered. This can be indirectly assessed from the percentage of registrations made on the basis of a death certificate only (%DCO) which occurs when this is the only record supporting the diagnosis of cancer. For these records, the date of diagnosis entered onto the cancer registry file is the date of death from the death certificate. These DCO cases are usually excluded from survival analyses as they have zero survival time and provide no other details relating to the cancer. A high %DCO rate suggests that incidence rates may be underestimated. In 1995, the %DCO for breast cancer in Scotland was low at 2.8%, suggesting high completeness.

2.1.2 INCIDENCE OF BREAST CANCER

There are approximately 720,000 new cases of breast cancer in the world each year (Parkin et al, 1993) with 34,500 cases registered in 1991 in the UK (Cancer Research Campaign, 1996). The number of cases registered with the Scottish Cancer Registry for the years 1986-1995 are given in Table 2.1.

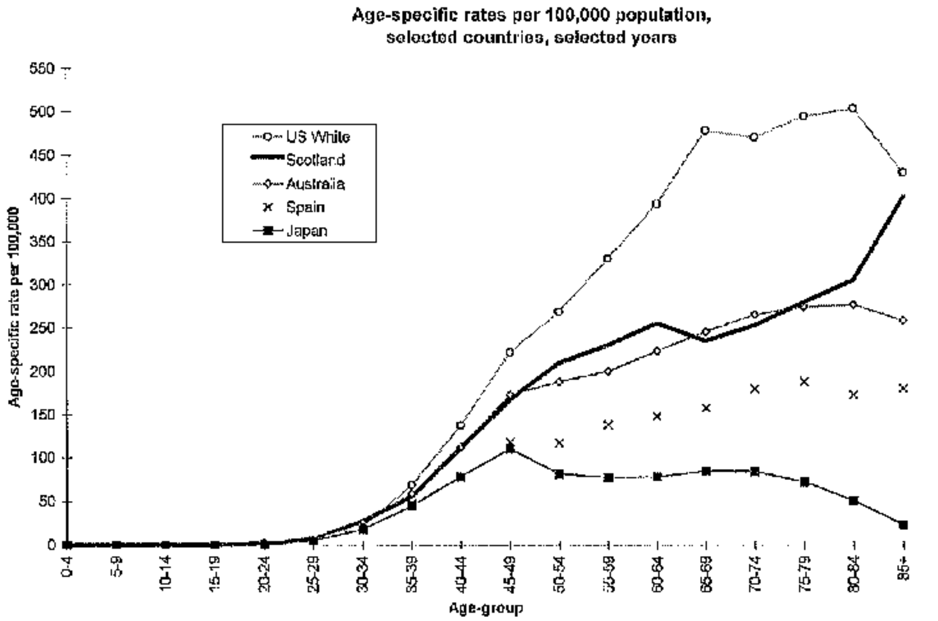
Year of Registration	Number of Women with Breast Cancer
1986	2617
1987	2684
1988	2680
1989	2775
1990	2969
1991	3171
1992	3233
1993	3110
1994	3071
1995	3156

Table 2.1: Numbers of cases registered in Scotland, 1986-1995.

This table shows an increasing trend in the annual number of reported cases with a marked jump around 1990-1992. Although the increase in numbers of cases may imply an increase in incidence, it is important to consider the population at risk and the rate of disease. To allow for population changes in age distribution over time, it is preferable to study age-specific rates (Boyle & Parkin, 1991; Sharp et al, 1993).

Age is the most important known risk factor for breast cancer, with the elderly, in general, being the high risk group (Henderson et al, 1996). The most dramatic increase for all countries is between the age-bands 30-34 to 50-54, between which the rate more than doubles. This can be seen in Figure 2.2 which gives the age-specific incidence rates per 100,000 population for breast cancer for Scotland and four other countries from the developed world.

For the Western World countries, the increase in the age-specific rate slows down for postmenopausal women but it is still present. In contrast, for Japan, the curve reaches a plateau and remains almost constant after the age of menopause. Hoel et al (1983) observed that the level of oestrogen in postmenopausal Japanese women is probably not very high as they have low body weight and not many excess fat cells (see Section 2.2.1).



US, San Francisco White: 88-92; Scotland 85-94; Australia, New South Wales: 88-92; Spain, Basque Country: 88-91; Japan, Miyagi: 88-92.

Figure 2.2: Age-specific rates for various countries for breast cancer (Parkin et al, 1997).

The increase in the number of registrations in 1990-1992 seen in Table 2.1 is primarily due to an increase in the age-specific rates for the screening age group 50-64 shown below in Figure 2.3 and is probably due to the introduction of the Scottish Breast Screening Programme. This was phased in throughout Scotland from 1988, attaining national coverage in 1994 (Scottish Breast Screening Programme Central Co-ordinating

Unit, 1997) and led to detection of a larger number of small early breast cancers. In 1995, 546 (17.6%) of cases of breast cancer were screen-detected.

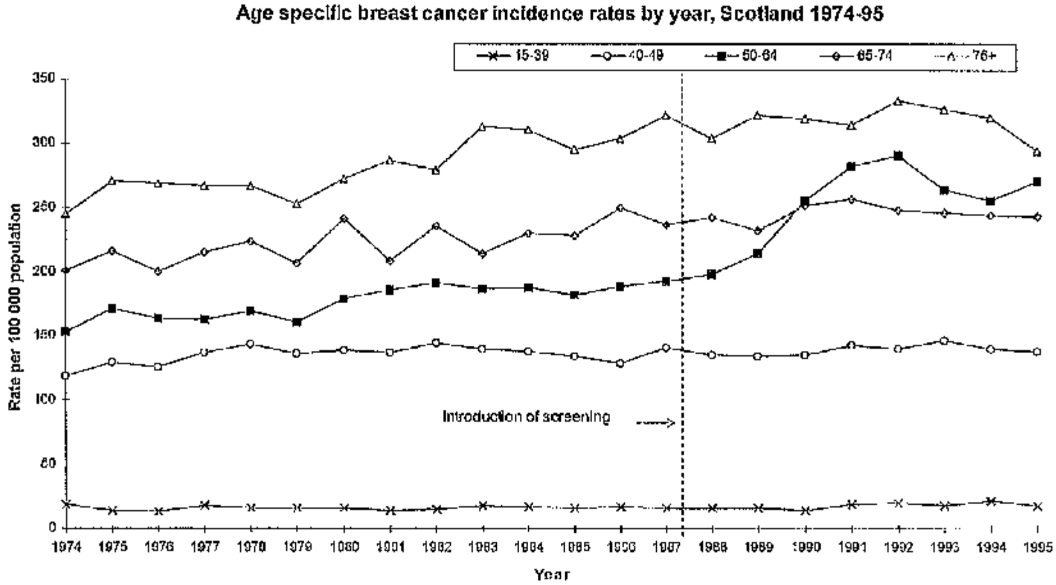


Figure 2.3: Age-specific rates in Scotland, 1974-1995, by age group (Scottish Breast Screening Programme Central Co-ordinating Unit, 1997).

Another important risk factor for breast cancer is the country in which the women grew up (Parkin et al, 1993), although this may in part be due to reproductive risk factors; for example, the age of menarche can be affected by temperature, climate and social welfare conditions in a country (see Section 2.2.2). To allow comparisons of incidence figures for countries with different age structures, the populations are usually standardised to an arbitrary standard population. Often this is either the World Standard Population (WSP) or the European Standard Population (ESP), both of which were first detailed by Doll et al (1966). Figure 2.4 shows the variation in World age-standardised rates (WASR) across several regional registries, countries and racial groups (Parkin et al, 1997) and it is clear that there is a six-fold variation in the rates, although the differences have been decreasing gradually over time (Lipworth, 1995).

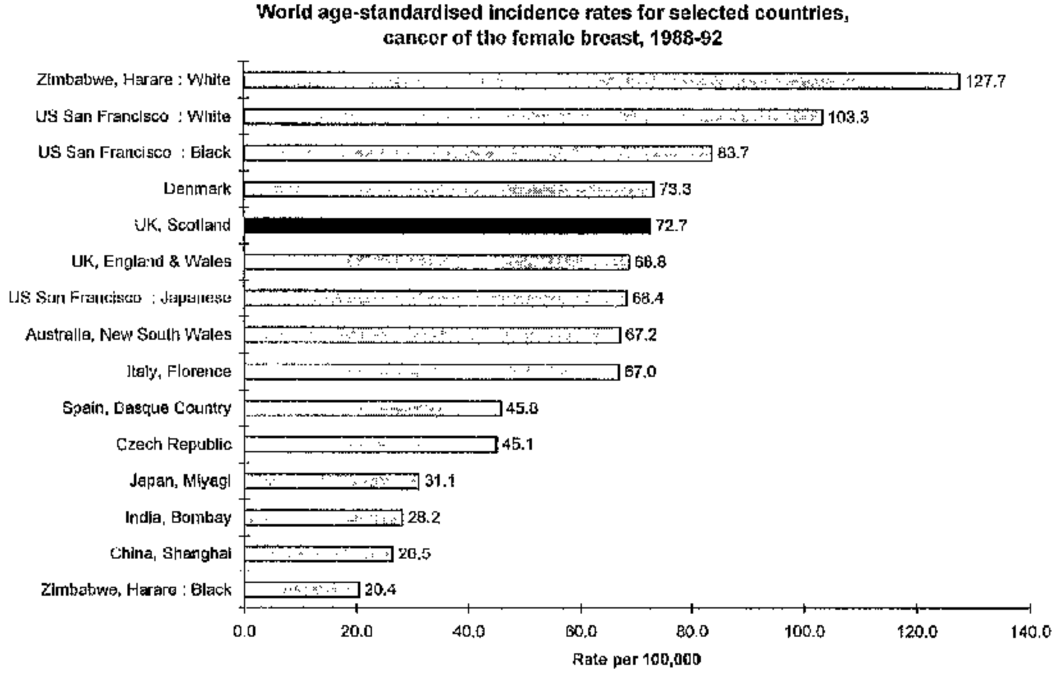


Figure 2.4: World age-standardised incidence rates for various countries (Parkin et al, 1997).

Doll et al (1966) point out that probably the most important bias when comparing incidence data from different countries is due to different methods of collection used to report the cancers. Some incidence rates may be low simply because known cases are not registered; alternatively, rates may be artificially higher because cases are registered without being verified pathologically to confirm the diagnosis.

Studies of migrants from Japan to the US have shown that these women have a marginally increased incidence rate compared with women of similar ages who have remained in Japan (Buell, 1973). However, Japanese women born in the US (i.e. descendants of these migrants) have very similar incidence rates to those of white US women (Shimizu et al, 1991), although this is not so obvious from Figure 2.4 for the period 1988-1992 in women in San Francisco. These findings indicate that environmental and social factors may be more important than genetic factors in altering the risk of getting breast cancer.

Breast cancer is also known to be a disease of the affluent, with a much higher incidence in women resident in areas of low deprivation or women in high social class (Sharp et al, 1993; Henderson et al, 1996). However, this pattern is not observed in either mortality or survival rates where both a higher mortality rate and worse survival figure are associated with a greater extent of social deprivation in some (Karjalainen & Pukkala, 1990; Schrijvers et al, 1995; Carnon et al, 1994) but not all (Twelves et al, 1998a) studies (see Section 5.2.3).

Figure 2.5 below shows the association between the risk of breast cancer and the Carstairs deprivation index (Carstairs & Morris, 1991). This measure of deprivation is often used for Scottish health statistics. The Carstairs classification of socio-economic deprivation was adapted to represent quintiles from the total Scottish population, based on the 1981 Census and updated for the 1991 Census. This measure is area-based and assigns to the populations living within small areas a score to reflect not readily measurable quantities, such as material well-being or poor access to amenities.

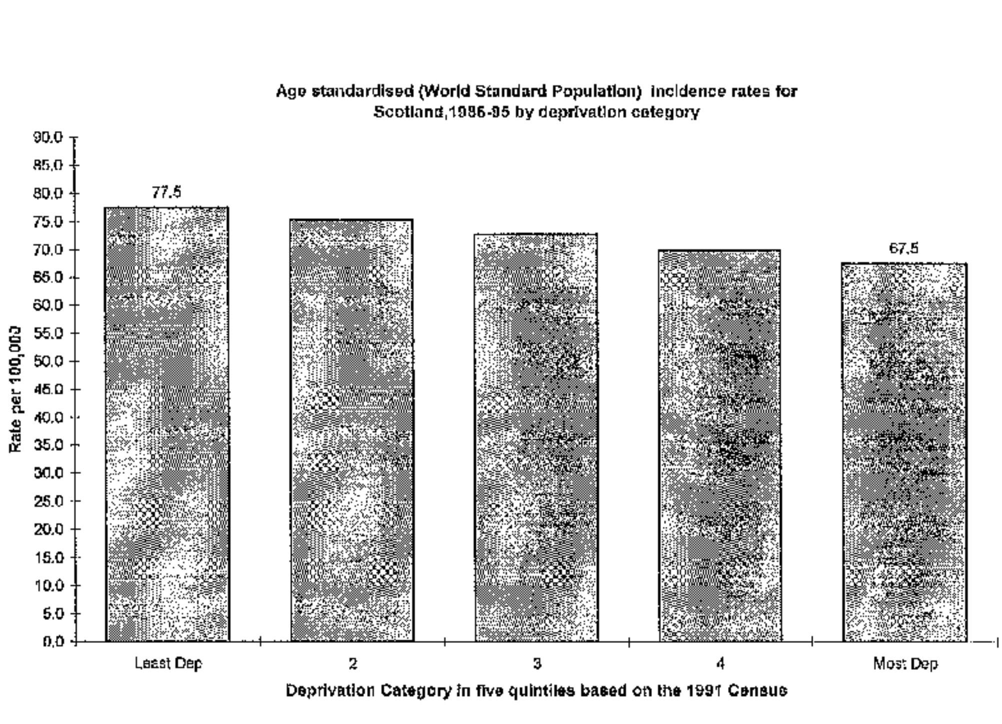


Figure 2.5: World age-standardised incidence rates in Scotland by deprivation category.

2.1.3 MORTALITY FROM BREAST CANCER

For the whole of the UK, breast cancer is the most common cause of female cancer mortality, representing 20% of all female cancer deaths (Cancer Research Campaign, 1996). Breast cancer was the cause of death for 3.9% of the 31,709 women who died in 1995 in Scotland (General Register Office for Scotland, 1996). Although breast cancer is more common in older women, it has the highest impact on mortality on women aged 35-54 (Figure 2.6). In this age range, breast cancer accounted for 15% of all female deaths, not just those due to cancer, during 1995 in Scotland compared to 13% due to ischaemic heart disease (ICD-9 410-414) and 8% due to lung cancer (ICD-9 162).

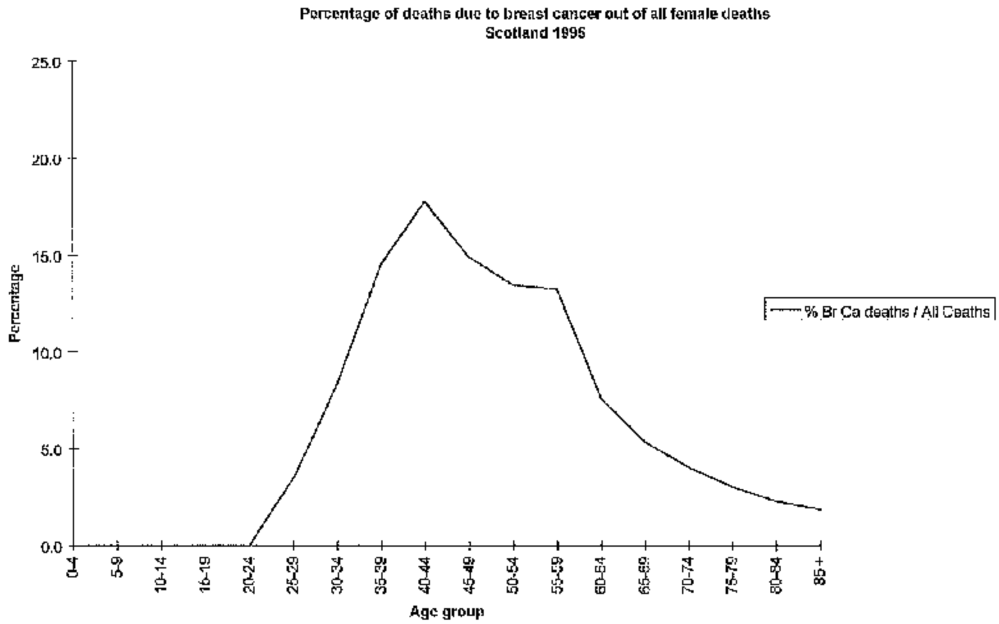


Figure 2.6: Percentages of deaths due to breast cancer by age group.

Scotland has one of the highest rates of breast cancer mortality and can be compared to those in other countries (Table 2.2; World Health Organisation, 1996).

World age-standardised mortality rates per 100,000 population from breast cancer, selected countries, most up-to-date years	
Netherlands (1994)	26.7
England & Wales (1994)	25.8
Scotland (1995)	25.2
Israel (1993)	24.5
US (1992)	21.4
Australia (1993)	20.4
France (1993)	19.8
Italy (1992)	19.8
Norway (1993)	19.4
Estonia (1994)	17.8
Finland (1994)	16.1
Japan (1994)	7.1
China, various urban (1994)	6.2
China, various rural (1994)	3.6

Table 2.2: World age-standardised mortality rates per 100,000 population from breast cancer for selected countries for the most up-to-date years.

One of the main aims of introducing the Scottish Breast Screening Programme (Scottish Breast Screening Programme Central Co-ordinating Unit, 1997) was to try to reduce mortality from breast cancer in the screened age group, 50-64 years old. Figure 2.7 shows the European age-standardised mortality rates per 100,000 for Scotland for 1950-1995 (Brewster et al, 1996a).

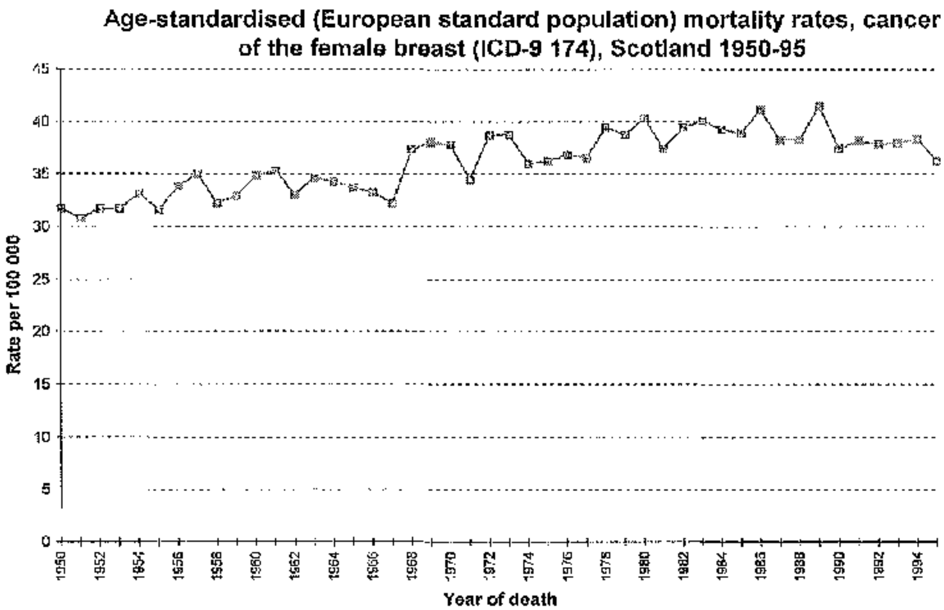


Figure 2.7: European age-standardised mortality rates for Scotland.

Mortality from breast cancer in Scotland has risen gradually since 1950, although it may now be levelling out, or even falling (Brewster et al, 1996a). A fall in mortality is supported by data for the whole of the UK (Quinn & Allen, 1995; Peto, 1998). These three studies suggest that the widespread introduction of adjuvant tamoxifen is the probable reason for this reduction in mortality. Death rates may be expected to fall further as benefits of screening are unlikely to have become fully apparent yet.

2.1.4 SURVIVAL FROM BREAST CANCER

In contrast to mortality, survival depends only on the number of deaths from the disease (or from any cause) among patients with the disease and therefore does not depend on the incidence (Berrino et al, 1995). Due to this difference, reduction in mortality should be the aim of any treatment (curative or preventative) or early diagnosis scheme, such as screening. However, treatment effects can best be assessed by examining survival.

CRUDE AND RELATIVE SURVIVAL
The relative survival figure for breast cancer tries to adjust the crude survival from breast cancer to correct for other causes of death. It does this by comparing the observed survival with the expected survival, based on the general mortality life tables for a population with the same age structure, for the same time period (Ederer et al, 1961). Relative survival is therefore age adjusted but does not allow for any variations in the numbers of deaths expected in the different deprivation categories, or Health Boards, say.

In Scotland, the crude and relative 5-year survival figures for all ages (0-84) have steadily improved since 1968 (Table 2.3). The figures are from Black et al (1993) and Harris et al (1998), except the 1988-1992 crude survival figure, which was calculated separately as this has not yet been published.

Period	Crude 5-year survival (%)	Relative 5-year survival (%)
1968-1972	49.5	56.4
1973-1977	52.0	59.7
1978-1982	55.1	63.1
1983-1987	56.3	64.3
1988-1992	63.8	70.1

Table 2.3: Crude and relative 5-year survival in Scotland.

Crude and relative survival figures may vary by age. For example, 5-year relative survival for the 35-44 group was 95.3% compared to 71.3% for the 55-64 group and only 18.2% for the 75-84 age group for the 1983-1987 cohort (Black et al, 1993).

Survival figures can also vary by country. Figure 2.8 shows 5-year relative survival values for the period 1981-1982 for twelve countries in the EUROCARE study (Berrino et al, 1995), for women diagnosed with breast cancer in the period 1978-1985. The 1981-1982 figures are shown because information was not available for all countries for the other years. Sant et al (1998) modelled these data and found variation by age, year of diagnosis and country, possibly due to variations in the quality of the data collected, or in the quality of the treatment women receive in the different countries (Sant et al, 1998; Berrino et al, 1995).

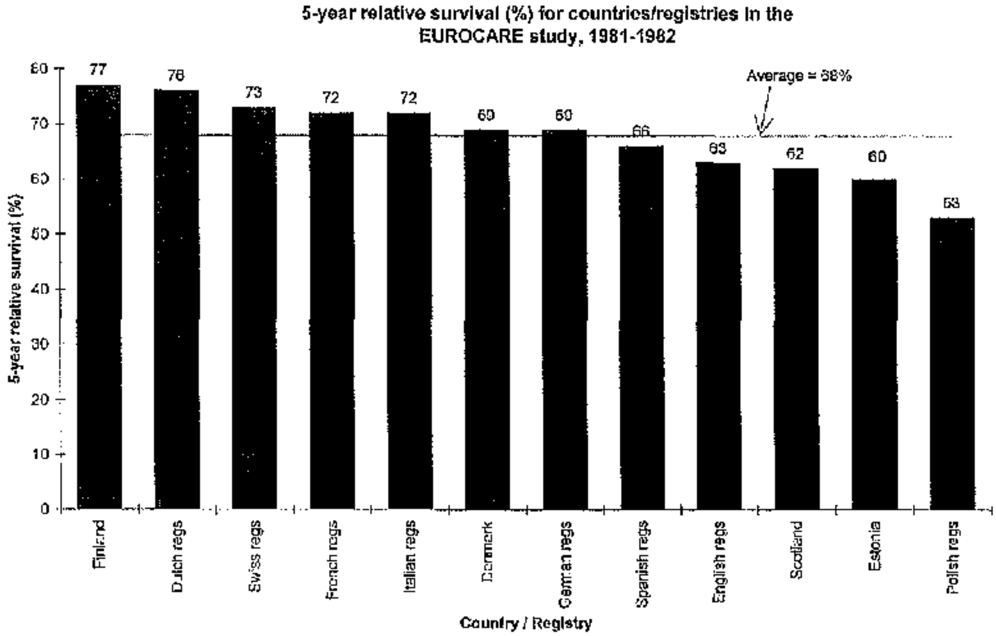


Figure 2.8: 5-year relative survival (%) for the EUROCARE countries / registries, 1981-1982. Note that only Denmark, Estonia, Finland and Scotland are national registries.

INTERPRETATIONS OF SURVIVAL FROM BREAST CANCER

Identification of prognostic factors for breast cancer survival (Miller et al, 1994) can:

- aid the decision of which treatment to give to a particular patient;
- allow treatments given to groups of patients with similar risks of recurrence or death to be compared;
- enhance understanding of breast cancer, which may lead to new treatments or strategies being developed;
- assess health education, to encourage earlier presentation;
- help evaluate the impact of the screening programme.

Differences in survival may be due to differences in case mix, by age, region, deprivation or variables associated with service provision from the National Health Service in Scotland, such as surgical case load. However, it is necessary to remember that case mix can vary for different levels of these variables as well as any treatment effects. That is, the case mix of a group of women may affect the survival chances for that group, irrespective of the treatment the women in the group receive. Thus, prognostic factors can influence the overall survival prospects of a group of women, as well as affecting the chances of an individual surviving.

For example, it may appear that a group of women detected by the screening programme, and therefore treated at a specialist centre (see Section 2.3.1), have better survival chances than women treated elsewhere. However, this group of screen-detected cases will almost certainly have a higher percentage of women with small, node negative tumours (see Section 2.3.2) and will, therefore, inherently have better survival prospects than the group of non screen-detected cases, notwithstanding the fact that these women may also receive superior treatment having been seen at a specialist centre.

2.2 AETIOLOGY OF BREAST CANCER

Some basic biological details relating to cancer in general, and specifically to breast cancer, are given in Section 2.2.1, whilst known and possible risk factors for developing the disease are discussed in Section 2.2.2.

2.2.1 BASIC BIOLOGICAL DETAILS

Normal cells in an organ, such as the breast, are continuously growing, reproducing and dying to allow normal function of the organ. Breast cell turnover is under partial control of circulating oestrogens, a group of female hormones. It appears that both oestrogen (Henderson et al, 1985) and progesterone stimulate cell division in the breast (Ferguson & Anderson, 1981; Henderson et al, 1996). As well as being produced by the ovaries in premenopausal women, oestrogen is also produced in smaller amounts from the conversion of the adrenal androgens to oestrogen in fat cells. Virtually all of the circulating oestrogens in postmenopausal women are produced via this route.

Cancer is the term used to describe the occurrence of a growth when this continual birth and death process goes wrong and abnormal cells develop and become invasive. If DNA in a cell becomes transformed, this leads to altered regulation of cell turnover, which then leads to cancer. The term malignant tumour or malignant neoplasm can also be used for cancer. The majority of breast cancers form in the epithelial cells lining the milk ducts in the breast (Henderson et al, 1996).

Boyle & Leake (1988) point out that breast cancer is not one disease, but several and prognosis and survival depend on various factors. Firstly, the tumour can be hormone sensitive or independent, which affects a woman's response to hormonal treatment. Secondly, tumours can be aggressive in nature or can be slow growing, and finally, that the tumour will either remain as a disease of the breast, being controlled locally, or it can metastasise very quickly; that is, the ability to "spread from the site of origin to distant tissues" (Souhami & Tobias, 1995). This is the main attribute which sets cancerous cells apart from normal cells. This spread of the cancer occurs when tumour cells invade local tissues, or are carried via blood or lymphatic systems to other organs throughout the body. Common secondary cancers from the breast include bone, the liver, the lungs and skin (Souhami & Tobias, 1995). In the case of breast cancer, it is usually the metastases that kill patients, not growth confined to the local tissue.

2.2.2 RISK FACTORS FOR BREAST CANCER

Oestrogens appear to play an important role in aetiology of breast cancer and may mediate the apparent effects of age and geography (Section 2.1.2). Other risk factors can be split into those related to reproductive life and those unrelated to it.

RISK ASSOCIATED WITH REPRODUCTIVE LIFE

Four risk factors which can be thought of as the natural reproductive factors (i.e. those linked to exposure to oestrogen and also to progesterone occurring naturally in the body) are age at menarche, length of menstrual cycle, age at menopause and age at first pregnancy. Two other factors: use of oral contraceptives and use of hormone replacement therapy are artificial reproductive factors. Table 2.4 overleaf gives the levels associated with higher risk of developing breast cancer for these factors.

OTHER POSSIBLE RISK FACTORS

Other known or possible factors for breast cancer are now discussed.

Height, weight and body mass index (BMI): Obesity is often measured using the Body Mass Index, defined as weight divided by height^2, measured in kg/m^2.

Vatten (1996) found strong evidence that height, weight and BMI are all positively associated with breast cancer risk for postmenopausal women. The relationship of height and weight with breast cancer risk is not so clear for premenopausal women, although there is some evidence to support the argument that being obese decreases the risk of getting breast cancer for premenopausal women.

Risk Factor	Higher Risk	References
Age at Menarche	**Early age.** Two year delay: relative risk 0.9 (95% CI: 0.85, 0.94).	Hsieh et al (1990); Titus-Ernstoff et al (1998)
Length of Menstrual Cycle	**Short cycle.** 28 day vs 33 day cycle: twice the risk	Henderson et al (1985)
Age at Menopause	**Late age.** Aged 55 vs under 45: twice the risk.	Hsieh et al (1990); Trichopoulos et al (1972)
Pregnancy	**Nullparity.** Nulliparous vs parous: 1.5 times the risk. **Late age at pregnancy.** Aged over 35 vs <18 years: 3 times the risk.	MacMahon et al (1970); Henderson et al (1996); Tavani et al (1997)
Use of Oral Contraceptives	**Currently taking the Pill.** Relative risk 1.24 (95% CI: 1.15, 1.33) **Within 10 years of stopping use.** Relative risk 1.07 (95% CI: 1.02, 1.13).	The Collaborative Group on Hormonal Factors in Breast Cancer (CGHFBC, 1996)
Use of Hormone Replacement Therapy (HRT)	**Currently taking HRT or within 5 years of stopping taking it.** For each year of use, risk increases by factor 1.02 (95% CI: 1.01, 1.04)	CGHFBC (1997)

Table 2.4: Risk factors for breast cancer due to reproductive life.

Family history: Evans et al (1994) showed that there is an increased risk of getting breast cancer if there is a history of breast cancer or other associated cancers (ovary, prostate, colon) in the family. This risk is even higher if these cancers occurred at an early age in the relative. Most of the cases related to family history occur at early age so that virtually all breast cancer cases that are diagnosed in women over the age of 60 are not due to inherited gene mutations.

Benign breast disease: This is a general term that is given to several different types of non-cancerous diseases that can affect the breast. There is some evidence to suggest an increased risk of developing malignant breast cancer for women who had benign breast disease compared to women who do not have any previous breast disease (Cancer Research Campaign, 1996).

Radiation: Tokunaga et al (1994) examined the incidence of breast cancer among the atomic bomb survivors of Hiroshima and Nagasaki in Japan. They found a strong linear dose response relationship of radiation exposure with breast cancer risk. This was much stronger for women aged under 20 years at the time of the exposure than for women aged 40 and over at the time of the bombings. Radiation given as chest *x*-rays searching for tuberculosis showed similar increased risk for getting breast cancer (Lipworth, 1995).

Diet: Many studies have examined whether there is any relationship with breast cancer risk and diet. These studies are difficult to conduct as it is hard to know what 'exposure' there was from food and energy levels ought to be taken into account. Since diet varies between individuals, across countries and across socioeconomic backgrounds, several components have been examined for their links with breast cancer risk. No firm evidence has been found to support the link between dietary fat (Cassidy, 1996); dietary fibre (Howe et al, 1990; Stoll, 1996) and vitamins A, C and E (Cassidy, 1996; Bohlke et al, 1999) with breast cancer risk.

Alcohol intake and smoking: A recent meta-analysis by Longnecker (1994) showed some evidence of a positive association of breast cancer risk with alcohol consumption, both in terms of some versus none and the amount of alcohol consumed. There does not appear to be much evidence to link breast cancer risk with smoking (Henderson et al, 1996). A weak inverse association between circulating levels of oestrogens and smoking are discussed by Michnovicz et al (1986).

2.3 MANAGEMENT OF BREAST CANCER

Whilst there have been some major breakthroughs in the treatment of breast cancer and the development of drugs which attack the cancer in an effort to prevent it spreading throughout the body, only about 60% of women survive for five years, many of whom have a relapse at some point. The organisation of breast cancer services in Scotland; the wide range of treatments available for breast cancer and the importance of participation in clinical trials are discussed in this chapter.

2.3.1 ORGANISATION OF BREAST CANCER SERVICES IN THE NATIONAL HEALTH SERVICE IN SCOTLAND

Traditionally breast surgery was not a separate sub-specialisation but was performed by most surgeons. Increasingly, treatment is now focused at a Breast Unit or one of the Screening Centres, with surgery being performed by a breast specialist. A policy document from the Chief Medical Officers of England and Wales (Expert Advisory Group on Cancer, 1995), known as the 'Calman/Hine' report, detailed plans of a network between primary care through Cancer Units at district hospitals to Cancer Centres for the provision of cancer services in England and Wales. One of the main points was that breast cancer can be managed at Cancer Units at district hospitals, but with Cancer Centres providing expertise in the management of all cancers and having additional specialist diagnostic and therapeutic resources, such as radiotherapy.

The Scottish Cancer Co-ordinating and Advisory Committee (SCCAC) proposed a similar network for Scotland (SCCAC, 1996). The Cancer Centres were identified as the locations where radiotherapy is given, namely: Raigmore Hospital in Inverness plus the four large teaching hospitals: Aberdeen Royal Infirmary; Ninewells Hospital and Medical School in Dundee; Western General Hospital (which in 1987 included the Longmore Breast Unit, now closed) in Edinburgh and the Western Infirmary/Beatson

Oncology Centre in Glasgow. In Scotland, 15 Health Boards provide health care for residents within their defined areas. The five Health Boards containing the Cancer Centres are Highland, Grampian, Tayside, Lothian and Greater Glasgow respectively.

These Cancer Centres are the bases for non-surgical oncology with many of the oncologists based at them visiting Cancer Units to aid decisions about the prescription of chemotherapy. Women are referred from the Cancer Units to the Cancer Centres for their radiotherapy. There is an increasing acceptance of the need for a multidisciplinary approach at specialist Breast Units with specialist surgeons, radiologists, breast care nurses, pathologists, oncologists and plastic surgeons.

Richards et al (1997) describe how the implementation of the Calman/Hine proposals has worked so far in the West Midlands. Dewar et al (1999) examined the increase in the use of radiotherapy and chemotherapy between the years 1987 and 1993 in Scotland (based on the Breast Cancer Audit data). They found that there had been an increase in the number of patients being referred to an oncologist from 1076 (50% of Audit population in 1987) to 1634 (64% of Audit population in 1993), which is a 52% increase. The number of patients receiving adjuvant radiotherapy and chemotherapy increasing by 72% and 215% respectively. However, there was only an increase from 32 to 37 consultant oncologists (16% increase) between the two years. Whilst the increase in the use of adjuvant therapy is necessary to ensure appropriate treatment for women with early breast cancer (see next section), there must be enough staff with the expertise needed to deliver this service. Richards & Parrott (1996) showed that oncologists currently only see half of the patients with cancer in Britain. The SCCAC report can only serve to increase the workload of these oncologists further.

The main purpose of both the Calman/Hine and SCCAC reports are that all women should have uniform access to high levels of specialist care to provide optimal treatment. Twelves et al (1999) point out that there were inequalities in determining whether or not a woman moved Health Board for her treatment (that is, she was treated at a hospital that is not within her Health Board of residence). They found that younger women and women living in affluent areas were more likely to move Health Boards for their treatment than more elderly or women living in social deprivation.

Several papers have shown a benefit from women being managed by specialist surgeons, and more importantly by a multidisciplinary team (Sainsbury et al, 1995a; Sainsbury et al, 1995b; Gillis & Hole, 1996; Twelves et al, 1998a; Twelves et al, 1998b) in terms of receiving more appropriate treatment, entering clinical trials and improving survival from breast cancer.

2.3.2 TREATMENT OF BREAST CANCER

INTRODUCTION

By the time the woman presents, the cancer may already have spread from the breast tissue into the lymph nodes, or formed secondary cancers in other organs. The management and prognosis of women with no evidence of metastases at presentation is very different to that for women where the cancer has spread. Figure 2.9 shows the Kaplan-Meier survival curves (see Section 5.1.1) for these two groups of women based on all of the women (n=2148) included in the Breast Cancer Audit in Scotland in 1987 (Scottish Breast Cancer Focus Group et al, 1996). The non-metastatic group has been broken down into those who underwent surgery and those who did not.

Of the 8% of women who had metastases at presentation, only about 10% of them were still alive at five years. The 16% of women who did not undergo surgery despite having no evidence of metastases at diagnosis had a better outlook, with approximately 35% of them surviving at 5 years. These women were mainly elderly and may have been deemed too unfit to have an operation. Alternatively, the tumour may have been too large, growing too quickly, or the women may simply have refused surgery. The remaining 76% of women did not have metastases at presentation and did undergo surgery. This is the subgroup that was included in all subsequent analyses, as this is the only group where there is a realistic chance of cure.

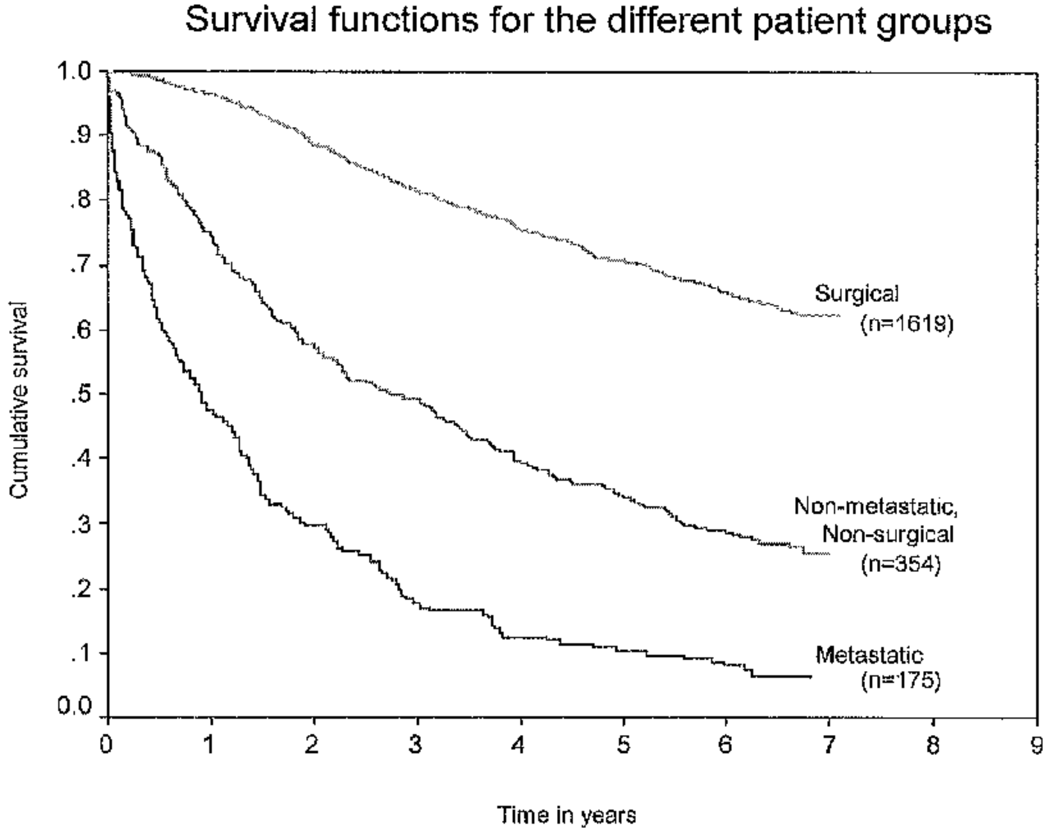


Figure 2.9: Estimated survival curves for the different types of breast cancer.

The Early Breast Cancer Trialists' Collaborative Group (EBCTCG) define early breast cancer to mean any cancer which is confined to the breast (or to the lymph nodes) and which can be removed surgically. Thus, disease that has metastasised beyond the breast and axilla (metastatic cancer) and tumours that are too large, too aggressive or located in an awkward location in the breast such that it cannot be excised (locally advanced cancer) are not included in the definition of early breast cancer. The two non-metastatic groups above (surgery and no surgery) fall roughly into the early and locally advanced categories respectively, although the non-surgical group may have included women whose disease was technically early, but who were too unwell with concomitant diseases to survive surgery or those who refused such treatment.

The EBCTCG have drawn together data from a large number of clinical trials for early breast cancer from around the world and have published several overviews. The Scottish Intercollegiate Guidelines Network (SIGN), in collaboration with the Scottish Cancer Therapy Network (SCTN), recently published clinical guidelines (SIGN/SCTN, 1998) for the management of breast cancer in Scotland (not only for early breast cancer, but also for locally advanced and metastatic breast cancer). How a woman is treated depends largely on her prognosis when she presents at the clinic.

DETERMINATION OF THE DISEASE AND OPTIMAL MANAGEMENT

When a woman has been referred to a breast clinic with a suspected breast tumour, usually fine needle aspiration followed by a biopsy is used to check whether the cells in the lump are malignant. Once the diagnosis has been confirmed, some basic staging investigations are performed. These include a blood test to check blood cell count and liver function, a chest *x*-ray and a clinical examination. The clinical stage and metastatic status are determined from results of these simple tests.

Cancers can be described in terms of clinical stage and metastatic status. The **clinical stage** describes the state of advancement of the disease. One widely used system is the four category TNM classification (UICC, 1987) which depends on the clinical tumour (T) size measured, whether there is any nodal (N) involvement (obtained by palpation) or any evidence of distant metastases (M). The **metastatic status** can also be determined by imaging techniques such as bone isotope, MRI and ultrasound scans. The presence of metastases means that the cancer is incurable and the survival prospects are therefore very poor.

If there is evidence of metastases then treatment is generally palliative (see below). If no secondary deposits can be found then the initial priority is local treatment to deal with the cancer in the breast.

TREATMENT FOR EARLY BREAST CANCER

The choice of treatment for early breast cancer should be made on the basis of the risk of recurrence of the disease, the menopausal status of the woman and the wishes of the patient. However, the initial management is normally **surgery**, either conservative (where only the lump is excised) or a mastectomy (where the whole breast is removed). The surgeon often removes some or all of the lymph nodes in the axilla (axillary sample or clearance), both for diagnostic and therapeutic purposes. Usually when only the lump is excised, **radiotherapy (RT)** is given to the breast, although it can also be given to the chest wall after a mastectomy or to the axilla to try to control the disease locoregionally.

EBCTCG (1995) showed that there was no survival benefit from performing a mastectomy as opposed to lump excisions plus RT for tumours <4cm in size. The results from this overview are based on approximately 28,000 women, entered into trials for surgery with or without RT, which began randomisation before 1985. In this case (size <4cm), the decision for surgery type should depend on factors such as the ratio of the size of the tumour to the size of the breast, the age of the patient and the patient's choice.

Although most women do not have clinical evidence of metastases at diagnosis, surgery is not usually sufficient because undetectable micrometastases may already be in the blood. These may lay dormant for a number of years until they develop into a clinically detectable recurrence which may eventually kill the patient. It is, therefore, necessary to identify the risk of relapse to guide the choice of **adjuvant systemic** treatments given at the same time as surgery. The aim of these therapies is to treat the whole body to try to prevent the disease recurring or spreading. The factors which identify the risk of relapse also, therefore, affect the overall prognostic chances of survival for a woman.

The single most important prognostic factor is **pathological node status.** This is obtained from tissue removed from the axilla. The tissue is examined to determine how many of the total number of nodes examined show tumour involvement. The status is often just given as positive or negative. Nodal involvement is often an indicator that the disease has micrometastasised, with tumour cells already spread into the blood supply but have not yet infiltrated other tissue and so cannot be detected. In general, if no nodes are positive, the prognosis is good. The outlook worsens as the number of positive nodes increases.

Miller et al (1994) show that survival decreases as the number of nodes involved increases, with 10-year survival rates being 65%, 38% and 13% for none positive, 1-3 nodes positive and ≥ 4 nodes positive, respectively. Alternatively, any nodal involvement had 10-year survival rate of 25%.

Another important prognostic factor is **pathological tumour size**, as measured by the pathologist. Ewertz et al (1991) found an increasing risk of death with tumour size, as did Miller et al (1994), Newman et al (1997) and Gordon et al (1992). When compared to tumours ≤ 2cm, the hazard ratios of dying for tumours of size 2.1-5cm and >5cm were 1.43 (95% CI: 1.09, 1.88) and 2.13 (1.33, 3.43) respectively (Newman et al, 1997). However, Carter et al (1989) reported an interaction between pathological size and pathological node status. That is, the effect on survival was larger for node status negative compared to node positive, when tumour size was large than when tumour size was small. Conversely, the survival effect was larger for small tumour size compared to large, when node status was positive rather than negative. They showed that in 71% of women with tumour size ≥ 5cm, at least one node is expected to have involvement.

Tumour differentiation or **grade** measures the degree of differentiation of the cells in the tumour. That is, how similar the cancer cells are in appearance, shape and structure compared to the normal cells in the breast. Miller et al (1994) showed that the grade of the tumour affects the survival figures. Grade I (well-differentiated) tumours had a 10-year survival of 85% compared to poorly-differentiated tumours (Grade III) having a 40% survival at 10 years. Ewertz et al (1991), Freedman et al (1979) and Sainsbury et al (1995a) all found a similar relationship of survival with grade.

Age is another factor which is both prognostic for survival and may influence choice of treatment. Some of the studies demonstrate a linear decrease of survival with age (Freedman et al, 1979; Karjalainen & Pukkala, 1990; Sainsbury et al, 1995a). However, other studies showed an increased hazard of death for women aged under 35 compared with the group of women aged over 35 (Miller et al, 1994; Richards et al, 1996) and under 40 (Newman et al, 1997) compared to 40-49, with the risk then increasing for the age groups greater than 40-49.

Clinical stage is not really used to determine choice of adjuvant treatment for early breast cancer, but it is another prognostic factor, since one of its components is the presence of metastases. Several staging methods have been included in reported survival analyses, some use the TNM staging method (UICC, 1987), whilst others use the extent of disease: local (no nodes or metastases), regional (nodal involvement) or metastases.

Shek & Godolphin (1988) found that survival decreased as the clinical stage increased. Sainsbury et al (1994) show that the survival at five years for Stages I, II, III and IV were 84%, 71%, 48% and 18% respectively. Stage IV disease represents metastatic breast cancer. Clinical stage obtained from the TNM system is not always reliable because the measurements are based on clinical assessment and not pathological details (Sainsbury et al, 1994). Bundred et al (1994) suggest that true prognostic information can only be achieved by histopathological assessment of the nodes removed from the axilla because only 70% of involved nodes can be detected clinically. Brewster et al (1996b) showed that there was poor agreement between clinical and pathological staging information. Sainsbury et al (1995a) and Schrijvers et al (1995) used extent of disease to stage the tumour with Sainsbury et al (1995a) quoting an increased hazard of death for nodal involvement of 1.99 (95% CI: 1.89, 2.09) and an even higher one for metastatic disease (4.39; 95% CI: 3.98, 4.85) when compared to local disease after adjusting for age, grade, deprivation, period of treatment and combination of treatment given. Schrijvers et al (1995) reported similar findings.

Oestrogen receptor (ER) status gives a measurement about the presence or absence of oestrogen receptors in the cells of the sample of tissue excised. Several ways of reporting the ER status mean that scores given by different labs cannot be directly compared. The scores are on a continuous scale, with a cut-off selected to separate negative from positive. Often only the binary variable is reported and used in statistical analyses, irrespective of the method used to obtain it.

The ER status of the woman is determined at the age when the transformation of the DNA in the cell takes place and not the age at diagnosis. Pujol et al (1998) found that 67% of peri- & postmenopausal group had ER positive tumours compared to 59% of premenopausal women. Souhami & Tobias (1995) suggest similar figures for postmenopausal (65% ER positive) but only 30% ER positive for premenopausal women. In general, having an ER negative tumour implies a poorer prognosis than having an ER positive tumour (Newman et al, 1997), with a hazard ratio for ER negative vs ER positive of 1.76 (95% CI: 1.35, 2.29). Similar findings were observed by Gordon et al (1992), Hawkins et al (1996) and Shek & Godolphin (1988). However, Miller et al (1994) suggest that the effect of ER status on survival weakens over time. They also point out that ER status and tumour grade are often correlated and give the example that most Grade III tumours are ER negative. Giuffrida et al (1992) observed that ER status was significantly associated with body mass index, with obese women more likely to have ER positive tumours in both pre- and postmenopausal women.

The importance of ER status is considerable for guiding whether or not hormone treatment would be useful. **Endocrine therapy** aims to prevent the cancer cells getting the hormones that they need to grow and survive. The usual drug of first choice is tamoxifen which is an anti-oestrogen, but endocrine therapy also includes aromatase inhibitors, which block the production of oestrogen from the fatty tissue in postmenopausal women.

Two EBCTCG overviews (EBCTCG, 1992; EBCTCG, 1998a) have examined the prescription of tamoxifen and the results are based on roughly 37,000 women in 55 trials. The annual reduction in the odds of recurrence in the two overviews were 25% (standard deviation (SD) 2) and 26.4% (SD 1.5) respectively. Similarly, there was an annual reduction in the odds of death of 17% (SD 2) and 14.5% (SD 1.7) for use of tamoxifen versus none for all ages. From the earlier report (EBCTCG, 1992), the corresponding 10-year survival figures for all deaths were 58.8% for the use of tamoxifen and 52.6% for none, which gives a very highly significant difference of 6.2% (SD 0.9). The benefits increased with longer duration of tamoxifen (i.e. use of drug for five years instead of two years) and SIGN/SCTN (1998) recommend that it be given for at least five years.

Ovarian ablation is another form of endocrine therapy. The purpose is to stop the normal function of the ovaries. This can be achieved surgically by removal of both of the ovaries; by irradiation of the ovaries; or by drugs which suppress their control of the menstrual cycle, hence altering the levels of oestrogen and progesterone circulating in the blood. Ovarian ablation can be used for early, locally advanced or metastatic breast cancer.

Two overviews (EBCTCG, 1992; EBCTCG, 1996) examined the results for about 3,000 women given ovarian ablation in total, roughly 2,000 of whom were under 50 years in age. The major finding was that ovarian ablation only provided a benefit in women aged under 50 years (a surrogate for premenopausal women), with an overall reduction in mortality per year of 18% (SD 5.7) for the under 50 group in the latter report. This was equivalent to a 15-year survival difference of 6% (SD 2.3) of 45.0% vs 39.0%.

An alternative systemic treatment is **chemotherapy (CT)**, used to kill cancer cells, both in the breast and the metastatic cells throughout the body. Chemotherapy is the term given to one or more cytotoxic drugs prescribed for this sole purpose and can be used for women with early, locally advanced or metastatic breast cancer. It can cause partial or complete ovarian suppression in premenopausal women (EBCTCG, 1996).

Two overviews examine the use of chemotherapy (EBCTCG, 1992; EBCTCG, 1998b) and are based on approximately 18,000 women. Chemotherapy was given either as single agents or in combinations and data are available from over 100 trials. The largest benefit was gained from giving polychemotherapy (multi-agents) for a prolonged period, although no additional benefit was gained from extending the period beyond 3-6 months (EBCTCG, 1992).

For women of all ages, the odds reduction in mortality was 11% (SD 2), which was highly significant (EBCTCG, 1992). The 5-year survival figures showed a benefit of 3.3% (SD 1.1) for chemotherapy vs none, and at 10 years the difference was very highly significant at 6.3% (SD 1.4). The benefit of giving CT was much greater for women aged under 50, although there was still a significant reduction observed for the 50-69 group. The gain for use of CT was higher for node positive women compared with node negative disease (EBCTCG, 1998b).

The exact choice of adjuvant systemic treatment for early breast cancer depends on the prognostic factors for risk of relapse and age or menopausal status of the patient, but it is generally accepted that either endocrine therapy or chemotherapy be given, either alone or in combination, following surgery (Richards et al, 1994).

TREATMENT FOR LOCALLY ADVANCED BREAST CANCER

In a minority of patients without evidence of metastatic disease at presentation, the woman cannot be operated upon because the disease has infiltrated the skin of the breast or chest wall; is in an awkward position; or is growing too rapidly. This cancer is known as locally advanced disease and the median survival for this group of women is about 24-30 months, with 5-year survival between 1% to 30% (Rodger et al, 1994).

The initial treatment is radical radiotherapy, followed by systemic treatment (SIGN/SCTN, 1998). Rodger et al (1994) point out that it may then be possible to perform some surgery if the systemic therapy reduces the bulk of the tumour. Often surgery is only performed on locally advanced disease to attempt to remove most of the tumour if it is fungating through the skin (Souhami & Tobias, 1995). A large number of patients with locally advanced cancer will develop uncontrolled disease of the chest wall. Patients given standard chemotherapy regimens have lower rates of recurrence than women not receiving CT, but they do not have improved survival.

TREATMENT FOR METASTATIC BREAST CANCER

This can either be for women with metastases at presentation or those who develop them as secondary cancers. "Currently, patients with distant metastases are incurable. The aim of treatment is therefore to maintain the highest quality of life and relieve symptoms" (SIGN/SCTN, 1998). Therefore, all of the treatment given at this stage is palliative and not curative in intent.

The median survival for women with metastases from breast cancer is about 18-24 months, although this varies considerably, depending on the site of the metastases, whether the tumour is hormone sensitive or not, and, for women with non-metastatic disease at presentation, the speed of progression of the metastases. Women with metastases in the bones and soft tissue (skin, other breast, lymph glands) have the best outlook, whereas patients with metastatic disease in the lungs, liver or brain may survive for as little as two months (Leonard et al, 1994).

Usually endocrine therapy is given first because it is not as toxic as chemotherapy. The exceptions to this are if the metastases are in sites such as the liver, lung or brain; or if there has only been a short interval between primary treatment (for patients without metastases at presentation) and the occurrence of metastases. In these situations, chemotherapy is the first line of treatment.

In some cases, surgery is performed to remove either the primary tumour and/or some of the metastases. The decision to operate or not will depend on site of the deposit and the overall health of the patient.

SIGN/SCTN (1998) recommend tamoxifen (or ovarian ablation for premenopausal women) as the first line of treatment, with progestogens (or aromatase inhibitors for postmenopausal women) as second line treatment if these fail (i.e. no response to the treatment or an initial response followed by disease progression). Only about 30% of women have an objective response (complete or partial) to hormone treatment, although women with ER positive tumours have a much higher rate of response of about 50-60% (Leonard et al, 1994).

When hormonal treatments no longer appear to have any effect on the cancer, chemotherapy is then considered. The first line response rates to CT (40-60%) are, in general, higher than for hormonal therapy, although they tend not to last as long and have more side-effects (Souhami & Tobias, 1995). Gregory et al (1993) found that it was not possible to predict which patients were likely to respond, but that women who responded to first line CT treatment were more likely to respond to second line CT treatment than those who did not respond to the first line treatment (24% vs 12%; P=0.04).

The survival benefit due to giving CT may be several months in a few women. These must be balanced against the toxic effects of treating women where CT gives no response (Souhami & Tobias, 1995; Ramirez et al, 1998).

When standard chemotherapy regimens fail to work, Leonard et al (1994) suggest that experimental CT drugs can be administered (with the patient's consent and adherence to the necessary guidelines for administering experimental treatments). In addition, the other symptoms of the cancer are treated to try to improve quality of life. Radiotherapy can be given as a palliative measure for women suffering from pain due to metastatic disease.

CLINICAL TRIALS FOR WOMEN WITH BREAST CANCER

There is, as described, wide variation in treatments given for breast cancer, especially among early, locally advanced and metastatic disease, but also within each of these three groups. With many new drugs becoming available, it is essential that they are tested both alone and in combination with other treatments. They can be assessed fully only through the use of Phase III randomised controlled trials (RCTs), although experimental Phase I and II trials provide the guidance for setting up RCTs.

Clinical trials are an integral part of defining better treatments, providing improved standards of care and optimising the standard therapy. They offer the opportunity of a major breakthrough in the treatment of cancer.

The overviews mentioned above demonstrate the large number of trials that have been available for early breast cancer. Despite this, participation in clinical trials for breast cancer is low. Tate et al (1979) estimated that on average 8% of patients with breast cancer in the UK entered clinical trials, whilst overall, 12% of women entered clinical trials in Scotland in 1987 (Twelves et al, 1998b) from analysis of the Breast Cancer Audit data.

From that Audit, 8.4% and 8.7% of patients in 1987 and 1993 respectively entered trials for early or locally advanced breast cancer in Scotland (Twelves et al, 1998b). In this study, it was found that being treated by a 'specialist' surgeon or seeing an oncologist implied that a woman was much more likely to enter a clinical trial. They also found that women treated on a clinical trial were more likely to have their tumour staged more thoroughly. For example, only 16% of women on a trial did not have their node status known compared to 32% of women not treated on a trial. Thus, patients treated on a trial may be managed more appropriately. Ramirez et al (1998) suggested that women on a clinical trial may have a better prognosis than women treated outwith the trial setting. Twelves et al (1998b) did not find a significant survival benefit for trial entry, although with only 58 deaths in women treated on a trial in the 1987 cohort examined for survival, the lack of significance may simply be due to a lack of power, as the hazard ratio suggested a benefit, 0.79 (95% CI: 0.59, 1.04; P=0.10) for those women entered onto a clinical trial.

Of the women diagnosed with breast cancer in Scotland in 1987, only 83 of the women included in the Breast Cancer Audit data (n=2148) were entered into trials for metastatic cancer (Twelves et al, 1998b). These women represent 12.8% of the 775 (175 with metastases at presentation plus 600 with non-metastatic disease at presentation, who had had a distant relapse by the time the data were collected) women eligible for entry to a trial for women with metastases.

It seems imperative that the number of women entering clinical trials increases. The reorganisation of the health service and support of trials provided by the Scottish Cancer Therapy Network mean it ought to be possible to achieve a similar level of participation for breast cancer trials in Scotland as that observed for entry of children into trials of acute lymphoblastic leukaemia (over 50%; Stiller, 1994). This should lead to improved treatments for women with breast cancer, and hopefully, ultimately lead to improved survival prospects.

This chapter has mainly focused on the risk of getting breast cancer, the optimal treatment available to women who do get the disease, and the survival chances for these women. The next chapter describes the purpose of a national retrospective audit of all women identified as having invasive breast cancer in Scotland in the years 1987 and 1993. This was performed to examine what treatment these women were receiving and what were their survival chances.

CHAPTER 3 SURVEY OF BREAST CANCER IN SCOTTISH WOMEN IN 1987

3.1 AIMS OF THE STUDY, METHODS OF DATA COLLECTION AND THE SUBSET OF PATIENTS SELECTED FOR ANALYSIS

AIMS

The main aims of the audit were to identify how women diagnosed with invasive breast cancer in Scotland in 1987 and in 1993 were managed and to investigate whether there had been any changes in the patterns of care between the two study years, during which time a national breast screening programme was introduced (Section 2.1.2). This thesis is only concerned, however, with analysis of the 1987 cohort. Analyses relating to the management patterns for these women and survival analysis results are discussed in the text. One issue to bear in mind is that this study is only a retrospective audit and not a controlled randomised clinical trial. Therefore, any conclusions reached can only be descriptive, indicating areas where a clinical trial might be appropriate or where further research could be beneficial.

METHODS OF DATA COLLECTION

A list of all women diagnosed with breast cancer in 1987 was obtained from the Scottish Cancer Registry.

All patients who were deemed ineligible were removed from this list. These included women who were DCO (death certificate only) registrations, because, by definition, only limited diagnostic information is held about such patients. Other women excluded were those who were diagnosed and treated outside Scotland; those women who in fact had non-invasive disease and also those women who had a previous diagnosis of breast cancer, identified using probabilistic record linkage (Kendrick & Clarke, 1993).

Having excluded the ineligible women, case notes were then sought for all eligible patients. However, not all of the case notes for these women could be found because some sets were either missing or had already been destroyed. From the case notes that were available, much additional information was collected, to supplement the data that had already been provided by the cancer registration system, by trained Data Managers from the Scottish Cancer Therapy Network (SCTN). A quality check was performed on a random sample of case notes to assess the accuracy of extraction of information from the case notes, using cross-checking of data extraction by Data Managers, and also checks were performed to assess the accuracy of the data entered onto the audit database.

NUMBERS INVOLVED IN THE 1987 COHORT

At the time that the list was drawn up from the Scottish Cancer Registry, there were 2,581 women who were registered in 1987 as having breast cancer. Out of these women, 79 were excluded as they were DCO registrations. Another 101 women were deemed ineligible because they were diagnosed and treated outside Scotland, their disease was non-invasive or they had had a previous diagnosis of breast cancer. This left 2,401 women who were considered to be eligible for inclusion in the study.

The Data Managers were unable to find 164 sets of notes and a further 89 sets had been destroyed. This meant, therefore, that information was available for 2,148 women in 1987. This represented 89% of the eligible cases. Data collection was undertaken during the years 1994-5.

It is important to be aware of the fact that there were significantly more notes missing (either not located or destroyed) for elderly patients. These patients were more likely to have died by the time of the data collection. Notes belonging to deceased people are more likely to have been destroyed or archived (where there may be a problem of retrieval). This possible bias cannot be accounted for in any subsequent analyses and needs to be remembered when interpreting any results.

ANALYSIS DATASET

A further 529 women were also excluded here. This was because only 1619 women were included in the cohort used in the survival analysis performed on the Breast Cancer Audit data (Twelves et al, 1998a). In that study, only those women who had no evidence of metastases at diagnosis and who underwent surgery were included in the analysis. One of the main purposes of this thesis was to investigate the effects of the unknowns on the results and conclusions from that survival analysis (see Chapter 5) and, therefore, the cohort of 1619 women was studied here.

The 529 women excluded were those with metastatic disease (175 in 1987), because these women would have been treated very differently from those women with early breast cancer (Section 2.3.2), and the women whose disease was non-metastatic, but did not undergo surgery (354 in 1987). This latter exclusion was because the three important prognostic factors: pathological node status, pathological tumour size and ER status can only be recorded if tissue is removed by surgery. Rather than perform the analysis on factors with even greater percentages of unknowns (see Section 4.2), the women who did not undergo surgery were not included in the analysis. It is possible that some selection bias may have been introduced by this exclusion, say for example, if different Health Boards had different policies for selecting women for surgery, leading to only the better prognosis women undergoing surgery.

Thus, the subgroup of women included in all analyses based on the Breast Cancer Audit data in this thesis relate to the 1619 women who had no evidence of metastatic disease at presentation and who underwent surgery.

OUTCOME INFORMATION

The initial plan had been to supplement outcome information collected from case notes by linkage to the death records from the General Registers Office (GRO). At the time of the initial analysis the latest death information available for linkage was up to the end of 1993.

It was realised that there would be a problem with the approach of using death information collected from case notes with data collection taking place during 1994 and the start of 1995. Bias would be introduced for those case notes examined towards the end of collection as these women would have had a longer time in which to have died or still have been seen alive. This bias could be regional systematic with Glasgow (being the largest region for data collection) taking the longest time to collect all of the information.

Therefore, it was decided that the most valid analysis would be to use the probability matching technique (Kendrick & Clarke, 1993) to link the cases with the GRO death records to obtain the date of death. For those women with no recorded death, the assumption that the women were still alive at 31/12/93 was made. Clearly, this would misclassify women who had migrated out of Scotland after diagnosis and died elsewhere. However, these women are likely to be few in number.

3.2 VARIABLES COLLECTED AND THE SUBSET SELECTED FOR ANALYSIS

INTRODUCTION

The information collected from case notes for all of the women in both years of the audit covered the referral patterns; the initial staging information collected at the clinic; the surgical procedures undertaken, by which surgeon and the date of diagnosis; other forms of treatment given; pathology details including extra staging information; and follow-up and outcome details. The data collected at each of these stages of management are discussed separately in Appendix 1.

Only information relevant to the analyses undertaken in this thesis are discussed subsequently. Analyses based on variables relating to non-relevant information can be found in 'Audit Report' (Scottish Breast Cancer Focus Group (SBCFG) et al, 1996).

VARIABLES SELECTED FOR ANALYSIS

The included variables were identified before undertaking the analysis. They fell into three categories: clinical, treatment and service(-related) variables. The clinical variables (Table 3.1) represent features of the patient and the disease at diagnosis. These are known to influence survival from prior clinical research (Section 2.3.2). The treatment variables (Table 3.2) are known from clinical trials to be of significant importance for determining outcome (Section 2.3.2). The service variables (Table 3.3) were chosen because they reflect the mode of service delivery by the National Health Service in Scotland (NHSiS). A social factor, deprivation was also included in with the service variables.

Clinical Variables:

Variable	**Variable Categories**
Age at diagnosis	<50, 50 - 64, 65 - 79, ≥ 80 years
Clinical stage	I, II, III, not known
ER status	positive, negative, not known
Pathological node status	positive, negative, not known
Pathological tumour size	≤ 2 cm, >2 cm, not known

Table 3.1: Clinical variables and definitions of the factors levels used in the analyses.

Age was divided into 15-year age bands so as to include the 50-64 range (the screening group) as one group. There were only 33 women aged under 35 years, so the original groups <35 and 35-49 were merged into one <50 group and analysed together in all analyses.

Clinical stage was derived from TNM staging. Since Stage IV patients are those with metastases, there are no Stage IV patients in the subgroup of 1619 women chosen for analysis.

ER status was considered positive if cytosolic protein ≥ 20 fmol/mg or staining $\geq 10\%$, otherwise it was taken to be negative.

Women were classified as unknown node status for three reasons: the information was not known as they had no axillary surgery; it simply was missing from the case notes;

the sample was inadequate in that it contained less than four nodes, all of which were negative.

The two original groups for pathological tumour size, >2-5 cm and >5 cm, were merged together to form one group, >2 cm, because there were only 79 women with tumours that were greater than 5 cm in diameter.

Treatment Variables:

Variable	Variable Categories
Type of surgery	mastectomy, breast conservation
Adjuvant chemotherapy	given, not given
Adjuvant endocrine therapy (including ovarian ablation)	given, not given
Any systemic treatment	given, not given
Adjuvant radiotherapy	given, not given

Table 3.2: Treatment variables and definitions of the factors levels used in the analyses.

As seen in Section 2.3.2, primary treatment of early breast cancer can include surgery, radiotherapy, chemotherapy and hormone treatments. Variables chosen for analysis are given in Table 3.2.

Women who had breast conservation followed by a mastectomy within three months of the initial surgery were coded as having had a mastectomy as their primary treatment.

Adjuvant endocrine therapy included both hormone treatment and ovarian ablation. The 'any' adjuvant systemic treatment group consisted of patients receiving chemotherapy or endocrine therapy or a combination of these treatments.

Adjuvant radiotherapy can be given to three different sites: to the breast, for women who have had breast conservation; to the chest wall, for those women who have had a mastectomy; and to the axilla, for all women, except those who have had an axillary clearance. As explained in Section 2.3.2, radiotherapy is a treatment of local control (trying to prevent local recurrences) rather than systemic control and has not been shown to have benefit in terms of overall survival. Therefore, in this thesis, site specific usage of radiotherapy is not considered.

Service Variables:

Variable	**Variable Categories**
Deprivation	I = least deprived, II, III, IV, V = most deprived
Health board of first treatment	A, B, C, F, G, H, I, L, N, S, T, V, Y
Referral to oncologist within 3 mths of diagnosis	yes, no, not known
Surgical caseload	1 - 9 patients per year, 10 - 29 patients per year, member of team or ≥30 patients per year, not known

Table 3.3: Service variables and definitions of the factors levels used in the analyses.

These variables represent the social background of the patient and organisational infrastructure of the NHSiS under which the primary treatment was administered. Table 3.3 gives the service variables chosen for analysis.

The Carstairs classification of socio-economic deprivation (Carstairs & Morris, 1991) was adapted to represent quintiles from the total Scottish population, based on the 1981 Census (Section 2.1.2). This is an area-based measure of socio-economic status, derived from the postcode of residence at the time of diagnosis.

The Health Board of first treatment was the Health Board in which primary treatment was administered. Although a few patients may have had neo-adjuvant treatment, the decision was made to derive the Health Board of first treatment to be the Health Board where surgery was performed for all of those women who underwent surgery (all of the cases in the chosen subgroup of analysis for this thesis). Health Board of residence was not used because the aim of the audit (SBCFG et al, 1996) was to examine management patterns and the effect on survival.

Due to the small numbers of women treated in the Health Boards covering the Islands (Orkney, Western Isles and Shetland), these three Health Boards were grouped together as the 'Islands' to represent off-mainland treatment. Appendix 2 gives the key to the

Health Board labels given in Table 3.3. These labels are those used in the virtually all NHSiS documents.

Referral to an oncologist for primary treatment meant that women for whom the date of referral was unknown had to be excluded because the referral could have been for primary treatment or later following a recurrence. The classification 'no' for referral to an oncologist included those women who saw an oncologist after three months as it was assumed that this referral was not as part of the primary treatment. The reason for the majority of women seeing an oncologist would have been for the prescription of radiotherapy, rather than chemotherapy.

The original surgical case load breakdown was 1-9, 10-24, 25-49 and 'team' or ≥50 patients per year. Here 'team' indicates a group of breast surgeons who collaborate and work together in a breast clinic. This was used for some of the analyses reported in the 'Audit Report' (SBCFG et al, 1996). However, the breakdown given in Table 3.3 was the one used in the initial survival analysis given in that report and also by Twelves et al (1998a) to allow comparisons with the recently published paper by Sainsbury et al (1995a). The number of cases a surgeon dealt with per year was based on the total number of patients with breast cancer diagnosed under their care, including those women who did not eventually undergo surgery. Women who were recorded as having had surgery in their case notes, but the surgeon's name was not stated had to be excluded from analyses involving surgical case load.

SECTION B :

ANALYSES

CHAPTER 4 DETERMINATION OF MISSING VALUES AND CHARACTERISATION OF VARIABLES

4.1 GENERAL CHARACTERISTICS OF VARIABLES SELECTED FOR ANALYSIS

INTRODUCTION

This section presents some basic descriptive statistics for the variables chosen for analysis in the Breast Cancer Audit. Also discussed are associations for pairs of variables with cross-tabulations given in Appendix 4. All of the variables used in the analysis were categorical.

BASIC DESCRIPTIVE STATISTICS

Clinical Variables: Table A3.1 in Appendix 3 gives the breakdown of the numbers and percentages of cases in the different levels for the factors for clinical variables. Figures 4.1 to 4.5 illustrate these breakdowns. Note that 'NK' stands for not known.

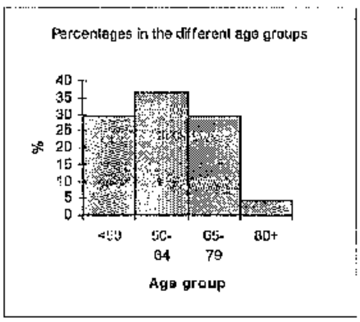


Figure 4.1: Percentages by age group.

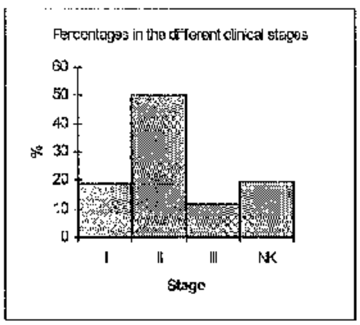


Figure 4.2: Percentages by clinical stage.

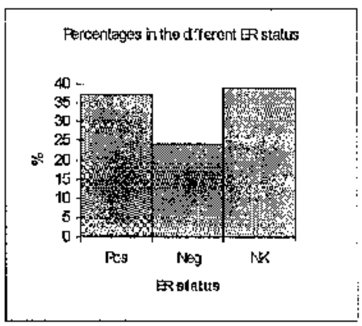


Figure 4.3: Percentages by ER status.

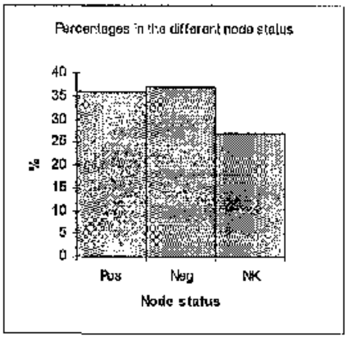


Figure 4.4: Percentages by node status.

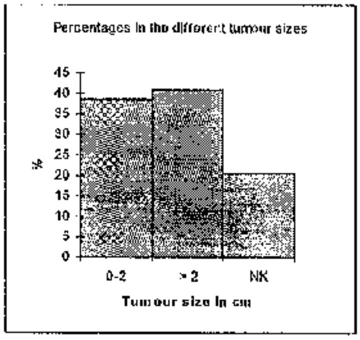


Figure 4.5: Percentages by tumour size.

For the 1619 surgical patients, both the mean and median ages at diagnosis were 58 years. The youngest and oldest women receiving surgery were aged 23 and 89 years respectively and the interquartile range was 48 to 67 years. The largest group for age was the 50 to 64 age group. Most of the tumours were clinical stage II, that is, either small tumours with node involvement or large tumours with no nodal involvement.

Menstrual status was collected for all but 85 (5.3%) women. This clinical variable was not examined further because it was found to be non-significant in the Cox's survival analysis (see Section 5.2.2).

Despite histological grade being an important prognostic factor for breast cancer (Miller et al, 1994), this variable was not included in the list of available clinical factors because 53% of the women did not have this information recorded. This decision is supported by Schemper & Smith (1990), who state that using covariate deletion is their chosen option when a large percentage, say 50%, of the data are missing.

Treatment Variables: Table A3.2 in Appendix 3 gives a corresponding breakdown of cases for the treatment variables. Figures 4.6 to 4.10 illustrate the percentage breakdowns.

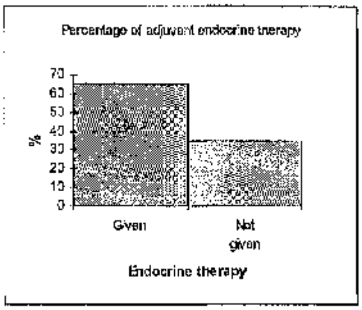


Figure 4.6: Percentages by adjuvant endocrine therapy.

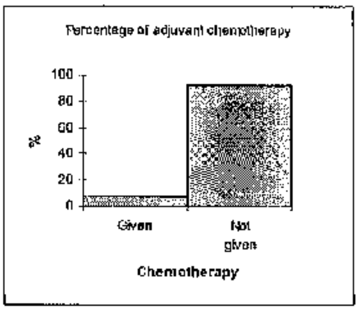


Figure 4.7: Percentages by adjuvant chemotherapy.

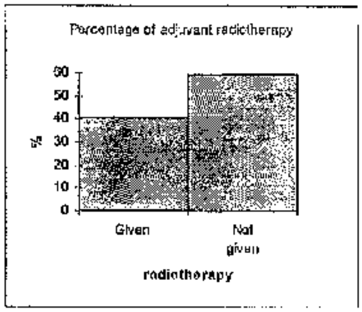


Figure 4.8: Percentages by adjuvant radiotherapy.

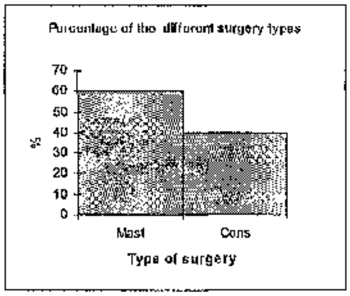


Figure 4.9: Percentages by type of surgery (Mast stands for mastectomy; Cons stands for conservation).

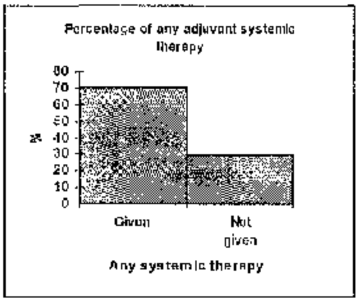


Figure 4.10: Percentages by adjuvant chemotherapy or endocrine therapy.

Mastectomy was the most common surgical procedure in this 1987 cohort, perhaps reflecting the fact that the majority of tumours were greater than 2 cm in size. However, the mastectomy group included the 117 women who had breast conservation, followed by a mastectomy within three months of the conservation surgery.

The majority of women getting some form of endocrine therapy received tamoxifen. Adjuvant chemotherapy was not widely prescribed for early breast cancer in 1987. The classification adjuvant chemotherapy or endocrine therapy is also known as any adjuvant systemic therapy in this thesis.

Service Variables: Table A3.3 in Appendix 3 provides a breakdown of the cases for each of the service factors. Figures 4.11 to 4.14 illustrate these breakdowns.

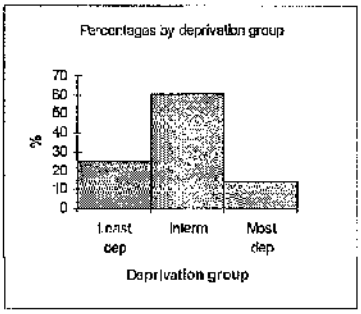


Figure 4.11: Percentages by deprivation group (dep stands for deprivation; Interm stands for intermediate).

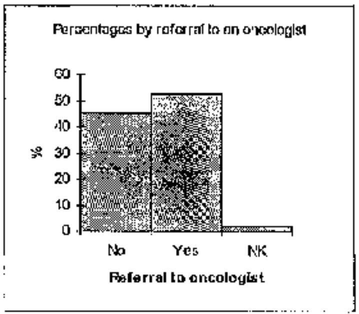


Figure 4.12: Percentages by referral to oncologist.

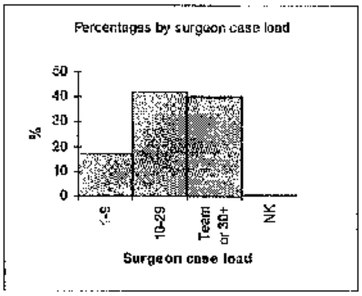
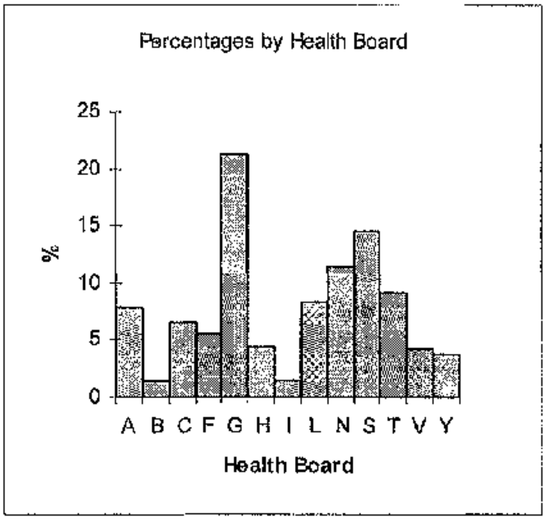


Figure 4.13: Percentages by surgeon case load.



A = Ayrshire & Arran
B = Borders
C = Argyll & Clyde
F = Fife
G = Greater Glasgow
H = Highland
I = Islands
L = Lanarkshire
N = Grampian
S = Lothian
T = Tayside
V = Forth Valley
Y = Dumfries & Galloway

Figure 4.14: Percentages by Health Board.

The fact that 25% of women were in the least deprived group, which was derived to be a quintile of the Scottish population, reflected the known higher incidence of breast cancer among women living in the less deprived areas (Harris et al, 1998). Only three categories were used, instead of five, to highlight the differences between the least (category I) and most deprived (category V) women more clearly.

There were 278 women in the surgeon case load grouping who managed only one to nine cases of breast cancer in 1987. Seventy-eight surgeons saw these women, with 23 surgeons only seeing one patient in that year. Thus 17% of the women in the cohort were operated on by surgeons who were relatively inexperienced with breast cancer, although there might be some underestimation of the case loads of surgeons who took up post or retired during the year of study (likely to be few in number).

The five Health Boards containing Cancer Centres are Highland, Grampian, Tayside, Lothian and Greater Glasgow respectively (Section 2.3.1) and are known in this thesis as Cancer Centre Health Boards (CCHB). The other Health Boards are known here as non-Cancer Centre Health Boards.

The number of women living in a CCHB was 885. Of these women, only 477 (53.9%) were actually operated on at the Cancer Centre.

ASSOCIATIONS BETWEEN PAIRS OF VARIABLES

The clinical variables were examined to investigate whether these variables, which represent the state of the tumour the clinician was faced with at the clinic, were related. A selection of relationships which were deemed to be interesting from a clinical point of view were also investigated. Table 4.1 presents the P values for the χ^2 tests of association which were performed.

C	**									
E	**	**								
N	**	**	**							
T	0.001	**	**	**						
H	0.001	**	**	**	**					
S	**	**	**	**	0.42	-				
R	**	**	-	-	-	-	-			
D	0.24	0.32	**	-	-	-	-	-		
CT	**	-	-	**	-	-	-	**	-	
TS	-	-	-	-	-	-	0.27	-	-	-
	A	**C**	**E**	**N**	**T**	**H**	**S**	**R**	**D**	**CT**

*Table 4.1: P values for χ^2 tests of association for various variables. Note that '**' indicates a P value <0.001; '-' means that the association was not tested. Age is given by A, clinical stage(C), node status(N), tumour size(T), ER status(E), Health Board(H), surgeon case load(S), referral to oncologist(R), deprivation group(D), chemotherapy(CT) and type of surgery(TS) respectively.*

All of the clinical variables were associated with each other and also there were differences in the levels of these variables amongst the Health Boards. These significant results for Health Board could be because there were differences in the proportions originally selected for surgery in the different Health Boards (data not given). This could lead to different distributions of the clinical factors for the patients treated in different Health Boards.

Neither age nor clinical stage at presentation were associated with the deprivation category assigned to the postcode of residence, although there was a significant association between deprivation and ER status. This is discussed further in Sections 5.2.2 and 5.2.3. Use of chemotherapy depended upon referral to an oncologist, age and node status. These results were expected with the majority of women receiving chemotherapy being aged under 50 years with node positive disease. This is the group of women, subsequently shown in the overview (EBCTCG, 1992), where chemotherapy has a beneficial survival effect over no use of chemotherapy (Section 2.3.2). The type of surgery performed was independent of the case load of the surgeon, although all of the clinical variables, except tumour size, were associated with surgeon case load.

Cross-tabulations of pairwise clinical variables: It is possible that some of the P values for the tests of association were only significant because of differences in the proportions of unknowns in the different levels. Tables A4.1 to A4.10 in Appendix 4 give the percentages in the cross-tabulations of the pairs of clinical variables. To examine whether different proportions of unknowns caused the significant results, tests of association on each of the pairwise-complete pairs of variables were performed (Table 4.2).

C	**			
E	0.001	0.13		
N	0.42	**	0.58	
T	0.09	**	0.001	**
	A	**C**	**E**	**N**

*Table 4.2: P values for χ^2 tests of association for pairwise-complete clinical variables. Note that '**' indicates a P value <0.001. Age is given by A, clinical stage(C), node status(N), tumour size(T) and ER status(E) respectively.*

Therefore, the observed significant associations for the pairs of variables: age by node status; age by tumour size; clinical stage by ER status and ER status by node status appeared to be due to differences in the proportions of unknowns in the different levels of the factors. Here, when all of the information was complete, the hypothesis of independence for each of these four pairs of factors could not be rejected. However, there were still significant differences between the numbers observed and expected for the different levels for the pairs of factors: age by clinical stage; age by ER status; clinical stage by node status; clinical stage by tumour size; ER status by tumour size and node status by tumour size. Some findings from the tables for these six cross-tabulations are now given.

From Table A4.1 in Appendix 4, there was a larger percentage of clinical stage I women in the under 50 age group, but a lower proportion of women with clinical stage III disease in this age group, than expected. There appeared to be larger number than expected of these stage III women in the age group 65-79.

There also appeared to be a larger percentage of ER negative women aged under 50 (Table A4.2 in Appendix 4) than in the older age groups. This agrees with Souhami & Tobias (1995; Section 2.3.2). However, there was a larger proportion of cases with ER status unknown for women aged over 65, especially those over 80 years.

There were many more women with pathological node negative disease with clinical stage I than expected from a statistical point of view, although not from a clinical point of view, with clinical node status being a component of clinical stage. That is, it would be expected that pathological node status would be related to clinical stage. Similarly, there were more women with pathological node positive disease whose clinical stage was stage II or III, especially III. There were more women with pathological node status not determined for women with clinical stage I disease or with clinical stage unknown (Table A4.6 in Appendix 4).

Similarly, there were many more women with pathologically small (≤ 2cm) tumours with clinical stage I than expected under the assumption of independence between these factors. However, with clinical tumour size being a component of clinical stage, it is not surprising from a clinical point of view that clinical stage was associated with pathological tumour size. There were more women with large tumours (>2cm) with clinical stage II or III. Again, there were more women than expected who had neither of these factors recorded (Table A4.7 in Appendix 4).

There were more women with small tumours which were ER positive than expected and more women with large tumours that were ER negative. There were also differences in the proportions with unknown tumour size across the levels of ER status (Table A4.9 in Appendix 4).

There were many more women with node negative disease who had small tumours than expected. Similarly, there was a larger percentage of women with large tumours that also had nodal involvement than would be expected by chance (i.e. under the assumption of independence). This supported the findings of Carter et al (1989; Section 2.3.2). There were more women with both of these factors missing than expected (Table A4.10 in Appendix 4).

Cross-tabulations of the clinical variables with surgeon case load: The associations of surgeon case load with the clinical variables were then examined to investigate whether significant results were due to differences in the proportions in the unknowns in the different levels. The cross-tabulations for each of the clinical variables with surgeon case load are given in Appendix 4 (Tables A4.11 to A4.15), although there was no evidence to reject independence of surgeon case load by pathological tumour size. These were examined because surgeon case load and specialisation has been linked to survival in several studies (Gillis & Hole, 1996; Sainsbury et al, 1995a; See Section 5.2.3).

There were more women aged under 65 years who were managed by a surgeon in the Team or 30 or more cases per year group. This group of surgeons (high case load) saw many more women with clinical stage II disease, but also more ER positive tumours. However, the most striking observation from the tables given in Appendix 4 is that this high surgeon case load group had much lower proportions of unknowns in the clinical variables. When the unknowns were excluded from the analyses, the pairs of factors became non-significantly associated (P=0.50 for clinical stage; 0.08 for ER status; 0.20 for node status). This suggests that the observed differences in the levels for the known factors by surgeon case load were because the high case load group was managing more women whose disease had been better staged. The possible influences on survival are discussed further in Section 5.2.3.

4.2 PATTERNS OF MISSING VALUES AND LOG-LINEAR MODELLING

The last section showed that the clinical variables were inter-related. The four variables: clinical stage, pathological node status, pathological tumour size and ER status have all been shown to have prognostic importance in terms of survival from breast cancer (Miller et al, 1994). There were some cases where this information was missing for these factors, and the patterns of the missing values in these data are now examined, both descriptively and by log-linear modelling.

4.2.1 THE VARIABLES: CLINICAL STAGE, PATHOLOGICAL NODE STATUS, PATHOLOGICAL TUMOUR SIZE AND OESTROGEN-RECEPTOR (ER) STATUS

For the 1619 patients, Table 4.3 shows the numbers and percentages which were known and missing for each variable.

Variable	Number (%) Known	Number (%) Missing	Total
Clinical Stage	1302 (80.4)	317 (19.6)	1619
Pathological Node Status	1184 (73.1)	435 (26.9)	1619
Pathological Tumour Size	1287 (79.5)	332 (20.5)	1619
ER Status	990 (61.1)	629 (38.9)	1619

Table 4.3: Numbers and percentages of known and missing values for each of the four variables of interest.

Only 578 (35.7%) of the women had all four variables known. These cases comprise the group known as the 'complete cases'. The number of cases where there was only one, two or three of the four variables missing were 546 (33.7%), 350 (21.6%) and 113 (7.0%) respectively. There were only 32 (2%) of the 1619 women who had no information recorded for any of the four variables. Therefore, about 30% of cases had two or more of these variables missing.

To investigate whether the missing values are related, log-linear modelling is performed. Some theory is now given for this technique.

4.2.2 THEORY OF LOG-LINEAR MODELLING

In the situation where several categorical variables have been cross-classified to give a contingency table, it is the counts of the individuals falling into the cells of this table that are modelled.

Let x_{jklm} represent the frequency of the (j,k,l,m)th cell, where the four variables take the values:

ER status (E): $j = 1, 2$

Tumour size (T): $k = 1, 2$

Node status (N): $l = 1, 2$

Clinical stage (C): $m = 1, 2$,

where each factor has level 1 meaning known and level 2 meaning missing.

Also, let ϑ_{jklm} represent the probability that a randomly selected individual falls into cell (j,k,l,m). Let the vectors $\underline{x}$ and $\underline{\vartheta}$ represent the 16 x_{jklm} frequencies and the 16 probabilities ϑ_{jklm} respectively, for simplicity.

Considering the total sample size to be fixed ($n = 1619$), the sampling distribution from which these counts are assumed to come can be shown to be multinomial (Dobson, 1990), since the assumption is made that the original counts are from independent Poisson variables, but these are constrained by the total fixed sample size and are thus from a multinomial sampling distribution.

The probability density function for the vector $\underline{x}$ conditional on $\sum_{jklm} x_{jklm} = n$, is given by

$$f(\underline{x}; \underline{\vartheta}|n) = n! \prod_{jklm} \frac{\vartheta_{jklm}^{\;x_{jklm}}}{x_{jklm}!}$$

with the constraints $0 \le \vartheta_{jklm} \le 1$ and $\sum_{jklm} \vartheta_{jklm} = 1$.

The expected value of a particular element of $\underline{x}$ is given by

$$E(x_{jklm}) = n\vartheta_{jklm}. \quad \text{(Eq 4.2.2\_1)}$$

In a log-linear model the logarithms of the expected frequencies are assumed to have a linear form

$$\log E(x_{jklm}) = \sum_{r=1}^{16} \beta_r^{\,T} z_r\,, \quad \text{(Eq 4.2.2\_2)}$$

where $\underline{\beta}$ is an 16 x 1 vector of unknown parameters and $\underline{z}$ is an 16 x 1 vector of indicator variables, z_r.

One hypothesis of interest is that of complete marginal independence. When this holds,

$$\vartheta_{jklm} = \vartheta_{j...}\vartheta_{.k..}\vartheta_{..l.}\vartheta_{...m}, \quad \text{(Eq 4.2.2\_3)}$$

where $\vartheta_{j...} = \sum_{klm} \vartheta_{jklm}$, the marginal probability of being in ER status level j for $j = 1, 2$.

Similarly, $\vartheta_{.k..}$, $\vartheta_{..l.}$ and $\vartheta_{...m}$ are the marginal probabilities for the tumour size, node status and clinical stage variables respectively.

Therefore, if the hypothesis of global independence holds then, from Eqs 4.2.2_1 and 4.2.2_3, the expected frequencies are given by

$$E(x_{jklm}) = n\vartheta_{j...}\vartheta_{.k..}\vartheta_{..l.}\vartheta_{...m}.$$

This can be written in terms of a main effects log-linear model with the structure

$$\log E\left(x_{jklm}\right) = \mu + \alpha_j z_2 + \beta_k z_3 + \gamma_l z_4 + \delta_m z_5,$$

with the appropriate constraints on the parameters $\alpha_j, \beta_k, \gamma_l$ and δ_m.

Similarly, the maximal (or fully-saturated) model can be written as the model with a constant μ; four main effects; six two-way interaction terms; four three-way interaction terms and a four-way interaction term.

Since the sampling distribution is assumed to be multinomial with n fixed, the log-linear model must include the corresponding parameter μ. In this analysis, a corner-point constraint is imposed, with cell (1,1,1,1) where all of the variables are known, taken as the reference cell. Also, all of the terms including the first levels of the variables are set to be zero. Thus, $\alpha_1 = 0$, $\beta_1 = 0$, $\gamma_1 = 0$, $\delta_1 = 0$, $(\alpha\beta)_{11} = 0$, $(\alpha\beta)_{12} = 0$ etc.

Every variable has only two levels and since all of the terms involving the first levels are set to zero, there is no need to include the subscripts on the parameter terms. Therefore, the unknown parameters given in Eq 4.2.2_2 can be written as:

$$\beta_1 = \mu;\quad \beta_2 = \alpha;\quad \beta_3 = \beta;\quad \beta_4 = \gamma;\quad \beta_5 = \delta;\quad \beta_6 = (\alpha\beta);\quad \beta_7 = (\alpha\gamma);$$
$$\beta_8 = (\alpha\delta);\quad \beta_9 = (\beta\gamma);\quad \beta_{10} = (\beta\delta);\quad \beta_{11} = (\gamma\delta);\quad \beta_{12} = (\alpha\beta\gamma);$$
$$\beta_{13} = (\alpha\beta\delta);\quad \beta_{14} = (\alpha\gamma\delta);\quad \beta_{15} = (\beta\gamma\delta);\quad \beta_{16} = (\alpha\beta\gamma\delta).$$

The main effects are represented by α for ER status; β for tumour size; γ for node status and δ for clinical stage respectively. The two-way, three-way and four-way interactions are denoted, for example, by $(\alpha\beta)$, $(\alpha\beta\gamma)$ and $(\alpha\beta\gamma\delta)$ respectively.

4.2.3 RESULTS OF LOG-LINEAR MODELLING

For the four variables of interest in the Breast Cancer Audit data, the breakdown of the observed values in the 16 cells is given in Table 4.4.

ER Status	Tumour Size	Node Status	Clinical Stage	
			Known	Missing
Known	Known	Known	578	81
		Missing	133	26
	Missing	Known	100	30
		Missing	32	10
Missing	Known	Known	232	73
		Missing	124	40
	Missing	Known	65	25
		Missing	38	32

Table 4.4: Observed numbers of cases in each of the 16 cells.

GLOBAL INDEPENDENCE

For the complete marginal independence model, the likelihood ratio statistic had a χ^2 value of 137.12 on 11 degrees of freedom (df) with $P<0.0001$. Therefore, the hypothesis that the missing values in the variables were independent of each other could be rejected. Thus, examination of the interactions between the variables was necessary and a search was made to try to identify the 'best' log-linear model fitting the data.

OTHER MODELS

The technique of backward elimination (Armitage & Berry, 1994) was used for the model selection. Only hierarchical models were sought. Table 4.5 below shows the results of this process starting with the maximal model, but only gives the highest generating classes for the model fitted at each step. A generating class is a way of describing what terms are in the model and is best illustrated through an example.

Example: Step 4 from Table 4.5 below has generating classes given by

E*T*C, T*N*C, E*N,

where E, T, C and N stand for ER status, tumour size, node status and clinical stage respectively.

E*T*C means that all of the terms $(\alpha\beta\delta),(\alpha\beta),(\alpha\delta),(\beta\delta)$, α,β,δ and μ are included in the model.

T*N*C with E*T*C means that the extra terms $(\beta\gamma\delta),(\beta\gamma),(\gamma\delta)$ and γ are also included in the model.

E*N with both E*T*C and T*N*C means that the extra term $(\alpha\gamma)$ is also included in the model.

Due to the hierarchical structure, only the highest order terms were assessed for removal at each step. Thus for example at Step 4, despite the fact that there were 13 terms in the model, only the interactions $(\alpha\beta\delta),(\beta\gamma\delta)$ and $(\alpha\gamma)$ were examined to see whether they could be removed from the model. At each step, the term which gave the smallest non-significant change of the likelihood ratio was removed. The model fitted (shown using the generating classes representation), along with the P values for removal of the highest order terms, are presented for each step in Table 4.5.

BEST FIT MODEL

Step 7 below shows that the best fit model included all of the two-way interactions, except for the one between node status and clinical stage. The likelihood ratio goodness of fit statistic for this model was 7.64 as χ^2_6 (P value =0.266) and hence this model could not be rejected.

Terms examined	Change in LR	P Value
Step 1: Generating Class: E*T*N*C		
$(\alpha\beta\gamma\delta)$	2.812	0.0936 - removed
Step 2: Generating Classes: E*T*N, E*T*C, E*N*C, T*N*C		
$(\alpha\beta\gamma)$	0.025	0.8755
$(\alpha\beta\delta)$	0.504	0.4778
$(\alpha\gamma\delta)$	0.000	0.9828 - removed
$(\beta\gamma\delta)$	1.105	0.2932
Step 3: Generating Classes: E*T*N, E*T*C, T*N*C		
$(\alpha\beta\gamma)$	0.024	0.8768 - removed
$(\alpha\beta\delta)$	0.509	0.4757
$(\beta\gamma\delta)$	1.105	0.2932
Step 4: Generating Classes: E*T*C, T*N*C, E*N		
$(\alpha\beta\delta)$	0.498	0.4806 - removed
$(\beta\gamma\delta)$	1.157	0.2820
$(\alpha\gamma)$	46.974	<0.0001
Step 5: Generating Classes: T*N*C, E*N, E*T, E*C		
$(\beta\gamma\delta)$	0.928	0.3353 -removed
$(\alpha\gamma)$	46.728	<0.0001
$(\alpha\beta)$	7.623	0.0058
$(\alpha\delta)$	26.143	<0.0001
Step 6: Generating Classes: E*N, E*T, E*C, T*N, T*C, N*C		
$(\alpha\gamma)$	46.725	<0.0001
$(\alpha\beta)$	7.621	0.0058
$(\alpha\delta)$	26.141	<0.0001
$(\beta\gamma)$	4.869	0.0273
$(\beta\delta)$	17.009	<0.0001
$(\gamma\delta)$	3.378	0.0661 - removed
Step 7: Generating Classes: E*N, E*T, E*C, T*N, T*C **BEST FIT MODEL**		
$(\alpha\gamma)$	51.204	<0.0001
$(\alpha\beta)$	7.326	0.0068
$(\alpha\delta)$	30.619	<0.0001
$(\beta\gamma)$	5.908	0.0151
$(\beta\delta)$	18.048	<0.0001

Table 4.5: The steps in the backward elimination process with the highest generating classes for each set of variables, along with the P values for removal of the highest order terms from model, based on change in the likelihood ratio (LR). Note that E stands for ER status, tumour size(T), node status(N) and clinical stage(C).

The parameter estimates with their standard errors for the best fit model are given in Table 4.6. All of the terms have 1 df. No standard error was calculated for the constant term because the multinomial sampling distribution was assumed and so this term was considered to be fixed.

Term	**Parameter**	**Parameter Estimate**	**Standard Error (se)**
Constant (all known)	μ	6.3497	---
ER status missing	α	-0.9036	0.0716
Tumour size missing	β	-1.7418	0.0949
Node status missing	γ	-1.4300	0.0837
Clinical stage missing	δ	-1.8752	0.0962
ER status missing by tumour size missing	$(\alpha\beta)$	0.3495	0.1287
ER status missing by node status missing	$(\alpha\gamma)$	0.8193	0.1148
ER status missing by clinical stage missing	$(\alpha\delta)$	0.7072	0.1278
Tumour size missing by node status missing	$(\beta\gamma)$	0.3328	0.1355
Tumour size missing by clinical stage missing	$(\beta\delta)$	0.6236	0.1437

Table 4.6: Parameter estimates and their standard errors for the terms in the best fit model.

All of the parameter estimates for the interactions represented being missing compared with being known. Since all of these parameter estimates were positive, then this implied that there was a positive association between the chances of the values being missing in both variables in each of the two-way interactions, except for the non-significant interaction between node status and clinical stage. Thus, it was more likely that the second variable was missing if the first variable was missing than when the first variable was known.

Table 4.7 below gives the estimated fitted values from the 'best' model for the 16 cells along with the observed values from Table 4.4. It can be seen that the estimated expected numbers of cases falling into each of the cells are fairly close to the observed numbers.

ER Status	Tumour Size	Node Status	Clinical Stage							
			K				M			
K	K	K	O	578	E	572.30	O	81	E	87.75
		M	O	133	E	136.96	O	26	E	21.00
	M	K	O	100	E	100.27	O	30	E	28.68
		M	O	32	E	33.47	O	10	E	9.57
M	K	K	O	232	E	231.85	O	73	E	72.10
		M	O	124	E	125.89	O	40	E	39.15
	M	K	O	65	E	57.62	O	25	E	33.43
		M	O	38	E	43.64	O	32	E	25.32

Table 4.7: Observed (O) and expected (E) numbers of cases in each of the 16 cells. Note that K and M stand for known and missing respectively.

LOOKING AT THE SUB-TABLES

To obtain a general picture as to why the significant two-way interactions were needed in the model it is possible to look at the sub-tables of observed values for the pairs of variables (easily obtainable from Tables A4.1 to A4.10 in Appendix 4). The percentages for being missing in the second variable given that the first variable was missing compared with being missing in the second variable given that the first variable was known are given for the six pairs of variables in Table 4.8. The P values for the differences between the proportions and for the corresponding interactions in the log-linear model are also given.

The associations between missing values in the pairs of variables are apparent from examination of the percentages. It can be seen that the proportion missing in the second variable when the first variable was missing is always larger than the proportion missing in the second variable given that the first variable was known. Caution is needed when interpreting the univariate results because the sub-tables are not derived from the log-linear model. The percentages are based only on the collapsed sub-tables of observed values for the pairs of variables and as such are only illustrative.

The reason for caution can be demonstrated by looking at the observed proportions for the clinical stage by node status sub-table. Simple comparison shows that the node status was missing in 34% of the cases when clinical stage was also missing but was missing in only 25% of cases when clinical stage was known. This observed difference was statistically significant (P=0.001) for the univariate test of differences in the proportions (Table 4.8). However, the interaction between this pair of variables dropped out of the log-linear model at Step 6, with a P value of 0.066 (Table 4.5). This P value is conditional on the other variables being in the multivariate model. There is, perhaps, some weak evidence to suggest that the missing values were related, although this was not statistically significant at the 5% level.

First variable	% Second variable missing given first variable missing	% Second variable missing given first variable known	P value test for differences in proportions	P value in log-linear model
ER status	Node status 37	Node status 20	<0.0001	<0.0001
ER status	Tumour size 25	Tumour size 17	0.0001	0.007
Node status	Tumour size 26	Tumour size 19	0.002	0.015
Clinical stage	ER status 54	ER status 35	<0.0001	<0.0001
Clinical stage	Tumour size 31	Tumour size 18	<0.0001	<0.0001
Clinical stage	Node status 34	Node status 25	0.001	0.066*
* not included in the log-linear model				

Table 4.8: Percentages missing in the second variable given that the first variable was either missing or known, along with the P values for testing that the proportions were the same in the univariate sub-tables and P values for the terms, conditional on the other terms, in the multivariate log-linear model.

CONCLUSIONS

The aim of this analysis was to find out whether there were any associations among the missing values in the variables. The hypothesis of no association among the binary variables was rejected. It was found that all two-way interactions were necessary in the model, except the interaction of node status by clinical stage. Thus, there was pairwise dependence between each pair of variables, although there was no evidence to suggest that each two-way interaction was affected by the values of the third and fourth variables. This interpretation holds except for the interaction of node status with clinical stage. For this term, there was insufficient evidence to reject conditional independence between these two variables, given the third and fourth variables.

In general, a log-linear model with a significant two-way interaction of two factors, each at two levels, suggests that the expected numbers of cases at level 2 of factor 1 are different for the two levels of factor 2. In this analysis, for example, there was a significant interaction for clinical stage with ER status, both at two levels either known or missing. Here the two-way interaction indicated that the expected number of cases with ER status missing was associated with whether clinical stage was known or missing.

CLINICAL INTERPRETATIONS

From a clinical point of view, it is not entirely clear how the missing values in the variables are expected to be associated with one another. Having discussed this matter with clinical colleagues prior to the analysis, two possible opposing hypotheses were given:

(i) it may be argued that there will be associations among the missing values of the three pathological variables: node status, tumour size and ER status. However, no associations are expected between whether or not clinical stage is missing with these three pathological variables being missing, except possibly with node status.

The reasoning behind this hypothesis is that there were three individuals involved in the process of recording the information about these four clinical variables in 1987. The surgeon determined clinical stage in his clinic prior to surgery and may or may not then have entered the details onto the case notes. The pathologist examined any material excised during the operation and recorded the pathological features (node status and tumour size) of the tumour. Thus, it might be expected that if one of these were missing, then so would the other. ER status was determined by the biochemist *if* a specimen was sent from the pathologist. However, there is a possibility that node status might be associated with clinical stage because the node status in the axilla can only be recorded by the pathologist *if* the surgeon actually removed some nodes from the axilla as part of the surgical procedure, perhaps because of the clinical node status.

(ii) it might be expected that all of the variables will be associated with one another in terms of the missing values.

The reasoning behind this hypothesis is that institutions may have agreed protocols, or at least informal practice agreements, for management of women with breast cancer. Thus, you might expect that hospitals which are less systematic in recording clinical information, may also have less well defined protocols for recording pathological data. The ideal situation is where there are multidisciplinary teams involved at all stages of the care of the woman and where all information is recorded by all of the specialists involved.

The fact that all of the two-way interactions were significant in the model, except for the clinical stage by node status interaction, appears to support the second hypothesis more strongly than the first hypothesis. For example, it was more likely that ER status and tumour size were missing when clinical stage was missing than when it was known. The non-recording of clinical stage had a strong bearing on the non-recording of the three pathological factors, although less so on node status. Overall, from Table 4.3 in Section 4.2.1, ER status was missing in 39% of cases; tumour size was missing in 21% of cases and node status was missing in 27% of cases. However, when only the women when clinical stage was not noted in the case notes (i.e. 317 cases) were included, these figures rose to 54%, 31% and 34% respectively, as can be seen in Table 4.8.

4.3 GENERAL DISCUSSION OF METHODS FOR HANDLING MISSING VALUES IN COVARIATES

INTRODUCTION

The last two sections examined the general characteristics of some of the variables in the Breast Cancer Audit data and also patterns of missing values in the four main prognostic factors. For each of the factors, extra categories to represent the cases with unknown values for each factor were created. This was the approach used in the analysis of the survival data (Twelves et al, 1998a). However, there are other techniques that can be used when analysing data with missing values in some of the covariates, as described below. Any discussions about the different techniques tend to focus on their applicability to analysis of survival data. Firstly, however, possible structures for missing values in data are reviewed. All statistical methods are likely to be affected when underlying assumptions made about the structure of any missing data are not valid.

STRUCTURE OF MISSING DATA

Missing Completely At Random (MCAR): The mechanism of missing values is said to be MCAR when the observations that are missing do not depend on any of the data, either those which are known or those which are missing.

Missing At Random (MAR): Data are said to be MAR when the observations that are missing do not depend on any of the unobserved values, either in the variable that is missing or in any other variable, but may depend on observed values in other variables.

Non Missing At Random (non-MAR): If, however, the probability that an observation is missing depends on its unobserved true value or on the true value of any other variable with missing information, then this mechanism for missing values is said to be non-MAR.

Patterns of Missing Values: There may be observed patterns among the missing values for several variables. However, this may not actually mean that the data are non-MAR.

Vach (1997) suggests the need to examine the assumption of the data being MAR using sensitivity analyses, but he makes the point that it is not possible to know if this assumption is valid using the available data. However, background subject knowledge may help to determine whether or not the assumption is reasonable.

METHODS OF ANALYSING DATA WITH MISSING VALUES

A number of techniques are now described. The first four are easily implemented using standard software. The remaining four are more complicated methods which require specialist software or fairly advanced programming skills.

Complete Cases Analysis: This method is the simplest approach of all (Greenland & Finkle, 1995). Here, only cases with complete information for all of the covariates are retained in the analysis. Cases with missing data for any of the covariates are simply discarded. This is very wasteful as it throws away information that has been recorded for some of the other covariates. When there is a large number of covariates, the number of cases that have to be excluded can be substantial, even if there are relatively few missing values for each covariate.

When this method is used, it assumes that there is no bias introduced by using only a subgroup of cases which has all of the information known and that this subgroup is representative of the whole population. However, this is a strong assumption to make and the estimates obtained based on these cases alone may be very biased (Schemper & Smith, 1990). Vach & Blettner (1991) investigate the situation of missing values in case-control studies. Using a simple context, they demonstrate that the estimate obtained for the odds ratio is not biased when MCAR can be assumed to be valid, but is biased when the data are MAR.

The complete cases method also produces estimates which have higher than necessary variances (Greenland & Finkle, 1995). The technique can be applied to survival analysis and is often the suggested method when there are only a few missing values in the data (Schemper & Smith, 1990). Whether the bias observed for MAR data in the case-control framework (Vach & Blettner, 1991) would be evident in survival analysis is not clear.

Available Cases Analysis: This method is another simple approach. It is described by Little (1992) in the context of multivariate normal data. Here the estimate of every element in the variance-covariance matrix is obtained separately. The value of element

(j,k) is estimated using the data which are complete for both variables j and k. One problem with this approach, however, is that the variance-covariance matrix is not necessarily positive-definite. This is a problem when the covariates are highly correlated. It is not obvious how this available cases approach could be applied easily to survival data since values of parameters in a model are not independent of the other variables included in the model.

Analysis Using Indicators For Missing Data: Greenland & Finkle (1995) outline this simple approach where indicator variables, m_j, to indicate missing values are created for every covariate, x_j, which contains some unknown information. Both m_j and x_j are then included in the analysis in the following manner. If x_j is missing then $m_j = 1$; otherwise $m_j = 0$. Then m_j is simply added to the model, whereas the variable x_j is replaced by the product $(1 - m_j)x_j$. The method is described for the regression problem, but the technique could be used in survival analysis.

Using this approach in the regression problem, information is obtained for the regression parameters, based on the subjects with known data. Whilst the extra terms involving m_j are used to obtain the regression fit, the parameters obtained for them are not reported when the results of the fit are given. Greenland & Finkle (1995) state that the estimates obtained for this method can be biased.

They also point out that when only one variable contains missing data, this method is the same as adding an additional category to represent these unknowns in the factor. This is the next method discussed.

Analysis Using Additional Categories For Missing Data: This method is another simple approach and was the one used in the survival analysis of the Breast Cancer Audit data. Additional levels were added to each of the factors with missing data to represent a category of unknowns in each variable. In their paper regarding missing values in case-control studies, Vach & Blettner (1991) demonstrate that the estimate for the odds ratio is biased for all of the different scenarios for the missing data mechanisms, including MCAR and MAR. They do not discuss the implications of these results for other study designs.

Imputation Methods: The idea behind imputation is to estimate and assign values for the missing data using the known data. There are several ways of obtaining these estimates.

(i) The simplest approach is to replace all of the missing values for a covariate with the overall mean for this variable based on the known values. One problem with this approach is that the variance will be underestimated if many missing values are allocated the mean value.

(ii) A slightly improved method involves the use of conditional means. Suppose variable j has some cases which are missing, but variable k is known for those cases. Then different estimates of variable j are obtained for the different values of k based on the cases where variable j and variable k are known. Usually, linear regression is used to obtain these estimates. For example, suppose age is not known for some people, but sex is known for all. Then the average ages for males and females would be calculated from the available data. Males with missing values for age would be given the average age for men and similarly, women with missing data on age would now take the female average age, obtained from those with age known.

Usually the assumption of MCAR is necessary. The theory has been developed for linear regression and is described in Little & Rubin (1987). The variance-covariance matrix is underestimated by the sample variance-covariance matrix. This method of imputing values based on conditional means could be used for survival analysis if a suitable model can be developed for the data. This would be similar to the approach of Schluchter & Jackson (1989) described below.

(iii) Vach & Blettner (1991) present a simple technique of filling the cells of a contingency table in the case-control context. They use knowledge of the proportions for the known observations to impute values for the missing data. They point out that this method can only be utilised when the assumption of MAR can be made. This method, along with all of those described above, is an ad hoc method, requiring specially derived formulae. The remaining imputation techniques and other methods discussed relate to modelling the data.

(iv) The Probability Imputation Technique was described initially in the paper by Schemper & Smith (1990) and updated in Schemper & Heinze (1997). The technique is presented only for binary variables, taking values 0 or 1. Based on the data for known cases, the probability, π, of getting a 1 is calculated. The missing values are then given either the value $1-\pi$ or π, instead of 0 or 1, depending on the values of the other covariates. It is similar to the conditional means approach.

Schemper & Heinze (1997) point out that their method is only to be used with binary 0/1 coded variables. It is not clear whether it would be possible to generalise this technique to categorical variables with more than two levels. Schemper & Smith (1990) say that the technique can be applied to the Cox model. In all situations, the method needs the assumption of MAR.

(v) In Multiple Imputation, rather than using average values to fill missing data, a set of imputed values is produced, possibly assuming a known distribution or conditional known distributions. For each of the cases with missing values, a random value is selected from the appropriate distribution. Parameter estimates are then obtained based on all cases, i.e. on the known cases and the missing cases, which have all been replaced by the random values. The process is then repeated many times, thus generating a set of parameter estimates. These are then combined in a variety of ways, details of which can be found in Little & Rubin (1987).

Maximum Likelihood Approach: Here, the approach is to use maximum likelihood (ML) to model the known data to obtain estimates for the parameters of the model and hence for the missing values simultaneously. To apply ML theory, a parametric model must be used for the joint distribution of the covariates (Vach & Blettner, 1995). They note that it is not always possible to obtain this. Little & Rubin (1987) question whether using the information matrix to calculate standard errors is valid in this context and point out that large sample normality of the likelihood function may not apply as the data will not necessarily be an independent, identically distributed sample.

(i) The simplest application of the ML approach involves the situation where the data have some special patterns of missingness and the likelihood function can be factored into components which can be easily maximised. Little (1992) gives an example.

(ii) When there are no specific patterns in the data, the likelihood function cannot be factored and it is necessary then to use an iterative maximisation procedure. Possibilities include the Newton-Raphson and the EM algorithms. These methods are computer intensive.

Little (1992) points out that the ML approach is valid for MAR data, but can also be adapted for some situations involving non-MAR data. The technique does not perform well when there is only a small number of cases and is mainly recommended for use with large samples.

Little (1992) states that this method is not very useful when the covariates with missing values are categorical. Schemper & Smith (1990) point out that it may not be possible to use ML for survival analysis due to the fact that the Cox model uses partial likelihood. Vach (1997) also comments on this fact and makes use of a logistic model with grouped survival data to permit use of the ML theory.

An Explicit Model: Schluchter & Jackson (1989) attempt to incorporate missing data in categorical covariates into a survival analysis. This paper uses ML theory. The joint distribution of the survival data and the covariates is modelled using a flexible log-linear model. The covariates are assumed to have a multinomial distribution, determining the probability of the observation taking a certain value, either known or missing. They state that they assume that the "hazard function, conditional on the covariates, is a stepwise function over disjoint time intervals. Thus, the survival times have piecewise exponential distributions". To employ this method, the assumption that the data are MAR must be made.

DISCUSSION

Not all of the techniques described above could easily be applied to survival data. The fact that the Cox model is semi-parametric in nature means that no explicit joint distribution can be written down between the survival times and the covariates. That is, the hazard for a set of covariates can be modelled, but the underlying baseline hazard function cannot. Thus, any approach which uses ML theory would be difficult to adapt for the Cox model. The semi-parametric nature of the Cox model also presents a problem for using Multiple Imputation. Schluchter & Jackson (1989) fitted a fully parametric model. Schemper & Heinze (1997) state that the probability imputation technique can be used with the Cox model but they also point out that only binary variables can be used in their method. No indication is given as to whether it would be possible to extend this to non-binary categorical variables.

Vach & Blettner (1991) point out that the method of additional categories is used extensively in published literature in many circumstances including case-control studies, despite the fact that the estimates obtained for this method are biased for all missing data mechanisms.

For all of the methods described above, the data need to be MAR, otherwise the results will potentially be biased.

4.4 POSSIBLE APPROACHES TO THE PROBLEM OF MISSING VALUES IN THE BREAST CANCER AUDIT DATA

INTRODUCTION

In the last section, various methods were presented for handling missing values in general situations. Here, the methods used to analyse the Breast Cancer Audit survival data are described.

The pattern of the missing values has already been presented in Section 4.2.3. It was shown that all but one of the two-way interactions for the four prognostic factors (clinical stage, node status, tumour size and ER status) were significant in a log-linear model describing the probability of being missing for each variable. Thus, a patient was more likely to have a missing value in node status when the tumour size was also missing. However, this only provided information about the structure of the missing data and not whether having missing data affected the survival results.

In the case of the Breast Cancer Audit survival data, it is not clear whether the missing information in the four main clinical prognostic covariates were MAR. If, for example, node status was missing because of the true unknown value of tumour size, then the data would be non-MAR. However, if the data were missing in both of these variables because of an external policy of recording pathological information within the hospital, then perhaps the Breast Cancer Audit data can be assumed to be MAR. It seems unlikely that the assumption of MCAR could be taken to be valid.

In the analysis of the data presented in the paper by Twelves et al (1998a), the missing values were included in the analysis using the method of additional categories. The decision to use this method was made for two reasons. The first was that use of the complete cases method was thought to be unacceptable due to the large amount of missing data present in the four main prognostic clinical factors (64%). The second reason was that many other studies have used the additional categories method (these are discussed in the next section, along with the other techniques employed in survival analysis of breast cancer data). It is not clear whether the estimates obtained and reported by Twelves et al (1998a) will be greatly biased using this method for analysing missing data in survival analysis.

To try to address this issue, the results obtained for a complete cases analysis, the simplest alternative to the technique employed, were compared with those from the additional categories method to see if there was any consistency between the findings for the two methods. The results of this comparison are given in Section 5.4.2. The main objection to the complete cases analysis method, that of wasting too much information, has already been stated. It is also not clear whether the subgroup of women with complete information would have been representative of the whole population. If it were not, then bias could be introduced into the estimates of hazard ratios. This would not be important if the survival for the women in the subgroup for a particular combination of factors was representative of the survival of all women in that particular combination. Another objection to this method of handling missing values is that there is the possibility of a loss of power due to the substantially reduced sample size thus reducing the possibility of detecting any relationships.

Another approach looked at briefly is an ad-hoc combination of the available cases, complete cases and additional categories methods. Here, two further subgroups of the 1619 surgical cases included in the Breast Cancer Audit were considered (known here as partial-complete cases analysis). The subgroups included those cases where:

(i) both node status and tumour size needed to be known, but the other factors clinical stage and ER status could be either known or missing; and

(ii) all three pathological factors (ER status, node status and tumour size) had to be known but clinical stage could be either known or missing.

The clinical reasons for looking at (i) were that node status and tumour size have been known to be important prognostic factors for a long time (Blamey et al, 1979) and are more likely to have been recorded than ER status, which is harder to determine, with several different methods used to analyse the specimens (Barnes et al, 1996). Indeed, examination of the recording of the Breast Cancer Audit data, given in Section 4.2.1, showed that nearly 40% of the cases did not have this information available.

The subgroup of cases given by (ii) was examined because clinical stage might not have been recorded in the notes if the pathological information was available. Also, the misclassification of clinical stage is a known problem (Bundred et al, 1994) and the agreement of clinical and pathological findings is not always very high (Brewster et al, 1996b). It was important to obtain a model based on the more important prognostic pathological factors and compare the estimates from this analysis with those based on all 1619 cases and just the 'proper' complete cases analysis.

The results from fitting Cox models to these subgroups of cases were compared with those from the additional categories and the complete cases analyses and are reported in Section 5.4.2.

It was decided that it was not feasible in the time available to apply the other methods discussed in the last section to the Breast Cancer Audit survival data. The main reason was the complexity and the need for specialist software. It is not certain that any of the techniques, other than the probability imputation technique (PIT; Schemper & Smith, 1990) could be applied to the Cox model. The PIT seems to require that all of the variables with missing information are binary. Three of the four clinical factors with missing information are indeed binary (ER status, node status and tumour size), but clinical stage is not, although it could be made binary. However, the technique still needs the missing data to be MAR. Whether this assumption is valid for the Breast Cancer Audit data remains unclear. However, this method would probably be the most appropriate to try to implement if the computer software and time were available.

No easy solution exists to the issues of missing data. However, analysis of retrospective cancer audit data still needs to be performed to provide some idea of the survival chances of the people with cancer in Scotland and the variation in survival across levels of treatment and other factors, such as Health Board. Therefore, it is necessary to move forward tentatively, and provide a set of results to inform decision making. Of course, any proposed solution must be interpreted cautiously in light of the potential bias, due to the imperfect nature of the data. In the next section, how other relevant literature dealt with missing values in survival data is discussed.

4.5 EXAMINATION OF HOW OTHER BREAST CANCER STUDIES DEALT WITH MISSING VALUES

INTRODUCTION

To investigate how data with missing values were analysed in other studies of survival from breast cancer, 14 papers were examined. These papers highlight known prognostic factors for breast cancer, and support the findings of the survival analysis of the Breast Cancer Audit data (Twelves et al, 1998a). Any references that are not discussed here, but which are mentioned in sections 2.3.2 and 5.2.3, do not directly involve any analysis of survival data.

Eleven of the 14 papers are retrospective studies, similar in nature to the Breast Cancer Audit. Two of the studies (Gordon et al, 1992; Haybittle et al, 1997) involve clinical trial data. These papers reported the effects of socioeconomic data on survival rather than the primary results of the trial. One study (Hawkins et al, 1996) was a prospective study for prognostic factors. Both this study and Gordon et al (1992), have no missing information, except for three cases in Gordon et al (1992) and are, therefore, not discussed further. The data in Haybittle et al (1997) do contain missing information in variables not used as part of the randomisation process. The approaches to dealing with missing information in the 12 studies is now detailed.

The first observation is that none of the papers included a discussion about the assumption of the structure of the missing data, and whether or not the data are assumed to be MAR. Several of the studies which used the Complete Cases (CC) method did compare the characteristics of the CC subgroup with all of the cases or with the excluded cases (Gillis & Hole, 1996; Newman et al 1997; Shek & Godolphin, 1988). Most of the studies concluded that the CC were representative of the whole population because the proportions in the levels of each of the factors were similar for all cases versus complete cases. Thus, the implicit assumption was made that the unknowns that would also be in the same proportions across the levels of the factors and, therefore, that the data are MAR. However, the important fact is whether or not the survival of the women in CC for a particular combination of factor levels is representative of the survival of all women in that particular combination, although this cannot be known.

Several of the papers explicitly state that they have used the CC analysis, by reporting that women with missing information were excluded from the analysis. The percentages of cases dropped in these studies (Haybittle et al, 1997; Ewertz et al, 1991; Carter et al, 1989; Shek & Godolphin, 1988; Newman et al, 1997) were 12%, 18%, 26%, 26% and 66% respectively. This last figure is almost identical to that in the Breast Cancer Audit data.

Two of the papers (Basnett et al, 1992; Gillis & Hole, 1996) do not explicitly say what they did with the missing values, although it appears that the cases were dropped and CC analyses performed. Basnett et al (1992) only had missing information for stage, which was unknown in only 9% and 4% of cases seen in teaching and non-teaching districts respectively. Gillis & Hole (1996) had a much higher percentage of missing information, with 31% and 22% missing for tumour size for non-specialist surgeons (non-spec) and specialist surgeons (spec) respectively. Similarly, 38% and 17% of women had no node status recorded for non-spec and spec respectively. Gillis & Hole (1996) quote crude survival figures for each of the factors separately, based on the cases where information was known in each of the factors, and also for the women with missing information. It is not clear, however, whether the adjusted hazard ratios obtained from the Cox regression model were based only on a CC analysis or whether additional categories were used for the unknowns.

The remaining five studies adopted the same method as used by Twelves et al (1998a); that is, the additional categories method. The percentages of cases with missing values that would otherwise have been lost, had a CC analysis been performed instead, were 9% (Karjalainen & Pukkala, 1990); 11% (Richards et al, 1996); 20% (Schrijvers et al, 1995); 47% (Sainsbury et al, 1995a); and 70% (Freedman et al, 1979).

Therefore, from this limited sample of papers chosen as suitable references for prognostic factors, it appears that none of the sophisticated techniques, such as Multiple Imputation, the Probability Imputation Technique or Maximum Likelihood methods are being used for the analysis of survival data with missing information. Instead, the only two techniques used were CC analysis and the additional categories method. Which of these methods, if either, is the more appropriate remains unclear.

CHAPTER 5 SURVIVAL ANALYSES

5.1 INTRODUCTION TO SURVIVAL DATA AND METHODS OF ANALYSIS

INTRODUCTION TO SURVIVAL DATA

Data which represent time from a definite origin to a particular event, or end-point, are known as survival data. Often, as with the Breast Cancer Audit data, the end-point of interest is death.

The survival times are assumed to be observations from a random variable T. Since time to an event is always positive, the distribution of the data is not symmetrical, but is generally positively-skewed. Therefore, the standard techniques for modelling normally distributed data cannot be used and so other techniques for modelling survival data have been developed.

THE SURVIVOR FUNCTION AND THE HAZARD FUNCTION

The survivor function is defined as the probability that an individual survives up to or beyond time t. Thus,

$$S(t) = P(T \geq t) = 1 - F(t),$$

where $F(t)$ is the cumulative distribution function. The probability density function of T is therefore given by

$$f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt}. \qquad \text{(Eq 5.1\_1)}$$

The instantaneous death rate of an individual surviving to time t is given by the probability than an individual dies at time t, given that they survived to that time. This is known as the hazard function and can be written as (Collett, 1994):

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}\{\log S(t)\} \text{ from Eq 5.1\_1.}$$

Thus, $S(t) = \exp\{H(t)\}$ or $H(t) = -\log S(t)$, (Eqs 5.1_2)

where $H(t) = \int_0^t h(u)\, du$.

$H(t)$ is the cumulative hazard. Note that $h(t)$ is not a probability density function.

CENSORING

Survival analysis needs to take into account subjects for whom the end-point does not occur before the end of the period of observation of the study. Rather than discarding data for such subjects, the information that an event did not occur is retained in the analysis by a method known as censoring. There are several types of censoring, but only that known as right-censoring will be explained here.

A subject is right-censored when it is known only that an event has not occurred by a certain point in time, t_c say, from the time origin, t_0 say. The observation is censored at t_c, and has a right-censored survival time of $(t_c - t_0)$. The analysis of censored data is only straightforward if it can be assumed that the true unknown survival time, $t > t_c$, is independent of the reason why the individual was censored at time t_c.

METHODS OF ANALYSIS

The analysis of observed survival times provides estimates of both the survivor and hazard functions. Parametric, non-parametric or semi-parametric methods can be used to obtain estimates of these functions.

A parametric model is fully described by a set of parameters for which probability distributions can be specified. One example used in survival analysis is the Weibull model. The hazard function for this distribution is given by

$$h(t) = \lambda \gamma t^{\gamma-1}$$

with $\lambda > 0$ and $\gamma > 0$. The corresponding survivor function is given by

$$S(t) = \exp\left(-\lambda t^{\gamma}\right). \quad \text{(Eq 5.1\_3)}$$

The scale parameter is λ and the shape parameter is γ. The simplest form of the Weibull distribution is the exponential distribution, which has shape parameter equal to 1. Parametric models were not fitted to the Breast Cancer Audit data, but the exponential distribution was modelled in a theoretical exercise based on simulated datasets, described in Chapter 6.

In contrast, a non-parametric model makes no assumptions about the distribution of T. An example of this approach, the Kaplan-Meier technique, is discussed in the next section. The third method is the semi-parametric approach. Part of the model is specified by parameters which can be obtained from modelling the data. The Cox's proportional hazards model is an example of a semi-parametric model and is described in Section 5.1.2.

5.1.1 KAPLAN-MEIER THEORY AND THE LOG-RANK TEST

INTRODUCTION

The Kaplan-Meier method is a non-parametric technique for estimating the survivor function, $S(t)$, at time t. Rather than model the survival data, it obtains the survivor function for intervals between consecutive end-points (i.e. deaths in the present context) from the ratio of the number of subjects still at risk (i.e. alive here) at the end of an interval to the number of subjects at risk at the start of that interval.

The Kaplan-Meier estimate of the survivor function is a step function. Collett (1994) provides a full derivation of this, a brief summary of which is outlined below.

BRIEF DERIVATION

The observed survival times for the n individuals in the sample are assumed to be $t_1,\ t_2,\ \ldots,\ t_n$. These may include censored observations and ties at the same time points. Thus, there are only r death times among the n individuals, with $r \leq n$. The jth ordered death time is denoted here by $t_{(j)}$.

The probability that an individual survives past $t_{(j)}$ given that they were at risk just before $t_{(j)}$ can be estimated by

$$\hat{p}_j = 1 - \frac{d_j}{n_j} = \frac{n_j - d_j}{n_j},$$

where d_j is the number of deaths that occurs at $t_{(j)}$ and n_j is the number of people still at risk just before $t_{(j)}$.

As no deaths are assumed to occur in the interval from $t_{(j)}$ to just before $t_{(j+1)}$, then $\hat{p}_j$ is equivalent to the probability of surviving from $t_{(j)}$ to $t_{(j+1)}$.

With the time intervals spanning from one death time to the next death time, the probability of surviving past $t_{(k)}$ is equivalent to surviving through all of the intervals before $t_{(k)}$ and surviving through the interval from $t_{(k)}$ to $t_{(k+1)}$, where $k = 1,\ 2,\ \ldots,\ r$. Therefore, the overall Kaplan-Meier estimate of $S(t)$ is given by

$$\hat{S}(t) = \prod_{j=1}^{k} \frac{n_j - d_j}{n_j},$$

for $t_{(k)} \leq t < t_{(k+1)}$, $k = 1,\ 2,\ \ldots,\ r$, $\hat{S}(t) = 1$ for $t < t_{(1)}$.

The standard error for the Kaplan-Meier estimate of the survivor function for any value of t in the interval from $t_{(k)}$ to $t_{(k+1)}$ is

$$se\{\hat{S}(t)\} = [\hat{S}(t)] \left\{ \sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)} \right\}^{\frac{1}{2}}, \quad \text{(Eq 5.1.1\_1)}$$

and is known as Greenwood's formula (Collett, 1994).

POINT ESTIMATES WITH CONFIDENCE INTERVALS

An approximate confidence interval for the estimate $\hat{S}(t)$ can be obtained using its calculated standard error. This is derived under the assumption that $\hat{S}(t)$ comes from a normal distribution with mean $S(t)$ and standard deviation given by Eq 5.1.1_1. Survival curves can be generated for subgroups of individuals in each of the levels of a factor; such as, different age groups. The results of the Kaplan-Meier analyses of the Breast Cancer Audit data are given in Section 5.2.1.

TESTING FOR EQUALITY OF THE SURVIVAL CURVES

Several tests of equality of the survival curves can be carried out. The log-rank and Wilcoxon tests are discussed here for a factor with only two levels, but both can be used for factors with more levels.

The null hypothesis for both of these tests for two levels of a factor is

$$H_0: S_1(t) = S_2(t) \quad \text{for all } t > 0.$$

The log-rank test is powerful when the assumption of proportional hazards is valid (Gregory et al, 1997). If this seems questionable, it may be better to use the Wilcoxon test (Collett, 1994). To decide which of these tests to perform, it is good practice to examine whether or not the hazards for the levels are proportional. The survivor functions for the different levels do not cross when the hazard functions are proportional (Collett, 1994). Thus, examination of the estimated survival curves for the two (or more) levels gives an informal indication as to whether the proportional hazards assumption holds or not for the set of data being considered.

Having identified factors where there is evidence to reject the test of equality among the survival curves for the levels of the factors, multivariate survival analysis is then often employed to investigate whether or not the differences remain once other factors have been taken in account.

5.1.2 THEORY OF COX'S PROPORTIONAL HAZARDS REGRESSION MODELS

INTRODUCTION

The aim of modelling the survival data is to describe the dependence of the outcome on one or more of the covariates. Models with proportional hazards are often used in survival analyses. The assumption of proportionality implies that the ratio of the hazards between different levels of a factor, or different values of a continuous variable, are constant over time. The Cox model is one example of a proportional hazards model.

THE MODEL

A Cox's proportional hazards regression model is usually given in the general form

$$h(t;\underline{x}) = \exp\left(\underline{\beta}^T \underline{x}\right) h_0(t).$$

$h_0(t)$ is known as the baseline hazard and is assumed to be unknown and is not itself necessarily of interest. It represents the hazard function for an individual with all of the explanatory covariates taking the value zero. As no form is assumed for $h_0(t)$, this part of the model is non-parametric. A parametric component of $h(t;\underline{x})$ arises through $\exp\left(\underline{\beta}^T \underline{x}\right)$ and hence the model is referred to as semi-parametric.

BRIEF DERIVATION OF THE REGRESSION PARAMETERS

Cox (1972) showed how the regression parameters could be estimated without the necessity of calculating $h_0(t)$ using a method called partial likelihood. Collett (1994) provides details of this, a brief summary of which is given here.

Again, it is assumed that there are n individuals with survival times $t_1, t_2, \ldots, t_n$. There are $r \leq n$ death times, with the jth ordered death time given by $t_{(j)}$. Let the individuals at risk, i.e. those alive and uncensored, just before $t_{(j)}$ be denoted by $R\left(t_{(j)}\right)$ and be called the risk set at that time. Then, the probability that the ith individual dies at time $t_{(j)}$, conditional on $t_{(j)}$ being a time of death, is equal to

$$\frac{\exp\left(\underline{\beta}^T \underline{x}_{(j)}\right)}{\sum_{l \in R\left(t_{(j)}\right)} \exp\left(\underline{\beta}^T \underline{x}_l\right)},$$

where $\underline{x}_{(j)}$ represents the covariates for the individual who dies at time $t_{(j)}$ and $\underline{x}_l$ are the explanatory variables for individual l. Using independence, the partial likelihood function is given by

$$L(\beta) = \prod_{j=1}^{r} \frac{\exp\left(\underline{\beta}^T \underline{x}_{(j)}\right)}{\sum_{l \in R\left(t_{(j)}\right)} \exp\left(\underline{\beta}^T \underline{x}_l\right)}.$$

The method of maximum likelihood can be used to find the estimates for $\underline{\beta}$ by maximising the partial likelihood function. Thus, the estimates for the parameters for the covariates have been derived without knowing anything about the baseline hazard function, $h_0(t)$.

THE SURVIVOR FUNCTION

Having estimated the regression parameters, the baseline hazard function and corresponding survivor function (Kalbfleisch & Prentice, 1980) can then be obtained to provide an estimate of the survival curve for all times. The estimated survivor function is given by

$$\hat{S}_i(t) = \left[\hat{S}_0(t)\right]^{\exp\left(\hat{\underline{\beta}}^T \underline{x}_i\right)} \qquad \text{(Eq 5.1.2\_1)}$$

with covariate pattern $\underline{x}_i$ for the ith individual and $\hat{S}_0(t)$ is the estimated baseline survivor function, obtained from Eqs 5.1_2 in Section 5.1.

The formula for the standard error for $\hat{S}_i(t)$ can be obtained from Kalbfleisch & Prentice (1980). These estimates for survival and their standard errors are available in statistical packages such as SPSS and SAS.

MODEL SELECTION

In arriving at a final Cox model, variable selection methods used in multiple regression are commonly employed. Both forward selection and backward elimination can be performed, along with other stepwise techniques (Armitage & Berry, 1994). The forward selection stepwise method was used to obtain the model in the analysis of the Breast Cancer Audit data and this was checked using backward elimination.

The results of fitting a Cox model to the Breast Cancer Audit data are presented in the paper by Twelves et al (1998a) and are extended in Section 5.2.2.

5.2 SURVIVAL ANALYSIS OF THE BREAST CANCER AUDIT DATA

INTRODUCTION

A survival analysis was performed on data from the Breast Cancer Audit. Only the subgroup of women who had no evidence of metastases at presentation and who underwent surgery were included in the analysis (Section 3.1). The aim was to investigate any variations in survival after surgery for women diagnosed with breast cancer in Scotland in 1987.

CENSORING AND END-POINT IN THE BREAST CANCER AUDIT DATA

The time origin was taken to be the time of diagnosis and the end-point was death from any cause. Cause-specific survival was not analysed because cause of death information from death certificates has been shown to be unreliable (Maudsley & Williams, 1993). Deaths up to the end of 1993 were linked to the original audit data using probability matching (Kendrick & Clarke, 1993) with death data from the General Registers Office. Thus, subjects without a recorded death from the matching were assumed to be alive at 31/12/1993 and were censored at this point.

THE DATA

Table 5.1 presents the variables chosen for inclusion in the survival analysis. The levels for these factors have already been given in Section 3.2.

CLINICAL FACTORS
Age
Clinical stage
Pathological node status
Pathological tumour size
Oestrogen receptor (ER) status
SERVICE FACTORS
Health Board of first treatment
Deprivation
Surgical case load
Seen by an oncologist
TREATMENT FACTORS
Type of surgery
Adjuvant radiotherapy
Adjuvant chemotherapy
Adjuvant endocrine therapy
Adjuvant chemotherapy or endocrine therapy

Table 5.1: List of variables in the three categories.

ANALYSIS STRATEGY

Initially univariate log-rank tests were performed on all of the factors listed above to identify which were important. Kaplan-Meier survival estimates at five years were also obtained for the levels of each factor.

Subsequently, two Cox regression models were fitted. The first model included the significant clinical factors plus any significant service factors. The second model included the significant clinical factors, but this time incorporated any significant treatment factors. The treatment and service factors were not included in a single model because the study was retrospective and so the nature of treatment may have been determined, in part, by service factors and would therefore be confounded. For example, women were unlikely to receive chemotherapy if they did not see an oncologist.

USING SPSS FOR SURVIVAL ANALYSES

SPSS was used to perform both the Kaplan-Meier and the Cox survival analyses. An estimated 'average' survivor function for any of the death times can be obtained from the Cox model by using the means as the values of the covariates. This is straightforward for continuous variables. However, for categorical variables, the 'average' is obtained by using the relative frequencies of the numbers of cases in each of the levels as weights. These are then multiplied by the corresponding parameter estimates for each level and an 'average' survival obtained from this.

For example, for factor A with three levels, the 'average' risk, $\exp\left(\hat{\underline{\beta}}^T \underline{x}\right)$ from Eq 5.1.2_1 in Section 5.1.2, would be given by

$$\exp\left(\hat{\underline{\beta}}^T \underline{x}\right) = \exp\left(\frac{n_{a1}}{n_a}\hat{\beta}_1 + \frac{n_{a2}}{n_a}\hat{\beta}_2 + \frac{n_{a3}}{n_a}\hat{\beta}_3\right), \quad \text{(Eq 5.2\_1)}$$

where n_{a1}, n_{a2}, n_{a3} and n_a represent the numbers of cases in levels 1, 2 and 3 of factor A and the total number of cases in factor A respectively. $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$ represent the parameter estimates for levels 1, 2 and 3 of factor A, with $\hat{\beta}_1 = 0$ in the particular set-up of indicator variables. The 'average' survival estimate would then be obtained using Eq 5.1.2_1 from Section 5.1.2. This would give an indication of how the hazards calculated from the parameter estimates affect the percentage of patients surviving to a particular time point, in general.

It is also possible to derive model based survival estimates for particular risk factor profiles; that is, obtaining the risk for certain levels of the factors. For example, age group 50-64, clinical stage II, ER positive, node negative, tumour size ≤ 2cm. The survival estimate at five years for individuals with these characteristics could then be reported.

5.2.1 RESULTS OF UNIVARIATE ANALYSES

Separate Kaplan-Meier analyses were performed on a subgroup of the factors in the Breast Cancer Audit data. A full table of results showing the 5-year % survival figures can be found in Table 1 of the paper by Twelves et al (1998a). The overall Kaplan-Meier estimate of 5-year survival was 70.9% with a 95% confidence interval (68.6%, 73.1%). The P values for log-rank tests for equality of the survival curves are given below in Table 5.2.

Factor	**P value**
CLINICAL FACTORS	
Age	<0.0001
Clinical stage	<0.0001
Pathological node status	<0.0001
Pathological tumour size	<0.0001
ER status	<0.0001
SERVICE FACTORS	
HB of first treatment	0.02
Deprivation	0.03
Surgical case load	0.03
Seen by an oncologist	0.25
TREATMENT FACTORS	
Type of surgery	0.01
Adjuvant radiotherapy	0.49
Adjuvant chemotherapy	0.02
Adjuvant endocrine therapy	0.74
Adjuvant chemotherapy or endocrine therapy	0.28

Table 5.2: P values for the overall log-rank tests of equality of the survival curves in univariate analyses.

As can be seen, there were significant differences between levels for each of the clinical factors. Figures 5.1 and 5.2 give the Kaplan-Meier estimated survival curves for the factors pathological node status and pathological tumour size to illustrate the differences. The remaining Kaplan-Meier curves for the clinical variables are given in Section 5.3.2. There was also evidence of variation in survival among the levels of all of the service factors, except referral to an oncologist, and of the treatment factors type of surgery and use of adjuvant chemotherapy.

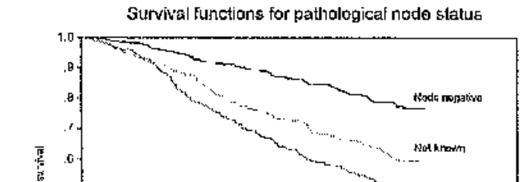


Figure 5.1: Kaplan-Meier survival curves for pathological node status.

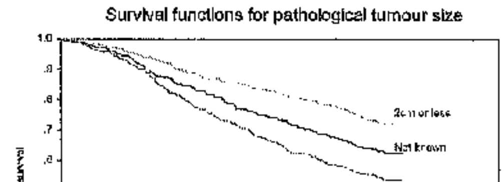


Figure 5.2: Kaplan-Meier survival curves for pathological tumour size.

Having identified the statistically significant factors in the univariate analyses, multivariate Cox's proportional hazards regression modelling was then used to investigate whether these factors remained statistically significant once other factors had been allowed for. The results of this analysis are described in the next section.

5.2.2 COX'S PROPORTIONAL HAZARDS ANALYSIS: THE 'CLINICAL FULL' MODEL

INTRODUCTION AND METHODS

The results of fitting two Cox's proportional hazards (PH) regression models are reported. The first Cox model involved testing all of the clinical and service factors

(Model 1), whereas the second model had all of the clinical and treatment factors available for selection (Model 2).

The forward selection stepwise technique (Armitage & Berry, 1994) was used to identify the 'best' model on the basis of the variables offered to it. Unfortunately, it was not possible to enter all of the service factors together in Model 1 without losing some cases; i.e. those with either surgical case load or referral to an oncologist unknown. These factors were therefore included separately into a model with the clinical factors and Health Board.

Variables were added to the model if they were significant at the 5% level and were removed if, with the addition of other variables, they became non-significant at the 10% level. The P values for the Wald statistics for the estimates in the model, conditional on the other variables being present, are given for the significant variables, whereas the score statistics for entering are given for those that were non-significant.

The P values given here for the non-significant factors are slightly different to those quoted by Twelves et al (1998a). In that paper, the P values shown for the Wald statistics were for *forced* entry, one at a time, for the non-significant factors with *only* the significant clinical factors included in the model. The reason for the difference is that a referee requested that the adjusted survival estimates be given for the non-significant factors to allow comparison with the Kaplan-Meier estimates. Health Board was not included in this procedure of forcing in variables as some of the service factors were correlated with this factor. However, it was included in the analysis reported below for Model 1, where the service factors were available for entry with the clinical factors and Health Board.

Hauck & Miike (1991) suggest a method for presentation of results when using the stepwise selection technique. Their technique illustrates the order of entry of the variables by highlighting which variable enters at each step. The P values shown in their suggested table format are for entry for those variables not included in the model and for removal for these from the model. They also identify variables whose significance changes greatly between steps $n-1$ and n as a result of another variable entering at step n indicating a high degree of correlation between the variables. Thus, these variables probably will not both enter the model, even though both may affect the dependent variable independently. Gordon et al (1992) make use of this mode of presentation in their paper with minor modifications. Elements of this and the original approach are used in this thesis.

Only point estimates for survival are given in this thesis for Cox regression models because an apparent problem was discovered during this research with the estimated standard error obtained for the survival estimate from Cox regression using Version 9.0 of the SPSS statistics package. Some examples of this problem are given in Appendix 5, along with the results of fitting a simpler model using binary variables only. This was used to compare the standard errors obtained from SPSS and SAS. The estimates are different (substantially in some situations) from the two packages. All of the other estimates and standard errors relating to survival analyses appear to be correct in SPSS, and it is only in the standard error on the survival estimate that there are differences between the two statistics packages.

Unfortunately, there was not enough time to repeat all of the Cox regression analyses given in this thesis using SAS due to the difficulties in dealing with categorical covariates in SAS. On the basis of the findings discussed in Appendix 5, it was decided not to include standard errors until the apparent discrepancies had been sorted out. However, this issue is currently unresolved and remains a subject under discussion with SPSS Inc.

RESULTS FOR MODEL 1

All of the clinical factors were required in the model as expected, along with the service factor Health Board (HB) of treatment. The P values given in Table 5.3 are for Wald statistics for presence in the model for that factor with all of the other significant variables in the model.

The other service factors: deprivation, surgical case load and referral to an oncologist were not significant in models with the factors given in Table 5.3, and were therefore not included in Model 1. The P values for the variables that were not entered into the model were 0.36 for deprivation, 0.88 for surgical case load and 0.34 for referral to an oncologist.

Significant Factors	P Value
Age Group	0.0017
Clinical Stage	0.0008
ER Status	<0.0001
Node Status	<0.0001
Tumour Size	<0.0001
Node Status by Tumour Size	0.0059
HB of Treatment	0.0160

Table 5.3: P values for Wald statistics for the significant factors in Model 1.

To present the hazard ratios and the 5-year % survival estimates for the interaction between node status and tumour size, a factor consisting of the nine possible combinations of the two factors was created. The numbers and percentages for the levels of this new variable are given in Table 5.4.

Combination	Number	%
Node not known, Tumour size ≤ 2cm	185	11.4
Node not known, Tumour size > 2cm	138	8.5
Node not known, Tumour size not known	112	6.9
Node positive, Tumour size ≤ 2cm	171	10.6
Node positive, Tumour size > 2cm	312	19.3
Node positive, Tumour size not known	100	6.2
Node negative, Tumour size ≤ 2cm	269	16.6
Node negative, Tumour size > 2cm	212	13.1
Node negative, Tumour size not known	120	7.4
Total	1619	100.0

Table 5.4: Numbers and percentages in each of the combinations of the interaction between node status and tumour size.

Table 5.5 shows the order of entry of the variables along, with the P values for removal for those variables already in the model and for entry for the non-significant variables.

As can be seen from this table, all of the clinical factors were highly significant and the order of entry of these factors is not important. The significance of deprivation altered upon the addition of ER status into the model at Step 2.

	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Step 7
A	0.0001	<0.0001	0.0001	e 0.0002	a 0.0012	a 0.0014	a 0.0035
C	<0.0001	<0.0001	0.0005	0.0026	e 0.0022	a 0.0009	a 0.0007
E	<0.0001	e <0.0001	a <0.0001	a <0.0001	a <0.0001	a <0.0001	a <0.0001
N	e <0.0001	a <0.0001	a <0.0001	a <0.0001	a <0.0001	a <0.0001	a <0.0001
T	<0.0001	<0.0001	e <0.0001	a <0.0001	a 0.0004	a <0.0001	a <0.0001
N*T	0.0093	0.0177	0.0097	0.0098	0.0040	e 0.0045	a 0.0059
H	0.0154	0.0251	0.0051	0.0134	0.0113	0.0141	e 0.0157
D	0.0410	0.1278	0.1860	0.1819	0.2485	0.2808	0.3628

*Table 5.5: Presentation of results from performing the stepwise selection on the variables that were initially offered to Model 1. Note that 'e' indicates which variable entered the model at that stage and that 'a' indicates that the variable has already been entered in the model. The P values represent entry for the ones not already in the model and removal for those factors in the model. Age group is given by A, clinical stage (C), ER status (E), node status (N), tumour size (T), their interaction (N*T), Health Board (H) and deprivation (D).*

The overall adjusted 5-year survival estimate was 74.3%. This represents the survival at five years for an 'average' subject. It uses the weighted risks for each of the levels in each of the factors. Thus, the 'average' subject does not represent any individual subgroup. Table 5.6 below gives the hazard ratios with 95% confidence intervals (CI) along with the adjusted 5-year % survival for the significant factors in Model 1. No CIs are given for the survival estimates due to the apparent problem with the estimates of their standard errors from SPSS (Appendix 5).

The adjusted survival estimates were obtained by specifying in SPSS that the survival estimates for each of the levels of the separate factors be given, whilst averaging over the other factors. These figures are the estimates for all women in one level of one particular factor, conditional on having weighted risks for other factors. The weighted risks of the other factors are assumed to be the same for all levels of that particular factor. For example, the estimate of adjusted survival at five years for age group 50-64 represents the survival for women in this age group, assumed to have 'average' characteristics for all of the other factors. The estimate does not represent the true risk for the women aged 50-64 in the cohort, all of whom will have meaningful levels of the other prognostic factors.

The purpose of reporting these figures is because the adjusted survival estimates simply re-express the hazard ratios on a scale that is easier to interpret; namely, percentage surviving at five years.

Variable	Hazard Ratio (95% CI)	Adjusted 5-yr % Survival
Age		
< 50 years	1	76.4
50 - 64	1.04 (0.84, 1.29)	75.5
65 - 79	1.18 (0.95, 1.47)	72.8
≥ 80 years	2.01 (1.39, 2.90)	58.2
Clinical Stage		
Stage I	1	80.9
II	1.41 (1.07, 1.85)	74.2
III	1.98 (1.42, 2.78)	65.6
Not known	1.54 (1.13, 2.09)	72.2
ER Status		
Positive	1	80.7
Negative	2.11 (1.69, 2.63)	63.6
Not known	1.45 (1.15, 1.82)	73.3
Node Status by Tumour Size		
N nk, T ≤ 2cm	2.28 (1.50, 3.48)	77.1
N nk, T > 2	3.53 (2.32, 5.38)	66.9
N nk, T nk	3.00 (1.93, 4.67)	71.1
N +ve, T ≤ 2cm	3.91 (2.62, 5.84)	64.1
N +ve, T > 2	4.37 (3.01, 6.35)	60.8
N +ve, T nk	4.46 (2.89, 6.88)	60.2
N -ve, T ≤ 2cm	1	89.2
N -ve, T > 2	2.72 (1.82, 4.07)	73.4
N -ve, T nk	1.45 (0.86, 2.44)	84.8
Health Board		
A	1.52 (1.10, 2.10)	66.9
B	1.46 (0.72, 2.93)	68.1
C	1.49 (1.06, 2.10)	67.5
F	1.55 (1.05, 2.29)	66.3
G	1	76.8
H	0.97 (0.61, 1.54)	77.4
I	0.64 (0.31, 1.34)	84.5
L	1.20 (0.86, 1.66)	72.9
N	0.95 (0.69, 1.31)	77.8
S	0.88 (0.65, 1.19)	79.3
T	1.33 (0.94, 1.87)	70.4
V	1.41 (0.90, 2.20)	68.9
Y	1.11 (0.71, 1.76)	74.5

Table 5.6: Hazard ratios and adjusted 5-yr % survival estimates, with 95% CIs for the hazard ratios. Note that N and T stand for node status and tumour size respectively. Also, nk stands for not known, +ve for positive and -ve for negative.

The significance of the node status by tumour size interaction highlights the importance of both of these factors in terms of the survival prospects of women with breast cancer. Women who were node negative with small tumours had a 89% adjusted 5-year survival estimate whereas women with large tumours and node negative disease only had a 73% chance of survival at five years. This indicates that if a woman had a favourable node status, she still had reasonable (though not as good) chances of survival when her tumour was large. However, women with node positive disease with small tumours only had a marginal advantage over women with large tumours and node positive disease (64% compared with 61% respectively). Thus, the effect of tumour size was less for women who were node positive compared with those with node negative disease.

The finding that Health Board of treatment was significant in the model is important. The magnitude of the apparent differences among the Health Boards were clinically significant when compared with the magnitude differences in the survival estimates between use and no use of treatments such as tamoxifen and chemotherapy drugs seen in clinical trials (Section 2.3.2).

UNIVARIATE VERSUS MULTIVARIATE RESULTS FOR MODEL 1

It does not make sense to directly compare the survival estimates obtained from the Cox model with those from Kaplan-Meier analysis because of the difference in the interpretations of the survival estimates. A Kaplan-Meier figure represents the actual survival for particular group of women with a particular level of a factor. The Cox regression estimate represents the risk for a particular level of a factor 'averaged' over the other factors.

As expected, all of the clinical factors were significant in both the univariate and multivariate analyses. The factors deprivation and surgical case load were significant in the univariate log-rank tests but not in the multivariate Cox's PH model, once the clinical factors had been adjusted for. It appeared that least deprived women had a better prognosis from the univariate analysis, but deprivation was not significant in the Cox model with clinical factors in it. Table 5.5 suggested that deprivation and ER status were correlated because the significance of deprivation changed from 0.04 to 0.13 when

ER status was added to the model. The test statistic for the χ^2 test for association between deprivation and ER status was significant with P value <0.001. Table 5.7 displays this association and reveals that there were more women who were ER negative and resident in an area of greatest deprivation then would be expected had these variables been independent.

The association remained when the women with ER status unknown were excluded from the analysis (P=0.004); that is, there was still an excess of women with ER negative disease resident in areas of high deprivation. Thus, the poorer observed 5-year survival for being most deprived may be due in part to this greater proportion of women who were ER negative. This is discussed further in Section 5.2.3.

		Least Deprived	**Intermediate**	**Most Deprived**
ER Positive	Obs	165	354	80
	Exp	149.5	363.3	86.2
	Res	1.8	-1.0	-0.9
ER Negative	Obs	90	218	83
	Exp	97.6	237.2	56.3
	Res	-1.0	-2.3	4.4
ER Not known	Obs	149	410	70
	Exp	157.0	381.5	90.5
	Res	-0.9	3.0	-3.0

Table 5.7: Observed (Obs) and Expected (Exp) numbers of cases under the assumption of no association between the variables ER status and deprivation. Note: Res stands for the adjusted standardised residual for the cell and can be tested as a Normal (0,1) deviate. Thus, cells with magnitude in excess of about +2 or -2 can be regarded as being significant.

Referral to an oncologist was not significant in either the univariate or multivariate analyses. This is not a surprising result as it is likely that whether or not a women saw an oncologist, as well as a surgeon, would depend on the nature of her disease. Some women will have had a good prognosis and will have had breast conservation followed by radiotherapy (thus seeing an oncologist). Others will have had a poor prognosis being node positive, but would be seen by an oncologist for the administration of chemotherapy.

RESULTS FOR MODEL 2

The second Cox model fitted was restricted to clinical and treatment factors. It was found that none of the treatment factors were significant. The P values for the score statistics for non-entry of these factors were 0.90, 0.60, 0.98, 0.42 and 0.25 for type of surgery, use of adjuvant radiotherapy (RT), use of chemotherapy (CT), use of adjuvant endocrine therapy and use any adjuvant systemic therapy respectively. The fact that all of these treatment factors were non-significant is not completely surprising with this being a retrospective study and not a randomised trial. The treatment would have been determined by the severity of the disease and thus would probably be strongly confounded by the presence of the clinical factors in the model.

Significant differences were observed between the survival curves in the Kaplan-Meier analyses for both use of CT and type of surgery (Figures 5.3 and 5.4).

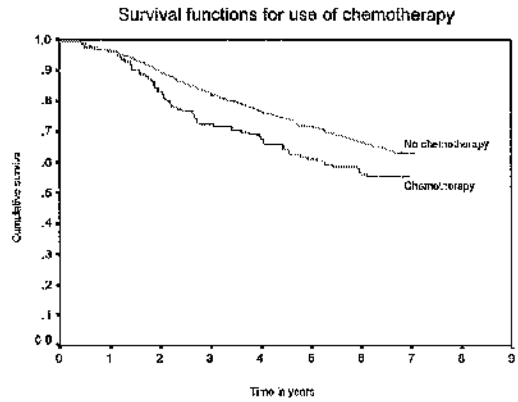


Figure 5.3: Kaplan-Meier survival curves for use of chemotherapy.

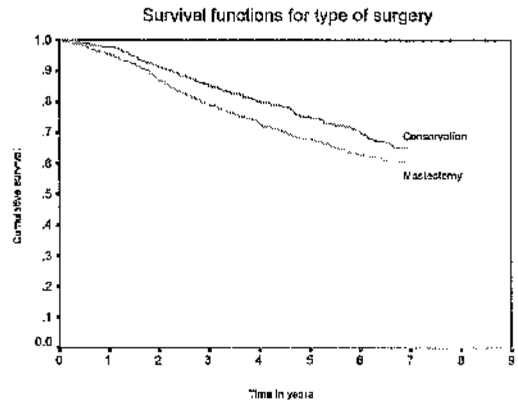


Figure 5.4: Kaplan-Meier survival curves for type of surgery.

Thus, the poorer Kaplan-Meier survival chances are for women who had a mastectomy or received CT. These treatments tend to be given to women with poorer prognosis tumours, i.e. women with a large tumour size or with node positive disease. Therefore, the women who needed to receive these treatments will have been expected to have the worse observed survival chances. This is accounted for in the full multivariate model, in which the treatment effects are no longer significant, once the clinical factors have been modelled.

THE 'CLINICAL FULL' MODEL

Model I was taken to be the 'best' model for these Breast Cancer Audit data and was the one used to obtain results presented in Twelves et al (1998a). The same model was obtained when the method of backward elimination was used (Armitage & Berry, 1994), when only the main effects and two-way interactions among the clinical factors were offered to the model. The only interaction that approached significance at the 5% level was that of clinical stage by node status, with P value 0.062.

Thus, the 'Clinical Full' model includes the factors: age, clinical stage, ER status, node status, tumour size, their two-way interaction and HB of treatment. Further discussion about the interpretation of these findings and comparison with the results of other studies are given in the next section and investigation into the use of the additional categories method for dealing with the missing values in the covariates is presented in Section 5.4.2.

5.2.3 DISCUSSION OF THE AUDIT RESULTS AND COMPARISON WITH OTHER RELEVANT STUDIES

INTRODUCTION

The findings of the analysis of the Breast Cancer Audit data are now compared with the published literature and discussed in the different groups: clinical factors, service factors, deprivation, treatment factors and patient characteristics.

CLINICAL FACTORS

All of the expected clinical factors were significant in the 'Clinical Full' model. It was found that there was a monotonically increasing hazard ratio of death from any cause with increasing age group. The 33 women aged under 35 years at diagnosis were combined with the 35-49 age group prior to the survival analysis, giving <50 years as the reference category. Therefore, it was not possible to examine whether the under 35 year olds had a worse prognosis, as suggested by Miller et al (1994) and Richards et al (1996).

Clinical stage was also a significant factor in the results from the Audit data, with an increasing risk of death with increasing stage. Similarly, ER status was significant in the model, with ER negative tumours having a worse outcome than ER positive tumours. Both node status and tumour size were significant in the Breast Cancer Audit data analysis, along with the interaction between them. This is similar to the finding of Carter et al (1989). The general effects of all of these factors on survival from breast cancer were discussed in Section 2.3.2.

SERVICE FACTORS

Basnett et al (1992) showed that survival was better for women treated in a hospital in a teaching district (T) than for women treated in a non-teaching (NT) district. The odds ratio of death for NT vs T, adjusted for age and clinical stage, was 1.74 (95% CI: 1.34, 2.27). They suggested that variations in use of different forms of adjuvant treatments in the NT and T districts may be the reason for the different survival figures observed. Similarly, Sainsbury et al (1995b) found variations in survival chances probably due to differences in use of adjuvant treatment in different regions in Yorkshire. Although not identical, use of Health Board of treatment in the analysis of the Audit data showed that none of the three Health Boards which had significantly higher hazard ratios (A, C and F) than Greater Glasgow Health Board (G) contained a Cancer Centre or a teaching hospital and appeared to have lower odds of use of adjuvant treatment (see Table 3 in Twelves et al, 1998a). However, use of any adjuvant systemic therapy was not statistically significant in the survival model, possibly because the Audit was not a randomised trial, with the clinical factors probably determining the treatment given. A fuller discussion is given in Twelves et al (1998a).

Instead of examining the facilities of the hospital delivering the care, Sainsbury et al (1995a) investigated the effect of the case load of the surgeon on survival. They found that surgeons treating more than 30 women a year had a risk ratio of death of 0.85 (95% CI: 0.77, 0.93) when compared to surgeons seeing less than 10 women with breast cancer per year. They also examined the rates of the usage of chemotherapy and endocrine therapy and found that high case load surgeons were more likely to prescribe adjuvant treatment than low case load surgeons, suggesting a possible reason for the observed improved survival rates for the high case load surgeons. The administration of chemotherapy suggests involvement of an oncologist implying, perhaps, that it was the multidisciplinary approach to care which improved survival.

Rather than using case load as a measure of expertise of the surgeon, Gillis & Hole (1996) coded each surgeon responsible for women with breast cancer in the West of Scotland in 1980-88 as being specialist surgeons or not. The surgeons were coded by "local perception". Gillis & Hole (1996) point out that each of the specialist surgeons "demonstrated the following indicators of specialist interest … setting up a dedicated breast clinic; a defined association with pathologists and oncologists; organising and facilitating clinical trials; and maintaining a separate record of all patients with breast cancer in their care." The hazard ratio for specialist (spec) vs non-specialist (non-spec) was 0.84 (95% CI: 0.75, 0.94), after adjustment for age, tumour size, deprivation and any nodal involvement.

In the analysis of the Breast Cancer Audit data, the variable described by Sainsbury et al (1995a) for surgeon case load was used, except that it was modified to include surgeons identified prior to analysis as working in breast clinic teams in the groups treating 30 or more women per year. This factor was found to be significant when a log-rank test was performed on the factor univariately (Section 5.2.1), but was not significant in the multivariate Cox model (Section 5.2.2). This makes sense because Section 4.1 showed that surgeon case load was significantly associated with each of the factors: age, clinical stage, ER status and node status in χ^2 tests of association for the pairs of variables. Initial examination of the breakdown of these pairs of variables in Appendix 4 shows that the surgeons with a higher case load had a better case-mix and saw patients with already improved prognoses.

However, these surgeons provided better staging of the disease for the women treated in their care as there was a lower proportion than expected in the unknown categories in the clinical factors for the high case load surgeons. When the unknowns in the clinical factors were removed from the analysis, none of the pairs of associations for the clinical factors with surgeon case load remained significant.

Thus, only the association between surgeon case load and age group was significant when unknowns were excluded. Therefore, the observed better survival chances for women seen by the high case load surgeons were partly due to these surgeons seeing a younger group of women (mainly <65 years) and partly because they staged the disease more extensively, thus providing the opportunity for the most appropriate treatment to be given.

Twelves et al (1998a) make the point that taking the results from Gillis & Hole (1996) and Sainsbury et al (1995a) with the findings from the Audit leads to the conclusion that the 'surgeon effect' probably translates into an effect of improved overall care, with treatment administered in a multidisciplinary team.

DEPRIVATION OR SOCIAL CLASS

Gillis & Hole (1996) showed that there were differences in the crude survival figures by deprivation, although the absolute figures varied by whether or not the women were treated by a specialist surgeon or not. Affluent women had 72% and 64% survival at five years for spec and non-spec respectively, compared to deprived women having 5-year survival figures of 65% and 54% for spec and non-spec respectively.

Sainsbury et al (1995a) also reported a higher hazard ratio for the most deprived category vs the rest (1.16; 95% CI: 1.10, 1.22), having adjusted for other factors in a Cox model. Schrijvers et al (1995) detailed a similar relationship for relative survival rates. Gordon et al (1992) used different area-based measures of socioeconomic status, such as percentage with higher education, mean family income, percentage in poverty. They also reported a higher risk of death with lower socioeconomic status. Carnon et al (1994) discussed a gradient in survival by deprivation category in approximately 7,500 women in the West of Scotland. They examined the association between deprivation category and the prognostic factors: tumour size, percentage of nodes positive, grade and ER status in about 1,300 women and found none of them to be significant. However, they did not examine survival in this subgroup of women where pathological information was available.

Rather than use the area-based measures of deprivation category, Karjalainen & Pukkala (1990) and Haybittle et al (1997) used social class, based on occupation, as a measure of material affluence (OPCS, 1975). Karjalainen & Pukkala (1990) found that the risk of death for being in a high social class (low deprivation) was 0.78 (95% CI: 0.68, 0.90) times that of being in the lowest social class. However, Haybittle et al (1997) did not find significant differences between the social classes with manual (IIIm, IV, V) vs non-manual (I, II, IIIn) having a relative risk of 1.07 (95% CI: 0.97, 1.19). The P value for the log-rank test between these groups was 0.12.

In the analysis of the Breast Cancer Audit data, deprivation was significant in the univariate analysis, but not in the multivariate Cox model. This suggests that the observed survival differences for the deprivation categories could partly be explained to the different proportions of ER status in the deprivation categories, with a larger number of ER negative women in the most deprived group (see last section). ER negative tumours have been shown to have a worse prognosis than ER positive tumours, both in the Audit and by other studies (Newman et al, 1997; Shek & Godolphin, 1988; Hawkins et al, 1996), thus perhaps explaining why women living in deprived areas had a poorer observed survival.

One possible explanation why women living in deprived areas have more ER negative tumours could be because a larger proportion of women resident in deprived areas also have low or average body mass index (BMI); that is not in the obese category. This hypothesis follows from Giuffrida et al (1992), who showed an excess of ER negative tumours in women with low or average BMI. However, the weight and deprivation relationship in the Breast Cancer Audit could not be examined as weight details were not collected.

Overall, although there is some evidence to support the observation that deprivation affects survival, it has still not been proved definitely.

TREATMENT FACTORS

Univariately, having a mastectomy or receiving chemotherapy indicated significantly worse survival in the analysis of the Breast Cancer Audit data. These factors were not significant in the multivariate model, however. This was as expected with the Audit being not being a randomised trial but a retrospective study of how women had their breast cancer managed in 1987. Thus, the treatment would probably have been driven by the prognostic clinical factors, with most of the women receiving chemotherapy having poor prognosis node positive disease, for example. The influence of treatment factors on survival have already been discussed in Section 2.3.2.

PATIENT CHARACTERISTICS

Haybittle et al (1997) found that survival was affected by the weight of women who were postmenopausal, with a highly significant risk ratio for women >60 kg vs ≤60 kg being 1.20 (95%CI : 1.08, 1.33). No relationship was observed for pre- or perimenopausal women of weight on survival. However, Ewertz et al (1991) found a different pattern with relative risks of 1.48 (1.03, 2.12) for <50 kg; 0.88 for 60-69; 0.99 for 70-79; and 1.02 (0.90, 1.55) for ≥80 kg with 50-59 taken as the reference category. Gordon et al (1992) found no relationship of survival with body mass index (BMI), which has weight as one of its constituents. However, Newman et al (1997) did find that BMI was related to survival, but only for women who were node negative. The hazard ratios relative to women with BMI <22.8 with no nodal involvement were 2.1 (1.1, 4.2) and 2.5 (1.2, 5.2) for no nodal involvement and BMI 22.8-28.9 and BMI >28.9 respectively. Neither weight nor BMI were available for analysis in the Audit. The relationship of these factors with survival remains unclear.

Overall, it would appear that the results from the analysis of the Breast Cancer Audit data discussed here and in Twelves et al (1998a) are similar to the findings of others and support the need for a multidisciplinary approach to the care of women with breast cancer.

5.2.4 EXPLORATION OF OTHER TWO-WAY INTERACTIONS

INTRODUCTION

When the original survival analysis was performed (Twelves et al, 1998a), the available computing facilities were not sufficiently powerful to allow for interactions with the factor for Health Board to be examined (because this factor has 13 levels). Each of the two-way interactions between the pairs of the clinical factors was examined for significance. None achieved the 5% significance level, except for the interaction between node status and tumour size. This factor makes clinical sense as discussed in Section 5.2.2.

However, the computing facilities are now available to allow the interactions between Health Board and the clinical factors to be examined. The results are presented below.

RESULTS

The interaction between node status and tumour size is already part of the 'Clinical Full' model. Two further interactions were significant (P=0.02 for clinical stage by Health Board; P=0.03 for node status by Health Board). However, the model including the interaction between node status and Health Board did not converge properly before the Information matrix became singular and therefore the result presented for that interaction relate to the model that had been fitted in the iteration before this happened and may or may not be acceptable. Thus, this possible interaction must be treated very cautiously.

The model including the interaction between clinical stage and Health Board did converge and so further investigation was necessary. Initially, the numbers in each Health Board for each clinical stage were examined. These numbers are shown in Table 5.8 below.

	Clinical Stage				
Health Board	**I**	**II**	**III**	**Unknown**	**Total**
A	17	53	11	45	126
B	6	8	2	6	22
C	21	44	14	28	107
F	31	48	9	3	91
G	73	180	29	61	343
H	17	31	7	17	72
I	3	8	8	6	25
L	16	66	22	31	135
N	31	73	38	44	186
S	34	162	33	6	235
T	25	79	3	41	148
V	11	31	5	21	68
Y	17	30	6	8	61
Total	302	813	187	317	1619

Table 5.8: Simple breakdown of numbers of cases in each clinical stage for each Health Board.

As can be seen, some of the clinical stage by Health Board combinations have very small numbers in them. These give concern about the stability of the model containing this interaction and, therefore, about the reliability of the estimates obtained for the hazard ratios for these combinations. In fact, when the standard errors were examined for some of the parameter estimates, it was clear that the model was unstable.

To investigate the two interactions further, some of the Health Boards were grouped together and new interactions fitted. The Health Board variable in groups kept the five Health Boards containing Cancer Centres separate (i.e. G, H, N, S and T) and combined the remaining Health Boards into one group to represent 'the rest'. This grouped Health Board variable with six levels was then fitted both as the main effect for Health Board and in interactions with the clinical variables.

Table 5.9 shows the significance for inclusion of these new interactions with the slightly modified 'Clinical Full' model and also the pairs of clinical variables in the original 'Clinical Full' model.

Interactions	P Value
Pairs of Clinical Factors	
A*C	0.815
A*E	0.720
A*N	0.804
A*T	0.749
C*E	0.166
C*N	0.062
C*T	0.859
E*N	0.709
E*T	0.326
Health Board with Clinical Factors	
A*HG	0.320
C*HG	0.085
E*HG	0.163
N*HG	0.647
T*HG	0.405

Table 5.9: Significance for inclusion of the interactions with Health Board in groups in the slightly modified 'Clinical Full' model and the pairs of clinical factors in the 'Clinical Full' model. Age group is given by A, clinical stage (C), ER status (E), tumour size (T), node status (N) and Health Board in (HG) groups respectively.

None of the interactions of the clinical factors with Health Boards in groups were significant.

DISCUSSION

The model including the interaction between node status and all of the Health Boards separately did not converge properly. Although the interaction between clinical stage and all of the Health Boards separately appeared to converge, the estimates of some of the standard errors implied that the model was unstable. When the Health Boards were grouped, the interaction of this variable with clinical stage was not significant (P=0.09). However, the group comprising the 'rest' of the Health Boards consisted of a mixture of very different Health Boards, which may have cancelled out any differences between these Health Boards in terms of the treatment of women with different clinical stage. It does not make sense to group any of the clinical stages together. The conclusion, therefore, was that when the Health Board was grouped, none of the interactions were significant. Overall it, therefore, cannot definitely be concluded that any further interactions were necessary in the 'Clinical Full' model, and therefore none were added.

5.3 MODEL CHECKING FOR ADEQUACY OF FIT AND VALIDITY OF PROPORTIONAL HAZARDS ASSUMPTION

INTRODUCTION

This section now investigates the adequacy of the fit of the 'Clinical Full' model and the validity of making the assumption of proportional hazards in that model.

The adequacy of the model is looked at using Cox-Snell residuals and is discussed more fully in the next section.

The proportionality of the hazards assumption is studied in Section 5.3.2 using two techniques. Firstly, informally through a plot of $\log\left\{-\log \hat{S}(t)\right\}$ versus $\log\{t\}$ and, secondly, by including a time-dependent covariate in the Cox regression model.

5.3.1 EXAMINING THE ADEQUACY OF THE FIT OF THE MODEL

INTRODUCTION

The 'Clinical Full' model defined in Section 5.2.2 is the model with the variables: age group, clinical stage, ER status, pathological node status, pathological tumour size, their two-way interaction and also Health Board of treatment. All of these variables were fitted as categorical factors. How well this model fits is assessed informally by examining the Cox-Snell residuals for all of the cases and then separately for each of the factors by plotting the different levels of the factors.

THEORY - DERIVING THE COX-SNELL RESIDUALS

Let the survival times for the n individuals be $t_1,\ t_2,\ \ldots,\ t_n$ and suppose there are r death times among the n individuals, with $r \le n$. The estimated hazard function from the Cox model for the ith individual with covariate $\underline{x}_i$, $i = 1,\ 2,\ \ldots,\ n$, is given by

$$\hat{h}_i(t) = \exp\left(\underline{\hat{\beta}}^T \underline{x}_i\right)\hat{h}_0(t). \qquad \text{(Eq 5.3.1\_1)}$$

The Cox-Snell residuals are defined, for the ith individual, as

$$\begin{aligned} r_{ci} &= \exp\left(\underline{\hat{\beta}}^T \underline{x}_i\right)\hat{H}_0(t_i) \\ &= \hat{H}_i(\mathrm{t_i}), \text{ by integrating Eq 5.3.1\_1} \\ &= -\log \hat{S}_i(t_i), \qquad \text{(Eq 5.3.1\_2)} \end{aligned}$$

where $\hat{H}_0(t_i)$ is the estimated cumulative baseline hazard, evaluated at the observed survival time for the ith individual.

The following mathematical result is needed to derive the Cox-Snell residuals.

> *Result 1:* If T is the random variable associated with the survival time of an individual with corresponding survivor function of $S(t)$, then the random variable $Y = -\log S(t)$ will have an exponential distribution with unit mean, irrespective of the form of $S(t)$.

If the fitted model is appropriate, then

$$\hat{S}_i(t_i) \approx S_i(t_i).$$

That is, the fitted value of the survivor function is close to the true value of the survivor function for the ith individual at time t_i. Therefore, from Result 1, $-\log \hat{S}_i(t_i)$ should be consistent with being a sample from a unit exponential distribution. The values $-\log \hat{S}_i(t_i)$ are the Cox-Snell residuals, r_{ci} (see Eq 5.3.1_2).

These residuals are unlike those obtained for linear regression as they do not relate the observed value to the expected value. Instead, they are useful for studying how well the residuals fit an exponential distribution with mean one. They are not symmetrically distributed, cannot be negative and are positively skewed.

THEORY - ASSESSING THE FIT OF THE MODEL

From the survivor function for a Weibull distribution, given in Eq 5.1_3 in Section 5.1, for the exponential distribution,

$$\log\{-\log S(t)\} = \lambda \log t .$$

Hence, for data from an exponential distribution with parameter $\lambda = 1$, a plot of $\log\{-\log \hat{S}(t)\}$ against $\log\{t\}$ should be approximately a straight line with intercept at zero and a slope of one (Collett, 1994). Note that $\log\{-\log \hat{S}(t)\}$ is the same as the estimated log cumulative hazard from the known relationship between the hazard and survivor function, given in Eq. 5.1_2 in Section 5.1.

Thus, by analogy, if a plot of $\log\{-\log \hat{S}(r_{ci})\}$ against $\log\{r_{ci}\}$ gives a straight line with slope one and zero intercept, then this implies that the Cox-Snell residuals can be assumed to come from a unit exponential distribution. This, in turn, implies that the fitted model is a good one (from Result 1).

METHODS

The Cox-Snell residuals were obtained by fitting a Cox regression model to the data and saving the cumulative hazard for each individual to give the Cox-Snell residuals, r_{ci}. These values were then taken as the 'survival times' in a Kaplan-Meier analysis and the values of $\hat{S}(r_{ci})$ obtained from this. A plot of $\log\{-\log \hat{S}(r_{ci})\}$, by transforming the $\hat{S}(r_{ci})$ obtained from Kaplan-Meier, against $\log\{r_{ci}\}$, with the r_{ci} obtained from the Cox regression, was examined to see whether the scatter plot of the observations lay roughly on a line with slope one and intercept zero.

It was also possible to look at the log cumulative hazard plots of the Cox-Snell residuals for different levels of each of the factors. If the fitted model is a good one, the points on the plot should be homogenous across the different levels of each factor. If, however, the points for the different levels are widely dispersed, then there would be a suggestion that this factor has not been fully taken into account in the model. The points on the curves only represent the Cox-Snell residuals for the event times.

RESULTS FOR THE 'CLINICAL FULL' MODEL

All Cases: The plot of the log cumulative hazards of the Cox-Snell residuals against the log of the Cox-Snell residuals for all of the cases (Figure 5.5) shows that the 'Clinical Full' model appeared to fit quite well, with only very slight departure from the line with unit slope and intercept zero at small values of $\log\{r_{ci}\}$.

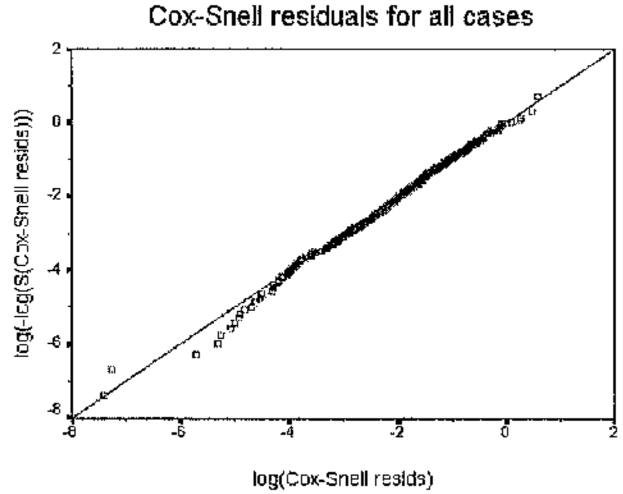


Figure 5.5: Log cumulative hazards plot for Cox-Snell residuals for all cases.

Individual Factors: However, looking at each of the plots for the separate factors: age group, clinical stage and ER status (Figures 5.6 to 5.8) suggests some departures from the line through zero with slope one. Some of the points for the different levels in the factors separated out, rather than overlapping each other. Most of the separation, however, occurred for the early event times and appeared to stabilise for the majority of the residuals.

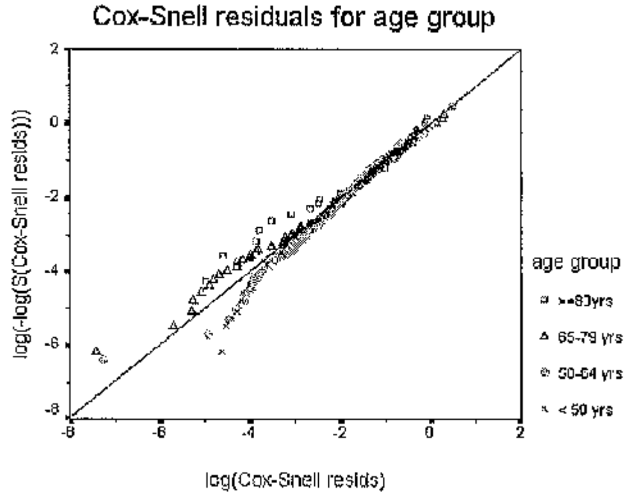


Figure 5.6: Log cumulative hazards plot for Cox-Snell residuals for age group.

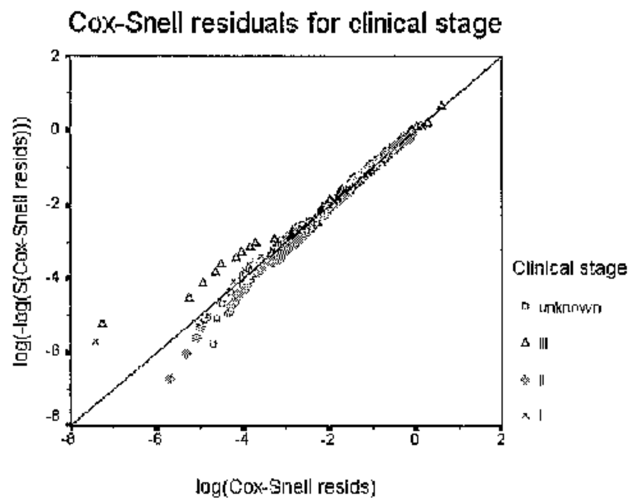


Figure 5.7: Log cumulative hazards plot for Cox-Snell residuals for clinical stage.

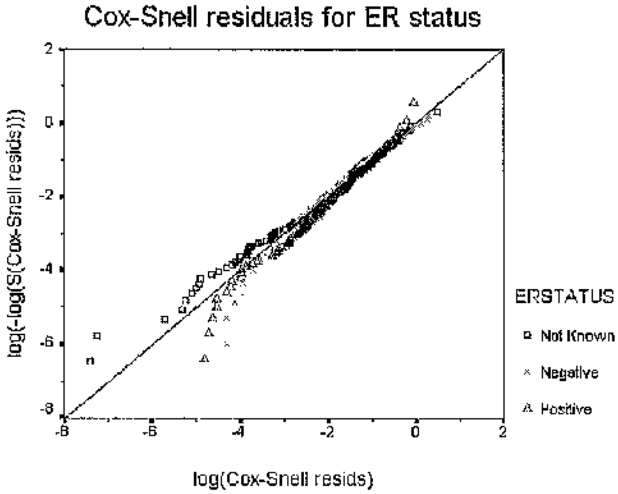


Figure 5.8: Log cumulative hazards plot for Cox-Snell residuals for ER status.

Due to the fact that an interaction term for node status and tumour size was included in the model, it seemed sensible to examine the plot for the combined levels rather than for the factors alone. However, this plot was too busy to make any sense of and, therefore, is not given.

The slight deviations noted from these plots perhaps suggests that there was some lack of fit of the model.

5.3.2 ASSESSING THE ASSUMPTION OF PROPORTIONAL HAZARDS IN THE COX MODEL

INTRODUCTION

This section concentrates on checking the crucial assumption of proportional hazards. This was assessed initially using two informal graphical methods. These plots were obtained from Kaplan-Meier analyses on the individual factors that were ultimately significant in the Cox model. The second method is a formal examination of the proportional hazards assumption using the technique of time-dependent modelling. Before the results are presented, some theory needs to be given.

Log cumulative hazard plots: The survivor function for a Cox model (Section 5.1.2) is

$$S(t) = \left[S_0(t)\right]^{\exp\left(\underline{\beta}^T x\right)}.$$

Therefore,

$$\log\left\{-\log S(t)\right\} = \log\left\{-\log S_0(t)\right\} + \underline{\beta}^T \underline{x}.$$

Thus, $\log\left\{-\log S(t)\right\}$ is a function of time plus a constant.

Therefore, plots of $\log\left\{-\log S(t)\right\}$ versus a function of time should be parallel across different levels of variables of $\underline{x}$. This suggests estimating $S(t)$ within subgroups and plotting $\log\left\{-\log \hat{S}(t)\right\}$ against t, say, for each subgroup, to look for departures from parallelism, indicating non-proportionality. Now $\log\left\{-\log S(t)\right\} = \log H(t)$, which is the log cumulative hazard function, indicating the name of the plot.

Kaplan-Meier survival curves: When the hazards are proportional,

$$h_a(t) = k\, h_b(t)$$

for a factor at two levels, say, a and b, so

$$H_a(t) = k'\, H_b(t)$$

or $$-\log S_a(t) = k'\left\{-\log S_b(t)\right\}.$$

Thus,

$$S_a(t) = \left[S_b(t)\right]^{k'}.$$

Therefore, the survivor function of one level is always greater than or equal to the survivor function of the other level for all times. This argument can be extended to factors with more than two levels.

It is therefore worth examining the Kaplan-Meier survival curves for each of the factors to check whether the estimated curves for the different levels cross or not. If they do cross repeatedly, the assumption of proportional hazards may be in question.

Alternatively, this could also suggest that there is no difference between the levels of the factor. The assumption of proportionality may also be doubtful if the curves diverge considerably.

Both of these informal plots are obtained using the Kaplan-Meier method. In contrast, the next subsection relates to the theory of some formal modelling, where the Cox model is fitted to examine the assumption of proportional hazards.

THEORY - TIME-DEPENDENT MODELLING TO ASSESS THE PROPORTIONALITY OF HAZARDS ASSUMPTION

The assumption of proportional hazards means that the ratios of the instantaneous risks of death for the individuals in each of the levels of a factor are assumed to be constant over time. The technique of fitting time-dependent covariates can be used to assess this assumption.

The idea is to fit the chosen Cox model with an additional term for the interaction between some function of time and one of the covariates. The significance of the parameter for the interaction in the extended model is then examined. This is straightforward for continuous and binary variables, but is more problematic for categorical factors at more than two levels.

Suppose, for simplicity, that all of the covariates are binary with $x_j = 0$ for level 1 and $x_j = 1$ for level 2 of the covariates x_j, $j = 1, \ldots, p$, where $\underset{\sim}{x} = \left(x_1, x_2, \ldots, x_p\right)^T$. Let the validity of the proportional hazards assumption be checked for covariate x_1. An additional term is created to represent the multiplicative interaction of this covariate with a function of time. So let

$$x_{p+1} = x_1 . g(t),$$

where $g(t)$ is any function of time, although it is usual to assume a monotonic form for g.

The hazard for the ith individual is given by

$$\hat{h}_i^*(t) = \exp\left(\underline{\hat{\beta}}^{*T} \underline{x}_i + \hat{\beta}_{p+1}^* x_{p+1}\right)\hat{h}_0^*(t).$$

Now $x_{p+1} = 0$ for level 1 of x_1 and $x_{p+1} = g(t)$ for level 2 of x_1. Therefore, the hazard ratio for being in level 2 versus level 1 for covariate x_1 is given by

$$\exp\left(\hat{\beta}_1^* + \hat{\beta}_{p+1}^* g(t)\right).$$

Thus, the hazard ratio depends on time t and, therefore, is no longer constant for all time, meaning that the hazards are no longer proportional.

The null hypothesis that $\hat{\beta}_{p+1}^* = 0$ is examined using the Wald statistic for this parameter when the interaction of the function of time multiplied by covariate x_1 is added into the model. The Wald statistic is compared to the χ_1^2 distribution for significance.

METHODS

Since the 'Clinical Full' model consisted entirely of factors at more than two levels, it was necessary to create dummy variables for every level for every factor. No unique method exists to circumvent this problem and, therefore, this way was chosen so that each contrast compared each level with the rest of the levels, for each factor, thus providing one degree of freedom tests for each of the interactions.

Thus, for example, for age group, four dummy variables were created as follows:

agelt50 =	1	if age < 50 yrs
	0	otherwise;
age5064 =	1	if age 50-64 yrs
	0	otherwise;
age6579 =	1	if age 65-79 yrs
	0	otherwise;
agege80 =	1	if age ≥ 80 yrs
	0	otherwise.

The dummy variables were named so as to indicate which level was being compared against the rest of the levels for each contrast in each factor. Dummy variables were created in a similar fashion for the other factors. Thus, for example, *stageI* x t gives the interaction of clinical stage I vs the rest (of the clinical stages) with time. For the interaction between node status and tumour size, nine dummy variables were set up to represent the nine different possible combinations for these two factors, each with three levels. Full details are given in the Results section.

The dummy variables were entered separately, one at a time, with the 'Clinical Full' model. Thus, for each factor, a series of Cox models were fitted, each assessing an assumption of proportional hazards. In this manner, it was possible to check whether there was any time-dependency in a particular level versus the rest of that particular factor, after allowing for the effects of the other explanatory variables being in the model. For each model, the other factors were assumed to be independent of time.

For example, for age group, four different Cox models were fitted. The first was the 'Clinical Full' model plus the interaction *agelt50* x t. That is, the interaction of the single contrast age less than 50 years versus the rest with time was added into the 'Clinical Full' model, where age group was already included with three degrees of freedom, with the other three levels compared to age less than 50. The next three Cox models fitted were the 'Clinical Full' model along with *age5064* x t, *age6579* x t and *agege80* x t (the 'ge' standing for greater than or equal to). The results of the modelling are given below, along with the informal plots, for each of the factors separately. The chosen function of time here was simply $g(t) = t$.

RESULTS

Age Group:

Firstly, the Kaplan-Meier survival curves and the log cumulative hazards plots for age group are presented (Figures 5.9 and 5.10 respectively). The Kaplan-Meier curves do not cross, except for the curves for the <50 and 50-64 groups. These two groups appear to be nearly identical as they cross several times.

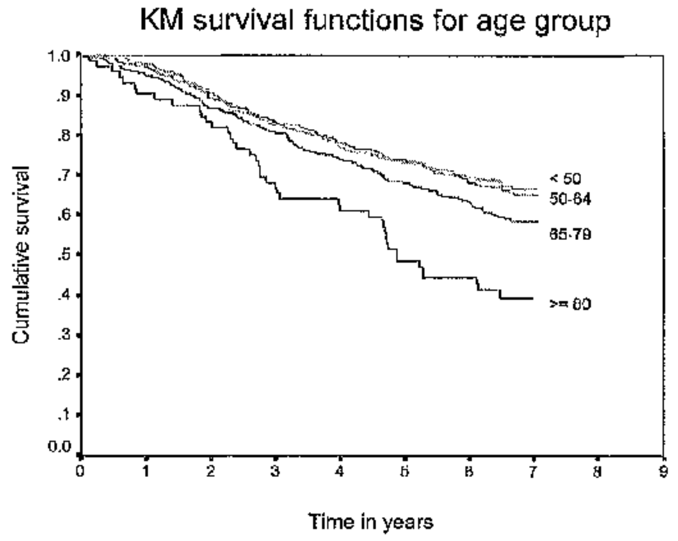


Figure 5.9: Kaplan-Meier survival curves for age group.

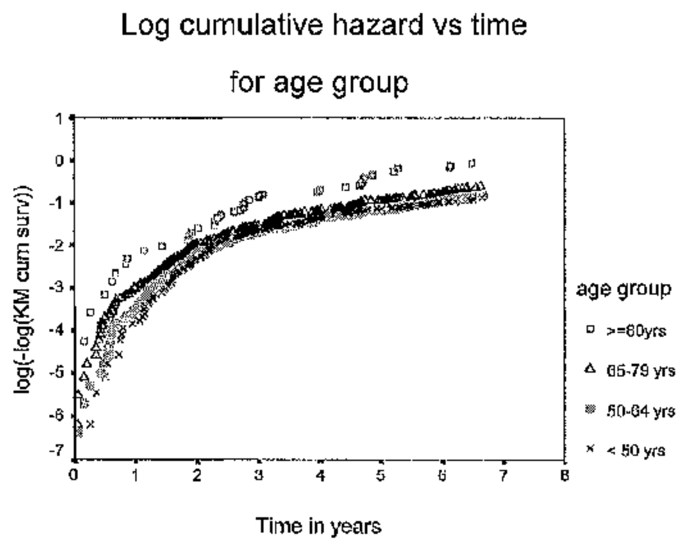


Figure 5.10: Log cumulative hazards plot for age group.

Looking at the log cumulative hazards plot shows that the scatter plots for each of the levels are nearly parallel. Perhaps, there is a slight suggestion that the curves come together. However, this could just be due to the fact that there were not many observations early on and that too much weight is being given to the data occurring in the first two years. In fact, there were only 58 deaths in total during the first year.

The dummy variables created to examine the proportional hazards assumption for age group were given in the Methods section above. Table 5.10 below gives the results for fitting the four interactions of the dummy variables with time.

	Parameter estimate	Standard error	P value for Wald statistic
agelt50 x *t*	-0.0400	0.0547	0.4649
age5064 x *t*	0.0165	0.0504	0.7437
age6579 x *t*	0.0098	0.0510	0.8479
agege80 x *t*	0.0274	0.0927	0.7676

Table 5.10: Results of time-dependent modelling for age group. Each of the contrasts represent one level of the factor vs the complementary levels in the interaction with time. Note that lt and ge stand for less than and greater than or equal to respectively.

None of these interactions were significant at the 5% level and, therefore, there was no reason to reject the assumption of proportional hazards for age group.

Clinical Stage:

Examination of the Kaplan-Meier survival curves for clinical stage (Figure 5.11) shows that the curves do not really cross except, perhaps, the unknown group and stage II, showing that the group of unknowns are very similar to the group with stage II disease. Alternatively, the unknowns could be a mixture of all three clinical stages and the mixture just happened to be similar to the stage II group. Figure 5.12 gives the corresponding log cumulative hazards plot for clinical stage.

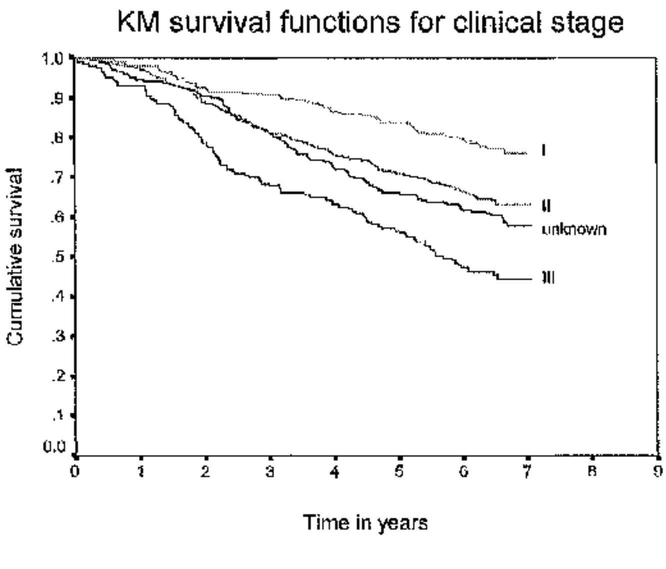


Figure 5.11: Kaplan-Meier survival curves for clinical stage.

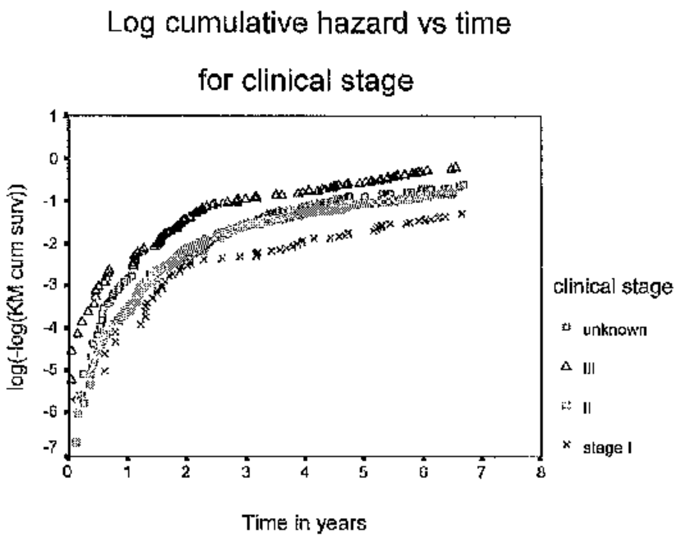


Figure 5.12: Log cumulative hazards plot for clinical stage.

Again, the curves seem to be reasonably parallel after the first two years.

None of the interactions were significant when the formal modelling for clinical stage with time was performed (range of P values for contrasts from 0.46 to 0.99). Thus, there is no evidence to assume that the hazards were not proportional for clinical stage.

ER Status:

The Kaplan-Meier survival curves for ER status is given in Figure 5.13. It does not show any serious crossing, except the ER negative curve drops down dramatically at about six months. In fact, out of 22 events that occurred before six months, 16 were in the ER status unknown group. This was compared with two in the ER positive group and four in the ER negative group. Discussion of this seemingly strange pattern is given below and in the next section.

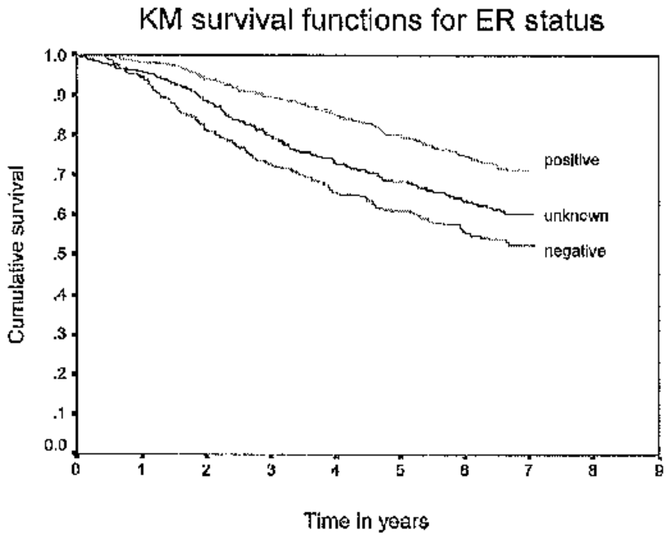


Figure 5.13: Kaplan-Meier survival curves for ER status.

The log cumulative hazards plot for ER status (Figure 5.14) reflects the pattern observed in Figure 5.13.

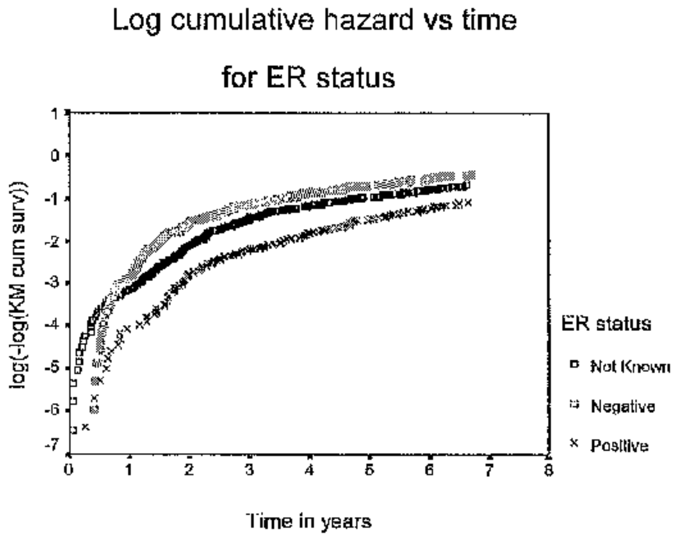


Figure 5.14: Log cumulative hazards plot for ER status.

The main problem occurs prior to the first year of survival. The curves are not parallel and the sudden drop in survival observed in the Kaplan-Meier plot (Figure 5.13) for the ER negative group is mirrored by the rapid increase in the log cumulative hazard for this group after the six months mark (Figure 5.14). After the first year of survival, however, the curves are nearly parallel, although there is a slight suggestion that they converge on each other. The occurrence of this pattern is due to the fact that until three months the hazard ratio for being ER negative vs ER positive is one, as no events happened for either group. There was an event at three and at five months in the ER positive group. Four events occurred between five and six months in the ER negative group. After this point, the hazard increases for both groups. This increase is more rapid for ER status negative than positive. Eventually, however, the hazard ratio for being ER negative vs ER positive begins to attenuate as the gradient of the estimated survival curve for ER negative, seen in Figure 5.13, becomes less steep.

The results of fitting the interactions of the dummy variables for this factor with time are given below in Table 5.11.

	Parameter estimate	**Standard error**	**P value for Wald statistic**
erpos x *t*	0.1904	0.0547	0.0005
erneg x *t*	-0.1153	0.0530	0.0298
eruk x *t*	-0.0559	0.0488	0.2518

Table 5.11: Results of time-dependent modelling for ER status. Each of the contrasts represent one level of the factor vs the complementary levels in the interaction with time. Note that pos, neg and uk stand for positive, negative and unknown respectively.

Both of the interactions of time with ER status positive vs rest and ER status negative vs rest were significant. Thus, there appeared to be a changing risk ratio of death over time for ER status.

The signs of the parameter estimates for the two interactions were positive for ER positive vs rest with time and negative for ER negative vs rest with time. Thus, the risk ratio for being ER negative decreased with time, although Figure 5.13 shows that women having ER negative tumours had a poorer prognosis than those with ER positive tumours. One interpretation of these results could be that being ER negative carried an important additional risk in the short term, but the magnitude of the risk decreased over time. These findings are discussed further in Section 5.3.3.

Node Status with Tumour Size:

Neither the Kaplan-Meier survival curves nor the log cumulative hazards plots are given here because the plots showing the nine combinations for this interaction were too busy to interpret meaningfully.

When the formal time-dependent modelling was performed for this factor, none of the interactions were significant and, therefore, the proportional hazards assumption could not be rejected for the node status by tumour size interaction with time.

Health Board:

Again, due to the large number of levels (13) for Health Board, neither the Kaplan-Meier survival curves nor the log cumulative hazards plots are given here. The time-dependent modelling for this factor revealed that two of the interactions were significant at the 5% level. These were the Islands Health Board (I) vs the rest (P=0.034) and Forth Valley HB (V) vs the rest (P=0.018). If these two interactions were not just due to chance, then this meant that the ratios of the risks of death for these two levels with the others were changing over time.

The hazard ratio for the Islands Health Board vs Greater Glasgow Health Board (G) in the 'Clinical Full' model was less than one, implying a decreased risk. However, because the parameter estimate for the interaction with time was positive, then women treated in the Islands HB appeared to have an increasing risk with time. In contrast, women treated in Forth Valley (V) appeared to have a decreasing risk compared to the rest over time. This Health Board had an increased risk of death compared to Greater Glasgow HB in the 'Clinical Full' model.

It is hard to interpret exactly what is happening when there are so many levels involved. This problem is heightened by the fact that in the 'Clinical Full' model the contrasts for the full Health Board factor, with twelve degrees of freedom, were all compared with Greater Glasgow, whereas in the interactions with time the thirteen individual contrasts of each level for time dependence were each compared to the rest of the twelve levels together. Possible interpretations for these two significant interactions are given in the next section.

5.3.3 CONCLUSIONS FROM THE MODEL CHECKING

INTERPRETATIONS OF THE RESULTS

Adequacy of the model: The previous two sections presented the results from examining the adequacy of the fit of the Cox model and of the validity of the proportional hazards assumption. From Section 5.3.1, the global plot of the Cox-Snell residuals for all cases seemed to fit adequately (Figure 5.5). However, when the subgroups were examined, the plots for the individual factors (Figures 5.6 to 5.8) showed that there was some suggestion that the model did not entirely fit adequately and that the factors had not fully been taken into account in the Cox model. However, when other interactions (the interaction of node status by tumour size is already present in the 'Clinical Full' model) were searched for in Section 5.2.4, it was found that none of the interactions were significant when added to the modified 'Clinical Full' model, when the Health Board variable was grouped.

Proportional hazards assumption: The finding that there were significant interactions with time for two of the three levels of ER status and two of the 13 levels of Health Board is perhaps of greater concern.

(i) **ER status:** Examination of Figures 5.13 and 5.14 in Section 5.3.2 showed that the ER status unknown group had the worst prognosis early on (up to about six months) and then the group became the intermediate prognostic group once deaths in the ER negative group became more abundant. It was found that 16 out of 22 women who had died before six months did not have the ER status of their tumours determined. This number appeared to be too great to simply be due to chance. Possible reasons to explain why this pattern was observed have been explored, following discussions with clinical colleagues.

One interpretation could be that some women had such poor prognoses that they did not have their ER status ascertained. This seemed an unlikely explanation, however, since the cohort for these analyses included only those women who were deemed fit enough for surgery and in whom no documentation of evidence of metastatic disease at presentation was found.

Another suggestion was that the 16 women with ER status not recorded and who died within the first six months were treated at rural hospitals, which may have low annual breast cancer case loads and no definitive protocols for staging the tumour. However, when the list of hospitals performing the surgery on these 16 cases was collated, this argument could not be supported (data not given).

A further possibility examined was that these women died suddenly of a cause of death not related to breast cancer; for example, a heart attack. However, ten of the cases had breast cancer as the primary cause of death; two cases had another cancer (lung and stomach, which probably were actually metastases from the breast cancer) and only four cases had deaths due to heart or pulmonary problems. Therefore, this suggestion was also not supported.

Thus, leading to the remaining possible clinical reason that all of the women dying before six months did in fact have metastases despite that there was no mention of them in the documentation. In many cases, metastases are not routinely searched for unless they are suspected at presentation or there is extensive lymph nodal involvement. Section 4.2.3 showed that ER status was more likely to be missing if node status was missing than if node status were known. However, examination of staging of node status and tumour size was similar in the women dying early, with and without ER status recorded. Thus, there is no obvious reason why having metastases not detected at presentation should be more prevalent in women with ER status unknown, than in those women where it was known.

None of these clinical reasons explained the observation and so it seems, therefore, that it was due to chance that 16 of the 22 deaths in the first six months were for women with unknown ER status.

It is plausible that the effect of the ER status covariate changes over time and the proportional hazards assumption does not hold. That is, ER status has an effect on outcome immediately after diagnosis and treatment, but this effect is not maintained over a long time. Miller et al (1994) support this. Collett et al (1998) found that the effect of ER status on a prognostic index they were deriving was strong in the first five years, but then weaker after that. They also highlighted a lessening importance of ER status on survival with time. However, many other studies have included ER status in Cox regression models (Gordon et al, 1992; Hawkins et al, 1996; Newman et al, 1997; Shek & Godolphin, 1988) without reporting any non-proportionality.

In terms of interpreting the results from fitting the 'Clinical Full' model, it is likely that the hazard ratios for ER status were not overly biased in terms of the order of the hazards and that the model probably gave a reasonable estimate of the average hazard ratios. It is possible, however, that it may over-estimate the long term predictions for the importance of ER status on survival from breast cancer.

(ii) **Health Board:** This is the other factor where significant interactions of two of the levels of the factor with time were found. One interpretation of the fact that the Islands (I) Health Board had a significant interaction with time with a positive parameter estimate but with the Health Board in the 'Clinical Full' model having a negative parameter estimate could be that this perhaps implies that these women had sub-optimal follow-up treatment. Maybe the women chose not to travel to either Glasgow or Inverness to receive radiotherapy subsequent to any breast conserving surgery on the Islands.

In the circumstance of the Forth Valley (V) Health Board interaction with time, however, the signs of the parameters probably indicate that the initial treatment was poorer than that received by women treated in Greater Glasgow (G) Health Board, but this increased mortality risk decreased over time.

These significant interactions could possibly reflect changing patterns of care over time in the Health Boards, which may have happened at different times in the different Health Boards, causing the effects on survival over time to change in these Health Boards.

However, the interactions with the two Health Boards could also be due to chance. After all, 33 tests in total were conducted and presented in the last section so that it would be expected that at least one would be significant at the 5% significance level merely by chance.

WEAKNESSES

Although the usefulness of the 'Clinical Full' model has been questioned by the findings from the model checking, several weaknesses of the methods employed should be borne in mind when interpreting the findings.

Cox-Snell residuals: Firstly, Collett (1994) points out that the use of the Cox-Snell residuals may not be appropriate if small samples are involved. This is because the distributional results relating to a unit exponential distribution may not be valid. However, this is probably not a necessary caveat in this particular situation, except perhaps for some of the Health Board levels.

The main problem with using the Cox-Snell residuals plots, however, is the informal nature of them. The interpretation of these plots is entirely subjective and it can be difficult to judge whether the observations lie within the margin of error expected due to fitting estimated values.

Kaplan-Meier and log cumulative hazards plots: Similarly, using the plots of the Kaplan-Meier survival curves and the log cumulative hazards plots to assess the proportional hazards assumption presents the same problem of subjectivity. It is not too difficult to spot survival curves that cross, although it is necessary to remember that the crossing may just be due to fitting estimated values. It is slightly more awkward to decide whether the estimated log cumulative hazard lines are parallel for most of the time. Although, in theory, the Kaplan-Meier curves would not be expected to cross and the curves on the log cumulative hazards plot to be parallel, in practice there would be some deviance from the expected positions because only the estimated values were being plotted.

Formal time-dependent modelling: Using formal time-dependent modelling has the benefit in that it can be assessed by formal tests derived using statistical inference. There are no problems with this method for the simple situations when the covariates are either continuous or binary. However, it is not entirely clear how to perform the modelling, or interpret the results, when the covariates are categorical factors with more than two levels, as there is no unique method in this situation. The use of the dummy variables seemed to be an acceptable method for partially assessing the proportional hazards assumption for each of the factors. This method should be powerful if one of the groups were different from the rest.

DISCUSSION

The validity of predictions made from the 'Clinical Full' model may have to be treated cautiously, although all of the weaknesses described above should be taken into account before the soundness of the model is ruled out completely. The 'Clinical Full' model probably provides acceptable average hazard ratios for the factors in the short term, but may be more questionable in the longer term.

One disadvantage of fitting a Cox model is that it does not allow the effect of a covariate on survival to diminish over time. Instead of fitting a Cox model, non-proportional hazards models could have been fitted to the data. Unfortunately, due to time constraints, this was not pursued here. However, Gore et al (1984) fit various non-proportional models to a series of nearly 4,000 women with breast cancer referred to one hospital between 1954 and 1964. They found that the hazard functions converged over time. Schemper (1992) examines, theoretically, violations of the proportional hazards assumption in a Cox model.

5.4 SURVIVAL ANALYSIS INTERPRETATIONS WITH RESPECT TO MISSING VALUES

Whether the missing values in the four main clinical variables were related to the Health Board of treatment is discussed in the next section. Approaches to handling the missing values are examined in Section 5.4.2 to investigate whether the method influenced the results and interpretations from the survival analyses. Possible explanations why the variables age and clinical stage had different results in the models based on all cases and complete cases only are discussed in Sections 5.4.3 and 5.4.4.

5.4.1 MISSING VALUES IN THE HEALTH BOARD OF TREATMENT

INTRODUCTION

Here, univariate associations between Health Board of treatment and having missing values in the clinical factors are examined.

RESULTS

The percentages of cases with complete data in each of four clinical factors are tabulated for the thirteen Health Boards (Table 5.12). The P values for the χ^2 tests of association for missingness of data for the clinical factors across Health Boards were all very highly significant (P <0.0001 for all tests). Therefore, whether or not the information was missing for each of the clinical variables depended on which Health Board the woman had her surgery in.

Health Board	%complete in C	%complete in N	%complete in T	%complete in E	%complete in all 4	Number of cases
A	64.3	67.5	75.4	83.3	34.9	126
B	72.7	90.9	81.8	9.1	0.0	22
C	73.8	78.5	79.4	62.6	31.8	107
F	96.7	69.2	81.3	69.2	45.1	91
G	82.2	76.4	80.5	77.0	47.5	343
H	76.4	79.2	87.5	4.2	2.8	72
I	76.0	64.0	88.0	4.0	4.0	25
L	77.0	67.4	77.8	43.0	23.7	135
N	76.3	82.3	79.0	72.0	40.3	186
S	97.4	84.7	85.5	88.5	64.7	235
T	72.3	66.9	64.2	35.1	15.5	148
V	69.1	52.9	73.5	1.5	1.5	68
Y	86.9	31.1	91.8	52.5	16.4	61
Overall	80.4	73.1	79.5	61.1	35.7	1619

Table 5.12: Percentages complete in each of the four clinical prognostic factors separately and in all four of them together by Health Board. Note that C, N, T and E stand for clinical stage, node status, tumour size and ER status respectively.

To try to simplify the findings, a variable was created to represent those Health Boards which contain the five Cancer Centres (HBs: G, H, N, S and T) as one level (CC) versus those Health Boards which do not have a Cancer Centre (No CC). The results for the grouped Health Boards are given below in Table 5.13.

	%complete in C	%complete in N	%complete in T	%complete in E	%complete in all 4	Number of cases
No CC	76.7	65.2	79.5	51.8	25.7	635
CC	82.8	78.3	79.5	67.2	42.2	984
Overall	80.4	73.1	79.5	61.1	35.7	1619

Table 5.13: Percentages complete in each of the four clinical prognostic factors separately and in all four of them together by whether or not there was a Cancer Centre (CC). Note that C, N, T and E stand for clinical stage, node status, tumour size and ER status respectively.

This table shows that the four main prognostic factors were available more frequently in the larger Health Boards, containing the Cancer Centres, than the smaller non-Cancer Centre Health Boards. Table 5.14 gives the corresponding P values for the χ^2 tests of association for Cancer Centre Health Board group against having missing values in the clinical factors.

	P Value
CC y/n with C kw or nk	0.002
CC y/n with N kw or nk	<0.001
CC y/n with T kw or nk	0.980
CC y/n with E kw or nk	<0.001
CC y/n with all four vars kw or nk	<0.001

Table 5.14: P values for χ^2 tests of association for Cancer Centre Health Board (CC) with the four clinical prognostic factors as either known (kw) or missing (nk). Note that C, N, T and E stand for clinical stage, node status, tumour size and ER status respectively and 'vars' for variables.

Thus, it appears that whilst there were differences among the ascertainment of pathological tumour size for all of the Health Boards, on average, there were no differences between the Cancer Centre Health Boards and the non-CC Health Boards.

5.4.2 MODELS FOR COMPLETE AND ALL CASES DATASETS

INTRODUCTION

One of the main aims of this thesis was to examine the influence of missingness of data on the results of the Cox regression analysis reported by Twelves et al (1998a), performed on the 1619 surgical cases. The model fitted in that paper is referred to here as the 'Clinical Full' or all cases model (ACM). The results of this were summarised in Section 5.2.2 and will be further discussed here.

The technique used by Twelves et al (1998a) for handling the missing data was to add extra categories for the unknown values in each factor. The assumption that the missing data were missing at random (MAR) was implicitly made when the model was fitted, although this cannot be tested directly, as discussed in Section 4.3. However, comparisons of this method with the complete cases method and also fitting the two partially-complete cases models, suggested in Section 4.4, are examined to investigate whether the results are consistent or disparate for the different models.

COMPLETE CASES ANALYSIS

Initially, a Cox model was fitted on the 578 cases for which there was known information for the four main clinical variables: clinical stage, node status, tumour size and ER status. However, examination of the results showed that the model produced parameter estimates which were unstable.

Table 5.15 below provides a breakdown of the numbers of cases in each of the Health Boards in the all cases model and those left when only the cases with complete information were retained. The percentage remaining for each Health Board is also given.

When a model based on the complete cases only was fitted excluding the Health Board variable, the standard errors obtained for the factors in the model were of similar magnitude to those obtained when the model was fitted on all cases.

Health Board	Number in all cases model	Number (%) with complete information	
A	126	44	(34.9)
B	22	0	(0)
C	107	34	(31.8)
F	91	41	(45.1)
G	343	163	(47.5)
H	72	2	(2.8)
I	25	1	(4.0)
L	135	32	(23.7)
N	186	75	(40.3)
S	235	152	(64.7)
T	148	23	(15.5)
V	68	1	(1.5)
Y	61	10	(16.4)
Total	1619	578	(35.7)

Table 5.15: Number of cases in each Health board when all cases and when only those with complete information were included.

Since the Health Board factor had not been expected to be significant a priori in the all cases analysis, it was important to try to ascertain whether it was present in that model only because of the presence of incomplete information in some of the other variables or because real differences existed among the survival chances of women treated in different Health Boards.

To try to address this, the four Health Boards (Borders (B), Highland (H), the Islands (I) and Forth Valley (V)) with only zero, two, one and one case respectively left with complete information were excluded and another Cox model obtained.

The variables present in the ACM were age, clinical stage, ER status, node status, tumour size, the interaction between these two variables, and Health Board of treatment. When the model was derived for the complete cases only, using the technique of forward stepwise selection, the variables age and clinical stage were not significant (P values for non entry were 0.30 and 0.14 respectively).

To allow comparison of the hazard ratios for these two factors between the two models, the factors were forced into the complete cases analysis. It was assumed that the addition of these non-significant factors into the model would not affect the results for the other factors noticeably. This complete cases analysis, with the reduced Health Board factor, was based on only 574 cases and is henceforth referred to as the complete cases model (CCM). The P values for ER status, node status, tumour size, their interaction and Health Board of treatment were all very highly significant (<0.001) in the CCM.

RESULTS - COMPARISON OF THE ALL CASES AND COMPLETE CASES MODELS

(i) Hazard Ratios

Table 5.16 below gives the hazard ratios with 95% CIs for the two models. The estimates for the unknown levels in the all cases model are not presented here as the objective was to compare the findings with the complete cases model. For the same reason, no estimates were given for the ACM for the four Health Boards which were excluded from the complete cases analysis. The results for the unknown levels for the ACM have already been detailed in Table 5.6 of Section 5.2.2.

Unfortunately, it was not possible to test formally whether the results were different using a statistical test because the two sets of estimates were not independent, since the women included in the complete cases model also belonged to the all cases model.

	All Cases Model		Complete Cases Model	
Variable	**Hazard Ratio**	**95% CI for Hazard Ratio**	**Hazard Ratio**	**95% CI for Hazard Ratio**
Age				
< 50	1	*	1	*
50-64	1.04	0.84, 1.29	0.95	0.67, 1.33
65-79	1.18	0.95, 1.47	1.01	0.68, 1.48
≥ 80	2.01	1.39, 2.90	3.25	0.75, 14.09
Clinical Stage				
I	1	*	1	*
II	1.41	1.42, 2.78	1.64	0.98, 2.73
III	1.98	1.13, 2.09	1.66	0.89, 3.09
ER Status				
Positive	1	*	1	*
Negative	2.11	1.69, 2.63	3.04	2.23, 4.16
Node Status by Tumour Size				
N+ T≤2	3.91	2.62, 5.84	4.73	2.72, 8.21
N+ T >2	4.37	3.01, 6.35	4.87	2.87, 8.27
N- T≤2	1	*	1	*
N- T >2	2.72	1.82, 4.07	2.64	1.52, 4.59
Health Board				
A	1.52	1.10, 2.10	1.20	0.67, 2.16
C	1.49	1.06, 2.10	2.74	1.56, 4.82
F	1.55	1.05, 2.29	2.70	1.57, 4.65
G	1	*	1	*
L	1.20	0.86, 1.66	0.97	0.51, 1.84
N	0.95	0.69, 1.31	1.16	0.70, 1.91
S	0.88	0.65, 1.19	0.87	0.58, 1.31
T	1.33	0.94, 1.87	1.87	0.90, 3.89
Y	1.11	0.71, 1.76	0.18	0.02, 1.29

Table 5.16: Hazard ratios (HR) with 95% CIs for the two analyses. Note that N and T stand for node status and tumour size respectively.

It is possible to obtain an idea about differences between the models by simple examination of Table 5.16 and Figure 5.15 below. The first observation to note is that some of the confidence intervals on the hazard ratios are very wide. Thus, qualitatively the findings appear to be quite similar, with almost the same patterns observable for the ordering of risk among the levels of the prognostic factors.

Quantitatively, there is a suggestion that there are more extreme hazard ratios in the complete cases analysis. For example, for the clinical factors, except for clinical stage, the poorer prognostic levels (ER negative, node positive and large tumour size) appeared to be more severe for the complete cases model. For example, for ER status, there was an increased hazard ratio of 3.04 (95% CI 2.23, 4.16) for the CCM compared to 2.11 (1.69, 2.63) for the ACM for ER negative relative to the baseline ER positive, although it is unknown whether the two were statistically different.

For the Health Board factor, Ayrshire & Arran Health Board (A) did not have a statistically significant hazard ratio compared with Greater Glasgow Health Board (G) in the complete cases model (Figure 5.15).

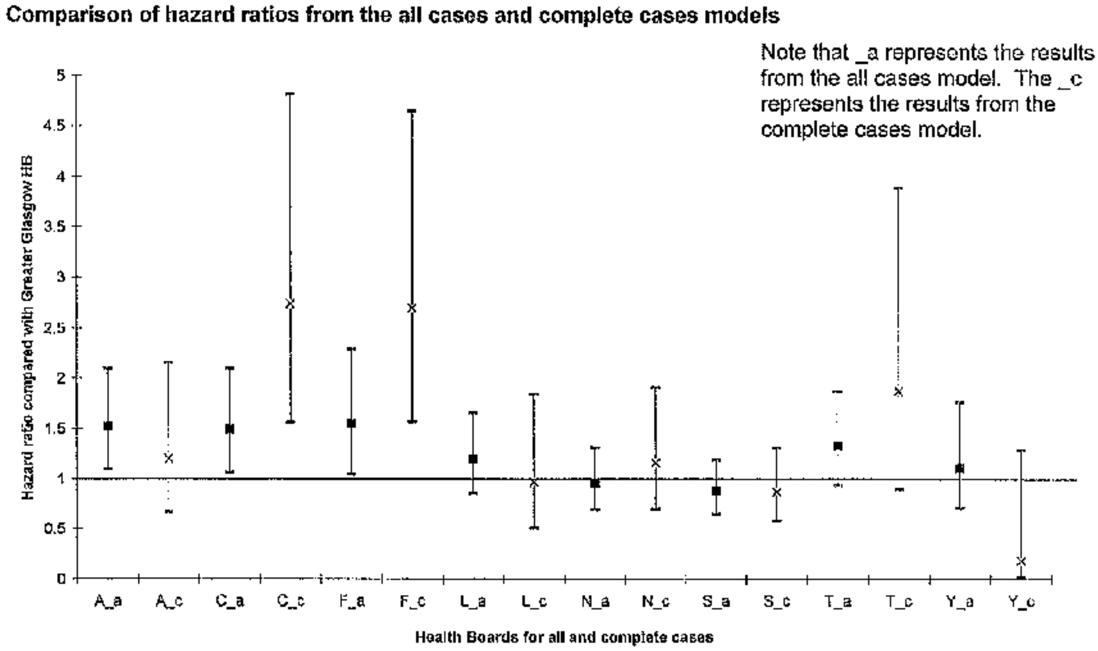


Figure 5.15: Hazard ratios with 95% CIs for the all cases and complete cases models for the nine Health Boards.

Examination of the ranks of the hazard ratios (Table 5.17), on the basis of the point estimates for the nine Health Boards shows some variation, but the ranks can be split into two distinct groups, ranks 1-4 and 5-9.

	Rank	ACM	CCM
High hazard	1	F	C
	2	A	F
	3	C	T
	4	T	A
	5	L	N
	6	Y	G
	7	G	L
	8	N	S
Low hazard	9	S	Y

Table 5.17: The ranks of the Health Boards in the ACM and CCM on the basis of the hazard ratios compared with Health Board G.

This does not demonstrate that only F, A and C had a statistically significantly higher risk than G in the all cases model, and in the complete cases model, only F and C were statistically different from G.

It is possible that including the four Health Boards which were dropped from the CCM (HBs: B, H, I and V) in the ACM altered the findings for the other variables. Therefore, another model with unknown values included in the prognostic factors was fitted. This model excluded the women in these four Health Boards, leaving 1432 cases. Table A6.1 in Appendix 6 gives the hazard ratios with 95% CIs for this model. Although there were some minor differences between the results for the two models with unknown values included (based on 1619 and 1432 cases respectively), none were very striking when compared with Table 5.16 in this section and Table 5.6 in Section 5.2.2. It appears, therefore, that including the four Health Boards in the ACM did not greatly influence the results for the ACM, in terms of making comparisons with the CCM.

Whilst the hazard ratio has the benefit of depending only on the parameter estimates calculated for the model, it can sometimes also be informative for clinicians to see the effect of differences in hazards on overall survival, say at a particular point in time, or on survival for each of the different levels of a factor. Therefore, the 5-year survival estimates are now considered.

(ii) 5-year % Survival Estimates

Before the results from the Cox models are presented, univariate Kaplan-Meier analyses are discussed. These were performed on the 574 women in the complete cases model and are compared with the results based on all cases discussed in Section 5.2.1 and by Twelves et al (1998a). Table A6.2 of Appendix 6 gives the Kaplan-Meier 5-year % survival estimates with 95% CIs based on all cases and only on the complete cases, although only the results for the known levels are given for the all cases Kaplan-Meier analyses. Table 5.18 below gives the P values for the univariate log-rank tests of equality of the survival curves for the different levels of the different factors from both the all cases and the complete cases analyses.

The overall Kaplan-Meier survival based on the complete cases was 72.7% (95% CI: 69.0%, 76.3%) compared with 70.9% (68.6%, 73.1%) for all cases. Thus, there is a suggestion that the subgroup with complete prognostic information had a slightly better survival, but this was probably not significant. Again, it is not possible to formally test whether they are different as the two groups are not independent.

Factor	**P value based on all cases**	**P value based on complete cases**
CLINICAL FACTORS		
Age	<0.0001	0.74
Clinical stage	<0.0001	0.0001
Pathological node status	<0.0001	<0.0001
Pathological tumour size	<0.0001	<0.0001
ER status	<0.0001	<0.0001
SERVICE FACTORS		
HB of first treatment	0.02	0.14
Deprivation	0.03	0.02
Surgical case load	0.03	0.13
Seen by an oncologist	0.25	0.18
TREATMENT FACTORS		
Type of surgery	0.01	0.01
Adjuvant radiotherapy	0.49	0.27
Adjuvant chemotherapy	0.02	0.0008
Adjuvant endocrine therapy	0.74	0.003
Adjuvant chemotherapy or endocrine therapy	0.28	0.16

Table 5.18: P values for the overall log-rank tests of equality of the survival curves in univariate analyses.

Examination of Table 5.18 reveals that there are differences in the sets of factors which were significant in the two cohorts, one of which is that Health Board was not significant univariately when based on the complete cases (P=0.14). However, it was significant in the multivariate Cox model based on complete cases when other factors had been adjusted for (P=0.0001). There was no evidence of differences among the survival curves for surgeon case load in the complete cases Kaplan-Meier analysis. This could be due to surgical case load for all cases having different survival prospects because the Team/30+ group staged their cases more thoroughly and, therefore, had less unknown information. Thus, the women under their care may have received more appropriate treatment.

It is interesting that there were statistical differences in survival for all of the treatment factors, except use of any adjuvant treatment, in the univariate log-ranks tests of equality of the complete cases. The differences observed in the complete cases situation univariately may be due to women with complete staging information receiving optimal treatment. This may mean that some women were not given treatment because their disease had been staged, who may have been given it had their staging information been unknown. However, when the treatment factors were added to a Cox model with the clinical factors, but not Health Board, based on 574 cases, none were significant with P values ranging from 0.06 for any adjuvant systemic therapy to 0.90 for type of surgery.

In the publication, Twelves et al (1998a), 5-year % survival estimates obtained from the Cox model were presented for each of the factors, by 'averaging' over the other factors. These included an 'average profile' by Health Board (Section 5.2.2). These estimates were made up of weighted averages of each of the levels of each of the other factors (Section 5.2). It would be possible to present similar figures for the nine Health Boards included in the CCM. However, there would be problems of interpretation due to differences in the frequency distributions for the factors for the two models (see Table 5.19 below). Thus, the 'average profile' by Health Board (HB) for the two models would not be comparing 'like' with 'like', because the weighting of the risks in the other factors would be different in the two models (Eq 5.2_1 in Section 5.2). This could lead to mis-interpretations of any differences observed between the results for the two models.

	All Cases Model			Complete Cases Model	
Variable	**Number of Cases**	**%**	**% when nks excluded**	**Number of Cases**	**%**
Age					
< 50 years	476	29	*	180	31
50 - 64	591	37	*	245	43
65 - 79	480	30	*	145	25
≥ 80 years	72	4	*	4	7
Clinical Stage					
Stage I	302	19	23	102	18
II	813	50	62	391	68
III	187	12	14	81	14
Not known	317	20	*	*	*
ER Status					
Positive	599	37	61	352	61
Negative	391	24	39	222	39
Not known	629	39	*	*	*
Node Status by Tumour Size					
N nk, T ≤ 2cm	185	11	*	*	*
N nk, T > 2	138	9	*	*	*
N nk, T nk	112	7	*	*	*
N +, T ≤ 2cm	171	11	18	101	18
N +, T > 2	312	19	32	183	32
N +, T nk	100	6	*	*	*
N -, T ≤ 2cm	269	17	28	157	27
N -, T > 2	212	13	22	133	23
N -, T nk	120	7	*	*	*

Table 5.19: Observed frequency distributions for the two analyses. Note that N and T stand for node status and tumour size respectively. Also, nk stands for not known.

One potential way to avoid comparing 'average profiles' by HB was to obtain survival estimates for particular levels of particular factors. To reduce the number of combinations (96 possible for the known values) to a more manageable set, the eight different combinations of ER status, node status and tumour size by Health Board were presented for age group 50-64 and clinical stage II (the largest levels in both factors) for the two models. Therefore, all of the weights for the risks (Eq 5.2_1 in Section 5.2) were now 1 in each of the eight groups for each of the 13 Health Boards separately.

When the 5-year % survival estimates for the two models are compared, it is necessary to be aware that any observed difference between the two models for a particular combination and a particular Health Board (known here as Health Board, prognostic groups) could be due to different parameter estimates for the Health Board from the two models; different values of the linear combinations of the other parameters from the two models; or simply due to the two models having different baseline survival estimates at that time point (or a combination of all three possibilities).

The results for the ACM and CCM are given in Tables 5.20a and 5.20b respectively. The eight prognostic groups given in these tables have been sorted into order of prognosis, based on the all cases model, from best to worst outcomes. The ↑ and ↓ in the body of Table 5.20b highlight which Health Board, prognostic group combinations had estimates of 5-year % survival in the CCM which were at least 10% **in absolute magnitude** greater than or smaller than respectively those of the ACM. The standard errors for the survival estimates for the Health Boards for the eight prognostic groups are not given due to the apparent problem with them using SPSS (Appendix 5).

Several comments can be made about Tables 5.20a and 5.20b. Note that all changes of percentage survival estimates stated below relate to absolute changes in percentages rather than percentage changes between the two models.

(1) The effect of the node status by tumour size interaction on survival can be observed clearly for both models. For example, comparing group E+, N-, T≤2 with E+, N-, T>2 shows roughly a difference of 10% for most Health Boards; whereas E+, N+, T≤2 vs E+, N+, T>2 has a difference of only about 3%. Thus, the large tumour size (poor prognosis) had more of an effect when node status was negative (good prognosis) than when node status was positive (poor prognosis).

Similarly, looking at group E+, N-, T≤2 compared to E+, N+, T≤2 demonstrates a difference of about 25% for most Health Boards. The corresponding difference for E+, N-, T>2 vs E+, N+, T>2 is about 12%. Therefore, being node positive (poor prognosis) had a larger effect when tumour size was small (good prognosis) than when it was large (poor prognosis).

All Cases Model: 5-year % survival for age group 50-64 and clinical stage II								
Health Board	**E+, N-, T≤2**	**E-, N-, T≤2**	**E+, N-, T>2**	**E+, N+, T≤2**	**E+, N+, T>2**	**E-, N-, T>2**	**E-, N+, T≤2**	**E-, N+, T>2**
A	89.9	79.9	74.9	66.0	62.8	54.3	41.6	37.5
C	90.1	80.3	75.3	66.6	63.4	55.0	42.4	38.3
F	89.7	79.5	74.8	65.4	62.1	53.6	40.8	36.7
G	93.2	86.3	82.7	76.1	73.6	66.9	56.1	52.4
L	92.0	83.8	79.6	72.1	69.3	61.8	50.1	46.1
N	93.6	86.9	83.5	77.1	74.8	68.3	57.8	54.2
S	94.0	87.8	84.6	78.6	76.4	70.3	60.2	56.7
T	91.1	82.2	77.7	69.6	66.6	58.7	46.5	42.5
Y	92.5	84.8	80.9	73.7	71.1	63.9	52.6	48.7

Table 5.20a: For the all cases model: 5-year % survival estimates by Health Board for the eight groups of the clinical factors. Note that E, N and T stand for ER status, node status and tumour size respectively. Also, note that the Health Boards: Borders, Highland, Islands and Forth Valley are not presented here as they were not included in the complete cases model, although they were included in fitting the all cases model.

Complete Cases Model: 5-year % survival for age group 50-64 and clinical stage II								
Health Board	**E+, N-, T≤2**	**E-, N-, T≤2**	**E+, N-, T>2**	**E+, N+, T≤2**	**E+, N+, T>2**	**E-, N-, T>2**	**E-, N+, T≤2**	**E-, N+, T>2**
A	93.7	82.1	84.2	73.6	↑ 72.9	59.3	39.3	38.2
C	86.2	↓ 63.7	67.6	↓ 49.6	↓ 48.6	↓ 30.4	↓ 11.9	↓ 11.2
F	86.4	↓ 64.2	68.0	↓ 50.1	↓ 49.1	↓ 30.9	↓ 12.2	↓ 11.5
G	94.8	84.9	86.7	77.5	76.9	64.8	↓ 46.0	45.0
L	94.9	85.3	87.1	78.0	77.5	65.6	47.0	46.0
N	93.9	82.7	84.8	74.4	73.7	60.5	↓ 40.6	↓ 39.6
S	95.4	86.7	88.4	80.2	79.6	68.7	51.0	50.0
T	90.4	73.5	76.6	62.0	61.1	↓ 44.4	↓ 23.3	↓ 22.4
Y	99.0	↑ 97.1	↑ 97.5	↑ 95.6	↑ 95.4	↑ 92.6	↑ 87.1	↑ 86.7

Table 5.20b: For the complete cases model: 5-year % survival estimates by Health Board for the eight groups of the clinical factors. Note that E, N and T stand for ER status, node status and tumour size respectively. Also, note that the Health Boards: Borders, Highland, Islands and Forth Valley are not presented as they were not included in this model. The ↑ and ↓ represent an increase and decrease, respectively, of absolute magnitude greater than 10% when compared with the corresponding cells of Table 5.20a.

(2) For the best prognostic group (E+, N-, T≤2) there were very few differences between the two models. Both models predicted high 5-year survival estimates for all Health Boards, with range 89.7% to 94.0% for the all cases model and 86.2% to 99.0% for the complete cases model. This was the only prognostic group where the estimate

for Dumfries & Galloway (Y) Health Board was not at least 10% greater in absolute magnitude for the CCM than it was for the ACM (as it was not feasible since the 5-year % survival estimate for Y for the ACM was 92.5%).

For the intermediate prognostic groups (E+, N+, T$\leq$2 and E+, N+, T>2), there were several substantial differences between the results obtained for the ACM and for the CCM. In Health Boards C and F (Argyll & Clyde and Fife), both groups had a much lower 5-year survival estimate for the complete cases model, with estimates that were nearly 15% lower, than those for these Health Boards in the all cases model. However, the estimate for Health Board Y was increased by about 20% for both prognostic groups when the CCM was compared to the ACM. Also, for E+, N+, T>2, Ayrshire & Arran (A) Health Board had a 10% higher survival estimate in the CCM than it had in the ACM.

In the poorest prognostic group (E-, N+, T>2), five out of the nine Health Boards had absolute differences of more than 10% between the two models. Health Board Y was again at least 10% higher for the CCM than the ACM (in fact, the estimates were 86.7% and 48.7% respectively). The other four changes were decreases of more than 10% in the Health Boards C, F, N (Grampian) and T (Tayside). These drops were all about 20% in size when the CCM was compared to the ACM.

(3) The estimates for the CCM appeared to be more extreme than the ACM estimates. For those Health Boards with either the better or the poorer survival figures in the all cases model, the complete cases model seemed to emphasise them. This finding is similar to the previous discussion that was given after the hazard ratios for the two models were compared in Table 5.16.

PARTIAL-COMPLETE CASES ANALYSIS

Suppose rather than limiting the cases to those where there was complete information in all four variables, the restriction was changed to (i) only being complete for node status and tumour size and (ii) being complete in the three pathological factors; namely: ER status, node status and tumour size.

The numbers remaining in each Health Board in these two situations are given below in Table 5.21.

The percentages relate to the numbers of cases remaining for each Health Board from all of the 1619 surgical cases. The differences between the percentages kept in each Health Board in the two columns were remarkable. The most notable was that when only node status and tumour size needed to be known, Borders Health Board (B) actually kept the greatest percentage of cases, with Highland Health Board (H) keeping the third highest percentage. This was in stark contrast to the percentages remaining in these Health Boards (0 and 2.5% respectively) when ER status also had be known in order to be kept in the analysis.

	N and T complete		E, N and T complete	
Health Board	**Number**	**Percentage**	**Number**	**Percentage**
A	65	51.6	57	45.2
B	16	72.7	0	0.0
C	66	61.7	45	42.1
F	55	60.4	42	46.2
G	225	65.6	188	54.8
H	51	70.8	2	2.8
I	14	56.0	1	4.0
L	70	51.9	36	26.7
N	123	66.1	95	51.1
S	167	71.1	156	66.4
T	64	43.2	26	17.6
V	31	45.6	1	1.5
Y	17	27.9	10	16.4
Total	964	59.5	659	40.7

Table 5.21: Number of cases and percentages in each Health board when only node status and tumour size, and when all three pathological factors, were complete. Note that E, N and T stand for ER status, node status and tumour size respectively.

Due to these differences, the four Health Boards which had to be dropped from the complete cases analysis (HBs: B, H, I and V) were kept in the first extra analysis but had to be dropped in the second extra analysis (thus losing four cases). Cox models were fitted to the 964 and 655 cases respectively.

For model (i), age was not statistically significant and for model (ii), neither age nor clinical stage were significant. These non-significant factors were forced in for consistency (as for the CCM) and comparisons were made with the all cases and complete cases models. When the results from these analyses were examined, it was found that the hazard ratios for models (i) and (ii) were only slightly different from those from the ACM and CCM (see Table A6.3 in Appendix 6).

5.4.3 RELATIONSHIP OF AGE WITH MISSING VALUES IN OTHER COVARIATES

INTRODUCTION

In the last section, it was observed that age was a significant factor in the all cases model (ACM), but not in the complete cases model (CCM). This section tries to identify possible reasons for this difference. One obvious explanation could be a lack of power in the CCM to detect a relationship between age and survival, as the same level of significance (5%) was used as the cut-off in both situations. Another possible reason could be that age was related to missingness of data in the other variables. This is investigated here.

RESULTS

Firstly, when the ACM was fitted, the Wald statistic for age was significant, both univariately and multivariately. However, it was not significant in either the multivariate CCM (P=0.30), or when the factor was fitted univariately (P=0.69) based on only the complete cases. The fact that there appeared to be no differences among the survival curves for the four levels of age univariately for the complete cases, but there were differences when all cases were included in a univariate analysis, supports the idea that there was some sort of association between age group, missing values in the other variables and outcome. Table A6.2 in Appendix 6 gives the Kaplan-Meier estimates at five years based on all cases and complete cases only.

To assess this in a simple manner, various tabulations were examined. Firstly, the simple distributions of age group for the complete and incomplete cases (the cases with at least one of the four variables missing) are given in Table 5.22 below.

	Complete		Incomplete		All	
Age Group	**Number**	**(%)**	**Number**	**(%)**	**Number**	**(%)**
<50 years	184	(31.8)	292	(28.0)	476	(29.4)
50 - 64	245	(42.4)	346	(33.2)	591	(36.5)
65 - 79	145	(25.1)	335	(32.2)	480	(29.6)
≥80 years	4	(0.7)	68	(6.5)	72	(4.4)
Total	578	(100.0)	1041	(100.0)	1619	

Table 5.22: Distributions of age group for the complete, the incomplete and all cases.

To demonstrate the difference between the two distributions more clearly, the levels of age group were merged into two groups representing under 65 and aged 65 and over (Table 5.23).

	Complete		Incomplete		All	
Age Group	**Number**	**(%)**	**Number**	**(%)**	**Number**	**(%)**
<65 years	429	(74.2)	638	(61.3)	1067	(65.9)
≥65 years	149	(25.8)	403	(38.7)	552	(34.1)
Total	578	(100.0)	1041	(100.0)	1619	

Table 5.23: Distributions of cases aged under 65 and 65+ for the complete, the incomplete and all cases.

There were big differences in the percentages in the two age groups between the complete cases, the women with incomplete cases and all cases (comprising the two groups of women). For example, only 25.8% of the 578 complete cases were aged 65 or over, compared with 38.7% of the 1041 cases with some incomplete information.

The next simple tabulation presented examines whether there was any relationship between the number of variables with missing information (out of the four clinical variables with missing information discussed in detail in Section 4.2) and age in two groups, along with the crude indicator of percentage dead at 31/12/1993. The number of variables with missing data were grouped into none (containing 578 women), one or two variables with missing data (896 cases) and three or all four variables with missing information (145 cases). Table 5.24 below gives the breakdown of observed numbers in these three groups by under or over 65 years with the expected number under a null hypothesis of independence in each cell along with the observed percentage dead by 31/12/1993.

	Complete		**1 or 2 vars missing**		**3 or 4 vars missing**	
	Obs	**(Exp)**	**Obs**	**(Exp)**	**Obs**	**(Exp)**
Age < 65 years	429 Dead=32.9%	(380.9)	556 Dead=32.9%	(590.5)	82 Dead=34.1%	(95.6)
Age ≥65 years	149 Dead=36.2%	(197.1)	340 Dead=45.3%	(305.5)	63 Dead=44.4%	(49.4)
Total	578 Dead=33.7%		896 Dead=37.6%		145 Dead=38.6%	

Table 5.24: Observed (Obs) and expected (Exp) numbers of cases and the crude percentages dead in the different age groups for the groups of numbers of variables (vars) missing (out of the four clinical variables with missing information).

There appeared to be no effect in the <65 group on percentage dead for differing amounts of missing data in the other clinical variables. There was a difference, however, in the women aged 65 and over group. The crude percentages dead were 36.2% for complete cases compared with approximately 45% of cases with some missing information. Therefore, there was a different relationship between having missing data and outcome by age.

CONCLUSIONS

Data were more likely to be missing for the women aged ≥65 and having any missing information for these older women was associated with poorer outcome. Thus, when the unknowns were included in the survival analysis, age affected the outcome, but it did not when the cohort was limited to those with complete information only. Whilst there will still be lack of power in the CCM to detect age, it is probable that the effect of age on survival in the ACM was partly due to a relationship between age and the presence of missing values in other covariates.

5.4.4 RELATIONSHIP OF CLINICAL STAGE WITH MISSING VALUES

INTRODUCTION

In Section 5.4.2, both age and clinical stage were found to be significant in the all cases model (ACM), but non-significant in the complete cases model (CCM). An exposition for the differing findings for clinical stage is given here.

One possible reason why clinical stage was not significant in the CCM could be lack of power. Another suggested cause could be an association of clinical stage with missing data in the other clinical variables and the influence on outcome.

RESULTS

Univariately, clinical stage was significant (P=0.0001) in the complete cases analysis, but became non-significant in the presence of other variables, with P value 0.14 in the multivariate CCM. This differs from the finding for modelling age in the complete cases situation, discussed in the last section. The P value of the Wald statistic for the presence of clinical stage with other factors in the ACM was <0.001.

For the complete cases, the fact that clinical stage was significant univariately, but not when other variables were included in the model, suggests that there must have been confounding in the multivariate CCM. To examine this and try to identify which variables were associated with it, a modified forward selection analysis was performed. (A simple stepwise selection was of no use because clinical stage never entered the model in the complete cases analysis.) The forward selection method was stopped as soon as clinical stage became non-significant upon the addition of a variable.

The analysis was based on the 574 cases with at least 10 cases remaining in each of the Health Boards (full details given in Section 5.4.2). The entering variable was selected on the basis of the Wald statistic for being in the model in the presence of other variables with the smallest P value for forced entry with clinical stage, and any other variables in the model. Assessment of the significance of clinical stage was also made on the basis of the P value for the Wald statistic. Table 5.25 below reports the findings of this forward selection. Step 1 gives only the P value for the Wald statistic for clinical stage as none of the other variables were offered to the model. The P values in Step 2 are those for the Wald statistic for forced entry with clinical stage, with each of the variables fitted separately in models with clinical stage. The P value given for clinical stage is that which was obtained in the model for the variable which was chosen for entry at Step 2. Similarly in Step 3, P values are for forced entry with clinical stage and the variable selected in Step 2 for all variables.

	Step 1	Step 2	Step 3
C	e 0.0003	a 0.0015	a 0.1110
A	*	0.5357	0.8360
E	*	e <0.0001	a <0.0001
N	*	<0.0001	e <0.0001
T	*	0.0004	0.0075
H	*	0.0230	0.0013

Table 5.25: Presentation of results from performing a forward selection on the variables with clinical stage. Note that 'e' indicates which variable entered the model at that step and that 'a' indicates that the variable has already been entered in the model. The P value for the variables already in the model are those which are obtained from the model for the new entering variable. The P values represent forced entry for all of the variables. Clinical stage is given by C, age (A), ER status (E), node status (N), tumour size (T) and Health Board (H) respectively.

Thus at Step 2, ER status was fitted with clinical stage. The significance of clinical stage did not alter upon addition of this variable. However, when either node status or tumour size were forced in with just clinical stage, the significance of clinical stage was greatly affected (data not given in Table 5.25). The P values for clinical stage with node status and tumour size were 0.033 and 0.039 respectively. This suggested that these two variables were associated with clinical stage. This was observed in Section 4.1 (and Appendix 4), where it was shown that these variables were not independent.

At Step 3, node status was fitted into a model including clinical stage and ER status. Clinical stage became non-significant (P=0.11), suggesting that the addition of node status in the presence of ER status caused clinical stage to lose its significance at the 5% level in the CCM.

The percentages of women who were dead by 31/12/1993 in the two age groups are given split by extent of missing data (Table 5.26).

	Complete	1 var missing	2 or 3 vars missing
	Obs (Exp)	Obs (Exp)	Obs (Exp)
Stage I	102 (123) Dead=18.6%	116 (111) Dead=22.4%	84 (68) Dead=27.4%
Stage II	394 (331) Dead=35.0%	274 (298) Dead=33.2%	145 (184) Dead=42.8%
Stage III	82 (76) Dead=46.3%	75 (69) Dead=61.3%	30 (42) Dead=60.0%
Unknown	81 (129) Dead=38.3%	129 (116) Dead=43.4%	107 (72) Dead=37.4%
Total	659 Dead=34.3%	594 Dead=36.9%	366 Dead=39.1%

Table 5.26: Observed (Obs) and expected (Exp) numbers of cases and the crude percentages dead in the different clinical stage groups for the groups of numbers of variables (vars) missing (out of the other three clinical variables with missing information).

Thus, for women with stage I disease, there was a progressive increase in death risk as the amount of missing information in the other three variables increased. For stage II disease, it appeared that having two or three of the other three variables unknown was a lot worse in terms of outcome than having either none or only one other variable with missing data. Stage III disease appeared to have much higher risk of death if any of the other variables were missing. No definite pattern was observed for women with unknown clinical stage.

DISCUSSION

In the complete cases, clinical stage became non-significant in the model started with only that factor in it, once ER status and node status had also been entered into the model. This makes some clinical sense because one element of clinical stage is *clinical* node status. It is expected, therefore, that clinical node status would agree reasonably with pathological node status (the node status available for analysis here). Thus, clinical stage would probably be expected to be partly redundant when pathological node status was determined and included in the analysis.

A table of pathological node status vs clinical node status was examined for the complete cases (Table 5.27) and revealed that 36.7% $\left(\frac{212}{578}\right)$ of the cases were not classified in the same way. However, it was not important to calculate the sensitivity, specificity or positive predictive values here because the majority of cases had the same code.

		Clinical		
		Positive	Negative	Total
Pathological	Positive	136	151	287
	Negative	61	230	291
	Total	197	381	578

Table 5.27: Numbers of cases in the groups with clinical and pathological node status, either positive or negative for the complete cases.

One possible reason that clinical stage was not in the complete cases model could be that there was enough of an overlap between the known pathological node status and the known clinical node status element of clinical stage to make clinical stage unnecessary. Similarly, another element of clinical stage is clinical tumour size, which would be expected to be similar to the pathological tumour size recorded, thus explaining why the introduction of known pathological tumour size in the model with known clinical stage appeared to affect the significance of clinical stage (as discussed in the paragraph after Table 5.25).

In contrast, in the all cases model, clinical stage was necessary in the multivariate Cox model even with these other variables in it. It, therefore, appeared that the introduction of the extra categories for the unknowns and inclusion of the cases with missing values in other variables, as well as clinical stage, allowed the variable for clinical stage to enter the model. Thus, clinical stage also appeared to be linked to the amount of missing data and outcome. However, it could also be due to the fact that the clinical stage variable is not the same in the two models, as the factor has an extra degree of freedom in the ACM.

The fact that clinical stage was significant on its own in the complete cases situation but it was not necessary in the presence of other variables with known factor levels (in particular ER status and node status), perhaps suggests that clinical stage was in the all cases model because when it had known clinical stage values, it is acting as a surrogate for the missing information in the other prognostic factors. For example, there were 435 cases with unknown pathological node status and only 317 cases with missing clinical stage in the ACM. In Section 4.2.2, it was shown that clinical stage and pathological node status were independent in the log-linear model fitted relating missing values in the clinical variables. Thus, it would be expected that some cases with pathological node status missing would have clinical stage known, thus providing an indication of the extent of disease for these cases. This reason could explain the presence of clinical stage in the ACM and the absence of it in the CCM, where the information about the other prognostic factors is obviously known.

This argument is supported by the fact that in the partial complete cases analysis, also described in Section 5.4.2, when only node status and tumour size had to be known, but ER status could be missing, clinical stage was necessary in the model (model (i)). However, once the analysis was limited to only those cases where all three pathological factors were complete, clinical stage was again no longer significant in the Cox model (model (ii)).

However, the reason for the absence of clinical stage in the CCM could just be due to the lack of power to detect it. This surmise is based on the fact that the P value for this factor was only marginally non-significant at 0.14 for non-entry.

5.4.5 GENERAL DISCUSSION

The results given in Sections 5.2.1 and 5.2.2 and by Twelves et al (1998a) support the findings of other relevant studies of breast cancer survival (Section 5.2.3), especially in relation to the need for management of this disease to be given in the setting of the multidisciplinary team approach.

When the 'Clinical Full' model was examined to assess the adequacy of the fit of the model (Section 5.3.1) and the assumption of proportional hazards checked (Section 5.3.2), it was found that there was a suggestion of non-proportionality for two of the levels of ER status. However, it was noted in Section 5.3.3 that there is no unique approach to assessing proportional hazards using time-dependent modelling for non-binary categorical factors.

In Section 5.4.2, the results from fitting the all cases and complete cases model were compared. One of the main problems with this approach is that it was not possible to test statistically whether any of the apparent differences were real on the basis of any known tests. All of the observations noted above about differences among the four analyses were informal.

It appears that the missing values, added as extra categories, caused some large absolute differences in the point estimates. These might lead to different interpretations of the importance of Health Board of treatment on survival, and indeed whether there were any true differences. However, it was consistently shown, by examining the different Health Board, prognostic groups (Tables 5.20a and 5.20b, Section 5.4.2), that women treated in some of the Health Boards had poorer outcomes than women treated in other Health Boards.

It is not clear whether using the complete cases technique for dealing with the missing values would have been more appropriate for the analysis of the Breast Cancer Audit data, although losing 64% of the cases appears to be wasting a great deal of information on other variables. Also, whether the data were missing at random cannot be tested and so it is unknown whether this was a valid implicit assumption to have made. The extent of any biases in the estimates for both the ACM and the CCM cannot be obtained.

On the basis of these two models, very different conclusions could be drawn in terms of differences in absolute magnitudes of survival for different Health Boards. These would perhaps then have different implications in terms of political and organisational structures of provision of services for breast cancer management in Scotland in the future.

CHAPTER 6 INVESTIGATIONS OF BIAS IN MODELS WITH ADDITIONAL CATEGORIES FOR MISSING VALUES

6.1 INTRODUCTION TO THE ABSTRACT PROBLEMS AND SOME GENERAL THEORY

INTRODUCTION

One method of handling missing values for categorical factors in proportional hazards models is by creating additional categories to represent the unknown levels. In Section 5.2.2, the results of fitting a Cox regression model to the Breast Cancer Audit data using this method were reported. The assumption of proportional hazards for the chosen model was investigated in Section 5.3.2.

It is not clear whether, in general, the assumption of proportional hazards for contexts involving these extra levels is consistent with the same assumption for designs with complete data (i.e. without these additional levels). This is the focus of this chapter. To avoid the complexities of the Cox regression model, exponential regression modelling is performed for the majority of the analyses.

THE ABSTRACT PROBLEMS

The exponential regression model is a very simple model with the proportional hazards property. To make the situation as uncomplicated as possible, the exponential regression model is assumed to have either one or two factors which have only two levels to represent the known values and an additional level to represent the missing values in each of the factors. Although the outcome is known for the missing values, the true levels of the factors are unknown. That is, the observations would have been classified as level 1 or 2 for the factors, had this information been available. As a further simplification, the problem of censored data is ignored and the context where all subjects are followed until their event time is considered.

Two different situations are examined. Firstly, in Section 6.2, the theoretical situation is explored where the observations falling in the third levels consist of random mixtures of two (or more) exponential distributions across the known levels. In Section 6.3, however, simulation models investigate the effects on bias of making the naive assumption that the observations falling in the third levels also have exponential distributions.

The aim of the initial theoretical exercise is to investigate whether or not the assumption of proportionality holds when the missing values are included as extra levels in a model which has proportional hazards for the levels for the known values.

GENERAL DISTRIBUTIONAL THEORY

The probability density function (pdf) for an exponential regression model is given by

$$f(y) = \exp\left(\underline{\beta}^T \underline{x}\right)\exp\left[-y\exp\left(\underline{\beta}^T \underline{x}\right)\right] \quad \text{(Eq 6.1\_1)}$$

and, letting the term $\exp\left(\underline{\beta}^T \underline{x}\right)$ be replaced by λ, it can easily be shown that the hazard function is given by

$$h(y) = \lambda = \exp\left(\underline{\beta}^T \underline{x}\right), \quad \text{(Eq 6.1\_2)}$$

which is constant for all values of y.

6.2 EXPONENTIAL REGRESSION MODEL WITH FACTOR(S) EXTENDED FROM TWO LEVELS TO THREE LEVELS BY ASSUMPTION THAT THIRD LEVEL IS RANDOM MIXTURE OF FIRST TWO LEVELS

In Section 6.2.1, a factor at two levels is examined, with outcomes assumed to satisfy an exponential regression model. A third level for missing values in this factor is created on the basis that the observations arise from a random mixture of the first two levels. The aim is to derive the hazard function for this third level. It is then of interest to assess whether this hazard function is proportional to the hazard functions for the observations in the first and second levels. The effects of changing the mixing parameter for weighting the pdfs of the two levels, and of changing the ratio of the hazards between the first two levels, are examined graphically.

The theory for two factors, both originally at two levels, is then examined in Section 6.2.2. The missing values are incorporated into the two factors as additional levels of the factors and are assumed to be random mixtures of the first two levels for both factors.

6.2.1 THE ONE FACTOR SITUATION

DERIVATION OF THE HAZARD FUNCTION

An exponential regression model for a single factor with two levels has pdf for the ith level, from Eq 6.1_1 in Section 6.1, given by

$$f_i(y) = \exp(\alpha_i)\exp\left[-y\exp(\alpha_i)\right] \quad \text{for } i = 1,\, 2\,.$$

Suppose the factor is then extended to three levels to incorporate missing values, where the observations in this third level are assumed to be a random mixture of data from the original two levels. This third level has a pdf, $f_3(y)$, which is a mixture of two exponential density functions, and is

$$f_3(y) = zf_1(y) + (1-z)f_2(y).$$

The hazard functions for the first two levels are

$$h_1(y) = \exp(\alpha_1) = \lambda_1 \text{ and } h_2(y) = \exp(\alpha_2) = \lambda_2. \qquad \text{(Eqs 6.2.1\_1)}$$

The hazard function for the third level can then be shown to be equal to

$$\begin{aligned} h_3(y) = \frac{f_3(y)}{S_3(y)} &= \frac{z\lambda_1 \exp(-\lambda_1 y) + (1-z)\lambda_2 \exp(-\lambda_2 y)}{1 - z[1-\exp(-\lambda_1 y)] - (1-z)[1-\exp(-\lambda_2 y)]} \\ &= \frac{z\lambda_1 \exp(-\lambda_1 y) + (1-z)\lambda_2 \exp(-\lambda_2 y)}{z\exp(-\lambda_1 y) + (1-z)\exp(-\lambda_2 y)} \end{aligned} \qquad \text{(Eq 6.2.1\_2)}$$

Thus, with the exception of the trivial cases $z = 0$ or 1, this is not proportional to either $h_1(y)$ or $h_2(y)$, as it is not constant for all values of y. This lack of proportionality can be illustrated graphically.

GRAPHICAL REPRESENTATION

An arbitrary value was chosen for the hazard function of the first group; namely

$$h_1(y) = \lambda_1 = 0.25.$$

To choose a sensible range for time, values were selected to represent time from zero to the 95th percentile for the exponential distribution of the first level. The range for y was taken to be $[0, y_{max}]$. Therefore,

$$\text{Prob}(Y \le y_{max}) = 1 - \exp(-0.25 y_{max}) = 0.95,$$

which implies that

$$y_{max} = 11.98 \approx 12.$$

Twenty-five equally-spaced points for y between 0 and 12 (i.e., at 0, 0.5, 1, ..., 11, 11.5, 12) were used.

For simplicity, the hazard ratio for level 3 vs level 1 was considered. This is obtained from Eqs 6.2.1_1 and 6.2.1_2 and is

$$hr_{3:1} = \frac{h_3(y)}{h_1(y)} = \frac{z\exp(-\lambda_1 y) + \gamma(1-z)\exp(-\gamma\lambda_1 y)}{z\exp(-\lambda_1 y) + (1-z)\exp(-\gamma\lambda_1 y)},$$

with λ_1 fixed at 0.25, λ_2 replaced by $\gamma\lambda_1$, and $z \in [0,\ 1]$ and $\gamma \in [0,\ \infty]$ being varying quantities.

The values selected to represent the mixing weight (z) of the two pdfs were 0.2, 0.4, 0.7 and 0.9. Values 0.1, 0.2, 0.3, 0.4, 0.5, 0.8, 1, 1.5, 2, 2.5, 4, 5 and 10 were chosen for γ, the value of the hazard ratio for level 2 vs level 1. The logarithm of $hr_{3:1}$ was then plotted for each of the combinations of z and γ. To illustrate the wide variation caused by changing the values of z and γ, six of the 48 possible curves were picked out and are shown in Figure 6.1 below.

RESULTS

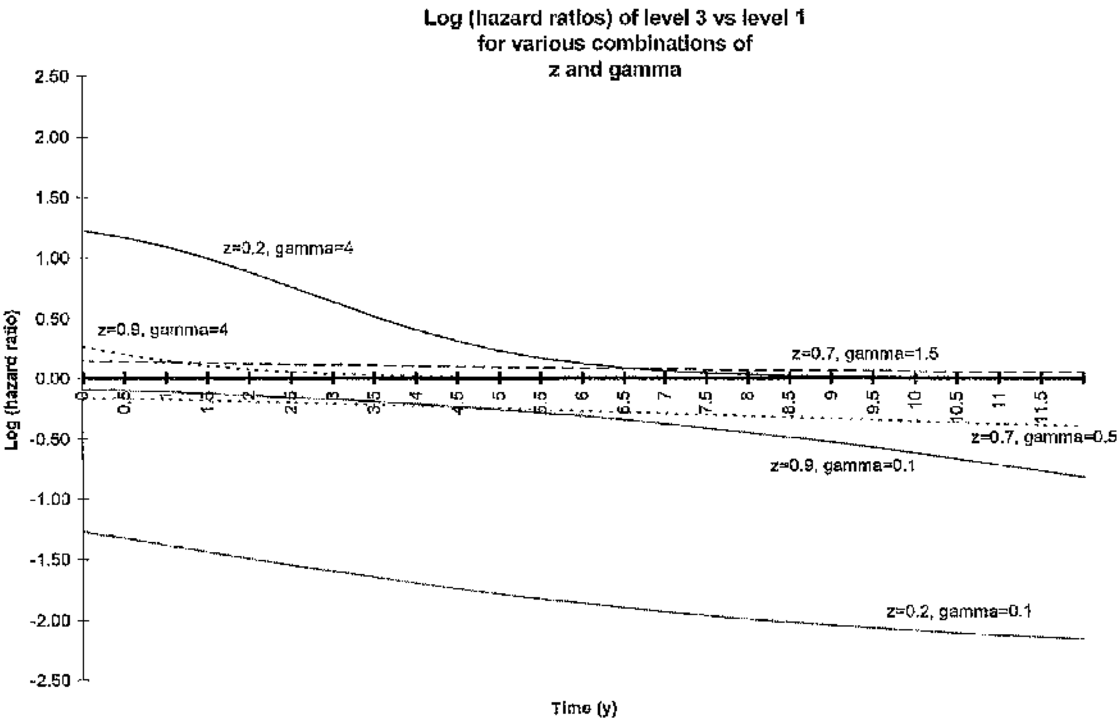


Figure 6.1: Chart showing the log of the hazard ratio for level 3 vs level 1 for six different combinations of z and γ.

When the scale factor, γ , is equal to one, the hazard ratio $hr_{3:1}$ is constant at 1, and is represented by the horizontal line through zero. When γ is greater than 1, both level 2 and level 3 have an increased risk relative to level 1, as the curves for $\gamma>1$ remain above zero on the log scale in Figure 6.1. Similarly, the curves with $\gamma<1$ remain below zero indicating that $\gamma<1$ leads to a reduced risk for both levels 2 and 3 compared to level 1.

Figure 6.1 has illustrated graphically the fact that when a third level was assumed to be a random mixture of the first two levels which do fit an exponential regression model, the hazard for this third level was not proportional to the hazards for these first two levels and that the non-proportionality could be quite considerable.

6.2.2 THE TWO FACTORS SITUATION

THE PROBLEM

The design is now extended to include two factors. Again, the aim is to derive the hazard functions for the missing categories and to check whether or not these hazard functions are proportional to the hazard functions for the known levels.

DESIGN WITHOUT MISSING VALUES

Suppose there are two factors at two levels with observations arising from a main effects exponential regression model, with y as the dependent variable. Let n_{kl} denote the number of observations in cell(k,l), where k and l represent the levels of factors F1 and F2 respectively. Figure 6.2 is a diagrammatical representation of the basic design, where there are no missing values.

		F2	
		1	**2**
F1	**1**	n_{11}	n_{12}
	2	n_{21}	n_{22}

Figure 6.2: Diagram to represent the design without missing values in the two factors.

The pdf for the n_{kl} observations falling in cell (k,l) is given by

$$f_{kl}(y) = \lambda_{kl}\exp(-\lambda_{kl}y), \quad \text{(Eq 6.2.2\_1)}$$

for $k = 1,\ 2$ and $l = 1,\ 2$, where

$$\lambda_{kl} = \exp[\mu_{11} + \alpha_k + \beta_l] \qquad \text{(Eq 6.2.2\_2)}$$

and the constraints $\alpha_1 = 0$ and $\beta_1 = 0$ are imposed.

The α_k parameters are main effects related to factor F1 and the β_l parameters are main effects related to factor F2.

The hazards for these four cells are constant and, therefore, proportional. The hazard function for cells $(1,2)$, $(2,1)$ and $(2,2)$ can be written, respectively, as

$$\lambda_{12} = \exp(\beta_2)\lambda_{11} = \gamma_1\lambda_{11}$$

$$\lambda_{21} = \gamma_2\lambda_{11}$$

$$\lambda_{22} = \gamma_1\gamma_2\lambda_{11}, \qquad \text{(Eqs 6.2.2\_3)}$$

where λ_{11} is the hazard function for cell $(1,1)$.

DATA WITH MISSING VALUES

Now suppose that the cases with missing values for F1 and/or F2 are to be included. Let the mixing parameter for factor F1 be p for level 1 and $(1-p)$ for level 2. Similarly, let the mixing parameter for factor F2 be q for level 1 and $(1-q)$ for level 2. It is assumed that the mixing operates independently in the rows and columns. Figure 6.3 shows the structure of the model with missing values.

		F2		
		1 q	**2** $1-q$	**3**
F1	**1** p	n_{11}	n_{12}	n_{13}
	2 $1-p$	n_{21}	n_{22}	n_{23}
	3	n_{31}	n_{32}	n_{33}

Figure 6.3: Diagram to represent the design for the two factors when missing values are included.

The pdfs for the observations falling in the four cells $(1,1)$, $(1,2)$, $(2,1)$ and $(2,2)$ are given by Eqs 6.2.2_1 and 6.2.2_2 as before. Using the mixing parameters, p and q, from above, the pdfs for the observations in the missing categories can be written as mixtures of two (or four) exponential density functions, as follows. The pdf for the n_{13} observations in cell $(1,3)$ is given by

$$f_{13}(y) = qf_{11}(y) + (1-q)f_{12}(y).$$

The pdfs for cells $(2,3)$, $(3,1)$ and $(3,2)$ can be written down in a similar manner, using the appropriate combination of the mixing parameters.

The pdf for cell $(3,3)$ is assumed to be

$$f_{33}(y) = pqf_{11}(y) + p(1-q)f_{12}(y) + q(1-p)f_{21}(y) + (1-p)(1-q)f_{22}(y)$$

Concentrating on the pdf for cell $(1,3)$, the hazard function for the observations falling in this cell can be calculated using

$$h_{13}(y) = \frac{f_{13}(y)}{S_{13}(y)} = \frac{q\lambda_{11}\exp(-\lambda_{11}y) + (1-q)\lambda_{12}\exp(-\lambda_{12}y)}{q\exp(-\lambda_{11}y) + (1-q)\exp(-\lambda_{12}y)}.$$

Then, by substitution using Eqs 6.2.2_3, the hazard function for cell $(1,3)$ can be rewritten as

$$h_{13}(y) = \frac{q\lambda_{11}\exp(-\lambda_{11}y) + (1-q)\gamma_1\lambda_{11}\exp(-\gamma_1\lambda_{11}y)}{q\exp(-\lambda_{11}y) + (1-q)\exp(-\gamma_1\lambda_{11}y)}.$$

Similarly, the hazard functions for the cells $(2,3)$, $(3,1)$ and $(3,2)$ can be derived.

The hazard function for cell $(3,3)$ can be shown to be equal to

$$h_{33}(y) = \frac{A}{B},$$

where

$$A = pq\lambda_{11}\exp(-\lambda_{11}y) + p(1-q)\gamma_1\lambda_{11}\exp(-\gamma_1\lambda_{11}y) + q(1-p)\gamma_2\lambda_{11}\exp(-\gamma_2\lambda_{11}y) \\ + (1-p)(1-q)\gamma_1\gamma_2\lambda_{11}\exp(-\gamma_1\gamma_2\lambda_{11}y)$$

and

$$B = pq\exp(-\lambda_{11}y) + p(1-q)\exp(-\gamma_1\lambda_{11}y) + q(1-p)\exp(-\gamma_2\lambda_{11}y) \\ + (1-p)(1-q)\exp(-\gamma_1\gamma_2\lambda_{11}y)$$

None of the hazard functions for the missing cells are constant for all values of y. Therefore, they cannot be proportional to any of the hazards defined for the known values falling in cells $(1,1)$ to $(2,2)$, given by Eq 6.2.2_2. Therefore, an additive exponential regression model would not fit satisfactorily for the additional levels created for the missing values in the factors F1 and F2 since the hazard functions for observations falling in the five cells are not proportional to the hazard functions for the observations with known levels for both F1 and F2.

DISCUSSION

The third level in the one factor exponential regression model did not satisfy the proportional hazards assumption and similar findings were observed in the two factors situation. It therefore seems reasonable to assume that the argument would carry through to the more complex Cox's proportional hazards model. The implications of this non-proportionality of the hazards are explored in the remainder of this chapter.

6.3 EXPONENTIAL REGRESSION MODEL WITH FACTOR(S) EXTENDED FROM TWO LEVELS TO THREE LEVELS BY NAIVE ASSUMPTION THAT THIRD LEVEL IS ALSO EXPONENTIAL

Having just shown that the assumption of proportionality of hazards for missing data categories can be inconsistent with the true proportional hazards assumption for the complete data context, the aim now is to investigate whether or not this invalidation of the assumption matters in practice. Exponential distributions are used to generate the complete cases data in such a manner that these cases satisfy a main effects additive exponential regression model. Here, the observations falling in the extra levels for the missing values are also naively assumed to have exponential distributions, although these observations were in reality generated from mixtures of two or four exponential distributions. Simulations were carried out to investigate whether it is a problem, in terms of parameter estimate bias, that an incorrect model is fitted and the assumption of proportional hazards violated.

6.3.1 SIMPLE THEORY FOR THE ONE FACTOR SITUATION

INTRODUCTION

Here, an exponential regression model on one factor is considered. Missing values are incorporated by creating an additional level which are obtained to be a random mixture of observations from the first two levels. The effect of taking the outcomes for the missing data category to naively be assumed to have an exponential distribution is investigated. The main interest is whether or not the inclusion of the extra level affects the parameter estimates for the first two levels.

SOME SIMPLE THEORY

Suppose that the observations in the ith level of the factor are y_{ib}, with $i = 1,\ 2,\ 3$ and $b = k = 1,\ \ldots,\ n_1$ for observations in level 1; $b = l = 1,\ \ldots,\ n_2$ for observations in level 2; and $b = m = 1,\ \ldots,\ n_3$ for observations in level 3 respectively. The pdf is given by

$$f(y_{ib}) = \exp(\alpha_i)\exp[-y_{ib}\exp(\alpha_i)].$$

If $\exp(\alpha_i) = \lambda_i$, then the joint pdf for the three levels is given by

$$f\left(\underline{y}_{1k}, \underline{y}_{2l}, \underline{y}_{3m}; \lambda_1, \lambda_2, \lambda_3\right) = f\left(\underline{y}_{1k}; \lambda_1\right).f\left(\underline{y}_{2l}; \lambda_2\right).f\left(\underline{y}_{3m}; \lambda_3\right),$$

by the independence of the three samples. This can be re-written as

$$f\left(\underline{y}_{1k}, \underline{y}_{2l}, \underline{y}_{3m}; \lambda_1, \lambda_2, \lambda_3\right) = \left[\prod_{k=1}^{n_1} \lambda_1 \exp(-\lambda_1 y_{1k})\right].\left[\prod_{l=1}^{n_2} \lambda_2 \exp(-\lambda_2 y_{2l})\right].\left[\prod_{m=1}^{n_3} \lambda_3 \exp(-\lambda_3 y_{3m})\right]$$

since observations within samples are also independent.

Clearly, in this simple case, the maximum likelihood estimates for λ_1 and λ_2, and hence α_1 and α_2, are independent of the outcomes (λ_3 and α_3 respectively) for the missing data category. They are exactly the same as they would have been had the missing data category been ignored.

Thus in the one factor situation, the fact that the hazard function for level 3 was not proportional to the hazard functions for levels 1 and 2 (Section 6.2.1) does not influence the parameter estimates obtained for levels 1 and 2 when an exponential regression model is fitted to these data, when it is naively assumed to fit the third level also.

6.3.2 INTRODUCTION AND STRATEGY FOR DEVELOPING THE THEORETICAL DATASETS IN THE TWO FACTORS SITUATION

INTRODUCTION

The remainder of this chapter concentrates on the two factors situation. The main aim is to investigate whether the non-proportionality of the hazard functions, shown in Section 6.2.2, affects the estimates of the parameters in the model describing the observations falling in the known levels when exponential regression models are fitted with two factors for various designs. It is assumed that both factors have two levels of known values and an additional level created for missing values. The missing value outcomes are again assumed to be a random mixtures of outcomes for the two known levels. The exponential regression model fitted to the data is taken to be additive so that there is no interaction term present.

Figure 6.4 shows the design with two factors at three levels. The observations with unknown levels fall into one of the five cells $(1,3)$, $(2,3)$, $(3,1)$, $(3,2)$ and $(3,3)$.

		F2		
		1	**2**	**3**
	1	n_{11}	n_{12}	n_{13}
F1	**2**	n_{21}	n_{22}	n_{23}
	3	n_{31}	n_{32}	n_{33}

Figure 6.4: Diagram representing the design for two factors, both at three levels.

When fitting models to the data, all of the observations falling into the nine cells are assumed to be exponential and so the assumed pdfs for all nine cells can be denoted by

$$f\left(y_{klm_{kl}}\right) = \lambda_{kl} \exp\left(-\lambda_{kl} y_{klm_{kl}}\right), \qquad \text{(Eq 6.3.2\_1)}$$

where

$$\lambda_{kl} = \exp\left[\mu_{11} + \alpha_k + \beta_l\right] \qquad \text{(Eq 6.3.2\_2)}$$

with $k = 1,\ 2,\ 3$ and $l = 1,\ 2,\ 3$ and the constraints $\alpha_1 = 0$ and $\beta_1 = 0$ are imposed. These λ_{kl} are the hazard functions from the assumed model for the nine cells.

The aim here is to derive two sets of normal equations to try to obtain the parameter estimates for the known levels. Firstly, when only the observations with known factor levels (i.e. cells $(1,1)$, $(1,2)$, $(2,1)$ and $(2,2)$) and, secondly, when all of the observations (i.e. known plus missing) are included in the design. Exponential regression models are fitted to observations in these contexts and the results are compared.

When the normal equations are obtained, using standard maximum likelihood techniques, for both the situations with and without the missing data categories, the equations cannot be solved analytically (see Appendices 7 and 8 for complete and all cases designs respectively). Therefore, simulation methods are used to study the properties of parameter estimates in the presence of missing information in the factors.

To perform these exercises, artificial datasets needed to be generated such that observations for the four cells with known factor levels arose from an exponential regression model with additive contributions from both factors but with no interaction. The missing values categories were created to have observations from random mixtures of the four known exponential distributions. The potential bias arising when the third levels are incorrectly taken to satisfy the exponential regression model is examined. All calculations, including random number generation, were carried out in SAS.

STRATEGY FOR DATA GENERATION

For The Complete Cases Designs: Figure 6.5 shows the four cells representing the known factor levels. These four cells will be referred to as the known cells and the context will be referred to as the complete cases design.

		F2	
	1	**2**	**3**
F1 **1**	n_{11}	n_{12}	
2	n_{21}	n_{22}	
3			

Figure 6.5: The known cells.

The observations are generated from four exponential distributions such that n_{kl} represents the number of observations falling in cell (k,l), where $k = 1,\ 2$ and $l = 1,\ 2$ and Eqs 6.3.2_1 and 6.3.2_2 are satisfied. The α_k parameters are main effects related to factor F1 and the β_l parameters are main effects related to factor F2.

The respective hazard functions for the four cells are given by

$$\lambda_{11} = \exp(\mu_{11}) \text{ for cell } (1,1),$$
$$\lambda_{12} = \exp(\mu_{11} + \beta_2) \text{ for cell } (1,2),$$
$$\lambda_{21} = \exp(\mu_{11} + \alpha_2) \text{ for cell } (2,1)$$

and

$$\lambda_{22} = \exp(\mu_{11} + \alpha_2 + \beta_2) \text{ for cell } (2,2). \qquad \text{(Eqs 6.3.2\_3)}$$

For all of the analyses described in the remainder of the chapter, the true values chosen for the parameters were:

$$\mu_{11} = 1.25,$$
$$\alpha_2 = 0.5$$

and $\quad \beta_2 = 0.3$.

The value for μ_{11} was chosen arbitrarily. The values for α_2 and β_2 were chosen such that the hazard ratio for level 2 versus level 1 of F1 was $\exp(\alpha_2) \approx 1.6$ and the hazard ratio for level 2 versus level 1 of F2 was $\exp(\beta_2) \approx 1.3$. These are similar in size to hazard ratios observed in the analysis of the Breast Cancer Audit data.

Thus, observations in cell $(1,1)$ were generated from an $Ex(\exp(1.25))$ distribution and so the mean of observations in the cell would be $\exp(-1.25)$. Similarly, the exponential distributions for the observations in cells $(1,2)$, $(2,1)$ and $(2,2)$ were $Ex(\exp(1.55))$, $Ex(\exp(1.75))$ and $Ex(\exp(2.05))$ respectively.

The SAS procedure *Lifereg* was used to perform all of the exponential regression modelling. The purpose of fitting the models to the complete cases only was to test the SAS program. As all of the four known cells were observations from exponential distributions, then when the exponential regression model was fitted, the parameter estimates $\hat{\mu}_{11}$, $\hat{\alpha}_2$ and $\hat{\beta}_2$ ought to have been very close to the true values set up for μ_{11}, α_2 and β_2.

For The All Cases Designs: Figure 6.6 represents the five cells corresponding to the missing factor level information. These are known as the missing cells.

		F2		
		1	**2**	**3**
	1			n_{13}
F1	**2**			n_{23}
	3	n_{31}	n_{32}	n_{33}

Figure 6.6: The missing cells.

Each observation falling in a missing cell was generated from one of the four possible exponential distributions. A mechanism was needed to decide from which distribution the observation should be generated. The mixing parameters for each of the five cells were chosen here such that the proportions of observations generated from the different exponential distributions in the missing cells were the same as the relative frequencies of the known cells. For example, the n_{13} observations in cell $(1,3)$ consist of observations from a mixture of two exponential distributions with the true mixture pdf for the n_{13} observations given by

$$f_{13} = \frac{n_{11}}{n_{11}+n_{12}} f_{11} + \frac{n_{12}}{n_{11}+n_{12}} f_{12}$$

$$= r f_{11} + (1-r) f_{12}$$

Similarly, for cell $(3,2)$, the n_{32} observations were created from a mixture of two exponential distributions such that the true mixture pdf is

$$f_{32} = \frac{n_{12}}{n_{12}+n_{22}} f_{12} + \frac{n_{22}}{n_{12}+n_{22}} f_{22}$$

$$= u f_{12} + (1-u) f_{22}$$

For cell $(3,3)$, the n_{33} observations are generated from a mixture of four exponential distributions such that

$$f_{33} = \frac{n_{11}}{n_{11}+n_{12}+n_{21}+n_{22}} f_{11} + \frac{n_{12}}{n_{11}+n_{12}+n_{21}+n_{22}} f_{12} + \frac{n_{21}}{n_{11}+n_{12}+n_{21}+n_{22}} f_{21} + \frac{n_{22}}{n_{11}+n_{12}+n_{21}+n_{22}} f_{22}$$

Full details are given in Appendix 9. Note that these mixing parameters are different from those used in Section 6.2.2. Clearly this is just one particular structure that could have been chosen for generating observations in the five missing cells. This process of using the relative frequencies was selected because it meant that the distributions of responses in the missing categories are weighted with respect to the observed numbers of subjects in the individual complete data categories.

THE ASSUMED MODEL

All observations are assumed to satisfy the exponential regression model, which is known to be incorrect for the missing values. This is known because the observations falling in the missing cells have been generated to be random mixtures from the two known levels. In Section 6.2.2, it was shown that including third levels which have observations which are random mixtures of the exponential distribution for the first two levels produces hazard functions for the third levels which are not proportional to those for the first two levels. The primary aim is to investigate whether the parameter estimates $\hat{\alpha}_2$ and $\hat{\beta}_2$, obtained when the missing values are also included in the dataset, are different from the true values of the parameters.

If the fitted model were true, then the log hazards for the cells would be those given by Figure 6.7. The values of $\hat{\mu}_{11}$, $\hat{\alpha}_3$ and $\hat{\beta}_3$ are unimportant here as it is really the values of $\hat{\alpha}_2$ and $\hat{\beta}_2$ that are of interest.

		F2		
		1	2	3
	1	μ_{11}	$\mu_{11}+\beta_2$	$\mu_{11}+\beta_3$
F1	2	$\mu_{11}+\alpha_2$	$\mu_{11}+\alpha_2+\beta_2$	$\mu_{11}+\alpha_2+\beta_3$
	3	$\mu_{11}+\alpha_3$	$\mu_{11}+\alpha_3+\beta_2$	$\mu_{11}+\alpha_3+\beta_3$

Figure 6.7: Log hazards for the nine cells when the model is true.

Here the estimated hazard ratio for being in level 2 of F1 versus level 1 of F1 is given by $\exp(\hat{\alpha}_2)$ and similarly for being in level 2 of F2 versus level 1 of F2, the estimated hazard ratio is $\exp(\hat{\beta}_2)$. Thus, if the parameter estimates obtained for the complete and all cases designs are different, then the hazard ratios will also be different, possibly leading to different interpretations and conclusions.

6.3.3 RESULTS FOR THE COMPLETE CASES DESIGNS BASED ON SIMULATED DATA

INTRODUCTION TO THE NINE GROUPS

Nine groups of designs for the complete data cells were created in three types according to the different numbers in the known levels. The most obvious type was the one with equal numbers in the known cells. The next two groups formed the second type, where the proportions of observations in level 1 out of the total for factor F2 were created to be the same for both levels of factor F1. The remaining six groups fell into the third type, where there was no definite pattern among the numbers in the known cells. Some of these six groups were chosen to have some extreme variations in the numbers in the four known cells.

Table 6.1 gives the numbers n_{kl} falling in cell(k,l) for the complete cases (see Figure 6.5 in Section 6.3.2). Note that all of the numbers are in units of a 1000. Large sample sizes were used to obtain biases in an asymptotic situation.

Type and Group	Numbers in known cells			
	n_{11}	n_{12}	n_{21}	n_{22}
I -- A	All 20			
II -- B	20	40	50	100
II -- C	25	35	50	70
III -- D	10	100	50	40
III -- E	15	75	20	90
III -- F	40	5	35	120
III -- G	2	90	100	8
III -- H	20	40	100	40
III -- I	30	40	60	20

Table 6.1: Numbers in the known cells for the nine groups.

THE COMPLETE CASES DESIGNS

Exponential regression models were then fitted to these nine groups for the complete cases. The biases for the three parameters μ_{11}, α_2 and β_2 were calculated by $(\hat{\mu}_{11}-\mu_{11})$, $(\hat{\alpha}_2-\alpha_2)$ and $(\hat{\beta}_2-\beta_2)$ and are denoted by $b(\hat{\mu}_{11})$, $b(\hat{\alpha}_2)$ and $b(\hat{\beta}_2)$ respectively. The standard errors, derived from the model fits, for the parameter estimates are given by $se(\hat{\mu}_{11})$, $se(\hat{\alpha}_2)$ and $se(\hat{\beta}_2)$ respectively and the results given in Table 6.2.

Group	$b(\hat{\mu}_{11})$	$se(\hat{\mu}_{11})$	$b(\hat{\alpha}_2)$	$se(\hat{\alpha}_2)$	$b(\hat{\beta}_2)$	$se(\hat{\beta}_2)$
A	0.0040	0.0061	0.0030	0.0071	-0.0036	0.0071
B	-0.0025	0.0051	0.0012	0.0048	0.0066	0.0046
C	-0.0041	0.0050	0.0029	0.0050	0.0080	0.0048
D	-0.0074	0.0060	0.0065	0.0052	0.0067	0.0057
E	0.0003	0.0059	0.0094	0.0045	-0.0013	0.0059
F	0.0025	0.0048	-0.0014	0.0065	0.0034	0.0056
G	-0.0013	0.0106	0.0021	0.0104	-0.0004	0.0103
H	-0.0014	0.0052	0.0033	0.0052	0.0049	0.0049
I	0.0002	0.0049	-0.0009	0.0055	0.0019	0.0056

Table 6.2: Estimated biases and standard errors for the complete cases designs for the nine groups.

All of the estimated biases were compatible with true values of zero bias, except for $b(\hat{\alpha}_2)$ for Group E. This was more than 2 standard errors from zero. However, it is not surprising to find one result significant at the 5% level, given that 27 tests have been conducted.

Thus, it appears that the complete cases designs gave the expected results, confirming that the SAS program generated the observations correctly and the estimation process worked successfully.

The estimated standard errors for Group G for all parameter estimates were much larger than those for the rest of the groups, probably because of the relatively smaller sample sizes in two of the four cells.

THE ALL CASES DESIGNS

For these nine groups with different numbers in the four known cells, a total of 79 different designs were then created with differing numbers in the missing cells. The numbers in these cells were chosen to investigate how the biases in $\hat{\alpha}_2$ and $\hat{\beta}_2$ changed depending on the overall percentage missing and the distribution of the missing observations in the five cells. Exponential regression models were then fitted to the 79 designs in the nine groups.

However, before the results of these 79 designs could be examined in detail, it was necessary to consider the validity of the model based standard errors. This was because the exponential regression model was known to be incorrect, as the hazards were not proportional (Section 6.2.2), when the missing values were included in this manner and, therefore, the standard errors obtained from an incorrect model would probably also be incorrect. The question, therefore, is by how much are the standard errors incorrect?

To tackle this problem, simulations were used to generate 20 replicates for 19 of the 79 designs. For six of these 19 designs, further iterations were performed to obtain larger numbers of replicates. The sample standard deviations in the parameter estimates obtained from these simulations are compared to the model derived standard errors for the 19 designs and the results are given in the next section.

6.3.4 SIMULATION STRATEGY AND RESULTS EXAMINING THE INACCURACIES OF THE MODEL BASED ESTIMATED STANDARD ERRORS FOR THE ALL CASES DESIGNS

INTRODUCTION

To investigate the validity of the estimated standard errors obtained from these designs, simulations were performed to obtain replicates. The numbers (in units of a 1000) in the nine cells for the 19 designs involved in this analysis are given in Table 6.3.

	Known cells				Missing cells				
Design	n_{11}	n_{12}	n_{21}	n_{22}	n_{13}	n_{23}	n_{31}	n_{32}	n_{33}
A2	20	20	20	20	5	10	15	25	35
B4	20	40	50	100	5	10	15	25	35
C1	25	35	50	70	35	25	5	2	50
D5	10	100	50	40	5	5	5	5	5
D6	10	100	50	40	2	5	10	5	2
D7	10	100	50	40	5	2	5	2	5
D9	10	100	50	40	5	2	2	2	5
D10	10	100	50	40	2	2	2	2	2
E1	15	75	20	90	150	150	150	150	150
E8	15	75	20	90	20	20	20	20	20
F5	40	5	35	120	10	10	10	10	10
F10	40	5	35	120	10	10	5	2	5
F12	40	5	35	120	5	2	2	10	5
G2	2	90	100	8	100	50	2	2	2
H9	20	40	100	40	10	10	10	10	10
H14	20	40	100	40	10	10	5	5	2
H21	20	40	100	40	2	2	10	2	2
I2	30	40	60	20	5	5	5	5	5
I8	30	40	60	20	2	2	2	2	2

Table 6.3: Numbers in the nine cells for the designs where replicates were simulated.

For each of these 19 designs, the 20 replicates were generated to identify any substantial deviations between the model based standard errors and the sampling standard deviations for a particular design. However, since 20 is not a large number of iterations, further replications were obtained for six of the designs. This was to try to identify more subtle differences. The number of additional iterations was limited by computer space, due to the size of the seeds file used to generate the observations and the sequences of random uniform numbers, and by the time taken to run the simulations. The larger numbers of replicates used are given in Table 6.4.

Design	Number of replicates
A2	285
D5	100
D7	100
G2	100
H21	100
I2	165

Table 6.4: The larger numbers of replicates generated for six of the designs.

From Cox & Oakes (1984) and Ford et al (1995), the variance-covariance matrix for the exponential regression model can be computed solely from the design matrix and does not depend on the parameter estimates. Therefore, the model based standard errors obtained for $\hat{\alpha}_2$ and $\hat{\beta}_2$ will be independent of the simulated data given a particular design.

The aim here was to compare whether this known asymptotic standard error, based on the assumption that the exponential regression model fitted the data, was similar to the sampling variability of the parameter estimates, due to the bias introduced by the fact that the exponential regression model was inappropriate. The values of the theoretical model based standard errors for $\hat{\alpha}_2$ and $\hat{\beta}_2$ obtained from fitting the model once (see Section 6.3.5) are given for each design, along with the values of the sampling standard deviation obtained on the 20 parameter estimates. For the six designs where larger numbers of replicates were obtained, similar values are also given for these analyses.

The standard errors for $\hat{\mu}_{11}$ are not given as the main interest was in whether or not the parameter estimates $\hat{\alpha}_2$ and $\hat{\beta}_2$ were affected by the inclusion of the missing values in the design. For all of the replicates, the results are given below in Table 6.5. The sampling standard deviations from the 20 (or larger numbers of) parameter estimates provide an idea about the true sampling variability for the designs.

Design	Parameter	Theoretical standard error	Sampling standard deviation obtained from 20 reps	Sampling standard deviation obtained from 20 reps
A2	$\hat{\alpha}_2$	0.0065	0.0052	0.0066
	$\hat{\beta}_2$	0.0058	0.0062	0.0060
B4	$\hat{\alpha}_2$	0.0047	0.0050	*
	$\hat{\beta}_2$	0.0042	0.0038	*
C1	$\hat{\alpha}_2$	0.0043	0.0045	*
	$\hat{\beta}_2$	0.0047	0.0052	*
D5	$\hat{\alpha}_2$	0.0050	0.0062	0.0052
	$\hat{\beta}_2$	0.0054	0.0056	0.0049
D6	$\hat{\alpha}_2$	0.0051	0.0062	*
	$\hat{\beta}_2$	0.0053	0.0058	*
D7	$\hat{\alpha}_2$	0.0051	0.0060	0.0050
	$\hat{\beta}_2$	0.0055	0.0053	0.0050
D9	$\hat{\alpha}_2$	0.0051	0.0060	*
	$\hat{\beta}_2$	0.0055	0.0060	*
D10	$\hat{\alpha}_2$	0.0051	0.0059	*
	$\hat{\beta}_2$	0.0056	0.0057	*
E1	$\hat{\alpha}_2$	0.0028	0.0028	*
	$\hat{\beta}_2$	0.0030	0.0027	*
E8	$\hat{\alpha}_2$	0.0041	0.0035	*
	$\hat{\beta}_2$	0.0050	0.0035	*
F5	$\hat{\alpha}_2$	0.0058	0.0059	*
	$\hat{\beta}_2$	0.0051	0.0051	*
F10	$\hat{\alpha}_2$	0.0059	0.0064	*
	$\hat{\beta}_2$	0.0053	0.0055	*
F12	$\hat{\alpha}_2$	0.0062	0.0060	*
	$\hat{\beta}_2$	0.0054	0.0062	*
G2	$\hat{\alpha}_2$	0.0048	0.0042	0.0051
	$\hat{\beta}_2$	0.0062	0.0050	0.0061

Table 6.5: Estimated standard errors for each of the designs (Theoretical standard error) and the sampling standard deviations obtained from the 20 (or larger numbers of) estimates. Note that '' indicates that only 20 replicates were obtained for that design.*

Design	Parameter	Theoretical standard error	Sampling standard deviation obtained from 20 reps	Sampling standard deviation obtained from 20 reps
H9	$\hat{\alpha}_2$	0.0049	0.0044	*
	$\hat{\beta}_2$	0.0046	0.0038	*
H14	$\hat{\alpha}_2$	0.0049	0.0044	*
	$\hat{\beta}_2$	0.0047	0.0040	*
H21	$\hat{\alpha}_2$	0.0051	0.0049	0.0049
	$\hat{\beta}_2$	0.0048	0.0038	0.0046
I2	$\hat{\alpha}_2$	0.0053	0.0045	0.0054
	$\hat{\beta}_2$	0.0054	0.0057	0.0055
I8	$\hat{\alpha}_2$	0.0054	0.0045	*
	$\hat{\beta}_2$	0.0055	0.0055	*

Table 6.5 cont: Estimated standard errors for each of the designs (Theoretical standard error) and the sampling standard deviations obtained from the 20 (or larger numbers of) estimates. Note that '' indicates that only 20 replicates were obtained for that design.*

The fact that the theoretical standard errors were similar to these sampling standard deviations, based on the 20 or more replicates, leads to the conclusion that the theoretical model based standard errors can be taken as being reasonable. Therefore, the results of fitting exponential regression models to the 79 designs in the nine groups can be discussed in the next section using the knowledge that the standard errors are probably acceptable.

6.3.5 EXAMINATION OF THE OBSERVED ESTIMATED BIASES AND ESTIMATED STANDARD ERRORS FOR THE ALL CASES DESIGNS

INTRODUCTION

The nine groups that were modelled in the complete cases designs were the basis of the exponential regression model fitted for 79 designs on all cases. Table 6.6 shows the numbers in the four known cells. All of the numbers are in units of a 1000.

Group	Numbers in known cells n_{11}	n_{12}	n_{21}	n_{22}
A	20	20	20	20
B	20	40	50	100
C	25	35	50	70
D	10	100	50	40
E	15	75	20	90
F	40	5	35	120
G	2	90	100	8
H	20	40	100	40
I	30	40	60	20

Table 6.6: Numbers in the known cells for the nine groups.

Having just shown that the standard errors (s.e.) for these designs appeared to be reasonable even though the model that was fitted was incorrect, it was then possible to test informally whether or not the parameter estimates obtained from these 79 designs were biased. This was possible by comparing the magnitude of estimated bias with its approximate estimated standard error. Only the values of $\hat{\alpha}_2$ and $\hat{\beta}_2$ were examined and the estimated biases $b(\hat{\alpha}_2) = (\hat{\alpha}_2 - \alpha_2)$ and $b(\hat{\beta}_2) = (\hat{\beta}_2 - \beta_2)$ are presented along with the estimated standard errors $se(\hat{\alpha}_2)$ and $se(\hat{\beta}_2)$.

The overall percentage of observations falling into the five cells with missing factor level information is given by '% missing' for each model. The results for the nine groups in the three types are now presented separately.

TYPE I – GROUP A

This group had complete symmetry with equal numbers in the four known cells (Figure 6.8).

	F2	
F1	20	20
	20	20

Figure 6.8: Numbers in the known cells in Group A.

Table 6.7 gives the numbers in the missing cells and the overall percentage missing in both designs in Group A, whilst the estimated biases and standard errors for $\hat{\alpha}_2$ and $\hat{\beta}_2$ for the two designs are given in Table 6.8.

	Number in missing cells					
Design	**n_{13}**	**n_{23}**	**n_{31}**	**n_{32}**	**n_{33}**	**% missing**
A1	20	20	20	20	20	56
A2	5	10	15	25	35	53

Table 6.7: Numbers in the missing cells and % missing in Group A.

Design	$b(\hat{\alpha}_2)$	$se(\hat{\alpha}_2)$	$b(\hat{\beta}_2)$	$se(\hat{\beta}_2)$
A1	<0.0001	0.0058	-0.0032	0.0058
A2	0.0024	0.0065	-0.0028	0.0058

Table 6.8: Estimated biases and standard errors for the designs in Group A.

There appears to be no evidence of bias in the parameter estimates obtained for this group.

TYPE II -- GROUP B

This group had equal proportions in the numbers in level 1 of factor F2 for both levels of factor F1, with $\frac{n_{11}}{n_{11}+n_{12}} = \frac{n_{21}}{n_{21}+n_{22}} = 0.33$. These proportions are the mixing parameters r and s (Section 6.3.2 and Appendix 9) for weighting the two exponential distributions used to generate the observations in the missing cells $(1,3)$ and $(2,3)$ respectively. These proportions are observed by examining the breakdown of the numbers in the known cells for this group (Figure 6.9).

	F2	
F1	20	40
	50	100

Figure 6.9: Numbers in the known cells in Group B.

Table 6.9 gives the numbers in the missing cells for the six designs in Group B.

Design	Number in missing cells n_{13}	n_{23}	n_{31}	n_{32}	n_{33}	% missing
B1	100	5	75	150	2	61
B2	40	40	40	40	40	49
B3	20	20	20	20	20	32
B4	5	10	15	25	35	30
B5	15	15	15	15	15	26
B6	5	5	5	5	5	11

Table 6.9: Numbers in the missing cells and % missing in Group B.

The estimated biases and standard errors for the parameter estimates are given in Table 6.10 for these designs.

Design	$b(\hat{\alpha}_2)$	$se(\hat{\alpha}_2)$	$b(\hat{\beta}_2)$	$se(\hat{\beta}_2)$
B1	0.0022	0.0045	0.0007	0.0032
B2	0.0007	0.0040	0.0033	0.0039
B3	-0.0007	0.0043	0.0049	0.0042
B4	0.0009	0.0047	0.0052	0.0042
B5	<0.0001	0.0044	0.0048	0.0043
B6	0.0010	0.0047	0.0059	0.0045

Table 6.10: Estimated biases and standard errors for the designs in Group B.

Again, there was no evidence of biased parameter estimates.

TYPE II -- GROUP C

This is another group in type II where there were equal proportions of observations for level 1 out of the total of factor F2 for both levels of F1 (Figure 6.10).

	F2	
F1	25	35
	50	70

Figure 6.10: Numbers in the known cells in Group C.

Table 6.11 gives the numbers in the missing cells and the overall percentages missing. The biases and model based standard errors for the parameter estimates are shown in Table 6.12.

Design	Number in missing cells					% missing
	n_{13}	n_{23}	n_{31}	n_{32}	n_{33}	
C1	35	25	5	2	50	39
C2	15	15	15	15	15	29

Table 6.11: Numbers in the missing cells and % missing in Group C.

Design	$b(\hat{\alpha}_2)$	$se(\hat{\alpha}_2)$	$b(\hat{\beta}_2)$	$se(\hat{\beta}_2)$
C1	0.0021	0.0043	0.0061	0.0047
C2	0.0013	0.0046	0.0059	0.0044

Table 6.12: Estimated biases and standard errors for the designs in Group C.

There was no evidence of bias in the parameter estimates.

It, therefore, appears that for both type I and type II contexts the parameter estimates were not biased, despite the fitted model being incorrect. These types have equal proportions in the known cells for $r = \dfrac{n_{11}}{n_{11} + n_{12}}$ and $s = \dfrac{n_{21}}{n_{21} + n_{22}}$; i.e. $r = s$ (see Section 6.3.2 and Appendix 9). In this situation, the expectations in each cell are compatible with a proportional hazards model even though the exponential regression model is not valid.

For example, the expectation for observations in cell $(1,3)$ is $\exp(-\mu)[r + (1-r)\exp(-\beta_2)]$ and the expectation for observations in cell $(1,1)$ is $\exp(-\mu)$. Therefore, the ratio of the expectations, and hence the hazards for the exponential distributions, for cell $(1,3)$ to cell $(1,1)$ is $r + (1-r)\exp(-\beta_2)$.

Similarly, the expectation for the observations in cell $(2,3)$ is $\exp(-\mu - \alpha_2)[s + (1-s)\exp(-\beta_2)]$ and for cell $(2,1)$, $\exp(-\mu - \alpha_2)$. Therefore, the ratio of cell $(2,3)$ to cell $(2,1)$ is $s+(1-s)\exp(-\beta_2)=r+(1-r)\exp(-\beta_2)$, since $r=s$. The ratio of the expectations for cell $(3,3)$ versus cell $(3,1)$ also has the same ratio.

The expectations have the form they would have had if the observations were from an additive exponential regression model. Thus, the expectations are proportional even though the exponential regression model is not appropriate, with the true hazards not being proportional. A sufficient statistic for the maximum likelihood estimate (MLE) is the sets of sums of observations in each of the cells. Hence, a MLE will be based solely on these quantities and in this case, when the mixing weights are in correct proportions, the sums are compatible with an additive exponential model and the estimates would be asymptotically unbiased.

TYPE III -- GROUP D

The numbers in the known cells for Group D are given in Figure 6.11.

	F2	
F1	10	100
	50	40

Figure 6.11: Numbers in the known cells in Group D.

Table 6.13 shows the numbers in the missing cells for the designs in this group.

Design	Number in missing cells n_{13}	n_{23}	n_{31}	n_{32}	n_{33}	% missing
D1	50	50	5	50	5	44
D2	35	25	5	2	50	37
D3	5	10	15	25	35	31
D4	10	10	10	10	10	20
D5	5	5	5	5	5	11
D6	2	5	10	5	2	11
D7	5	2	5	2	5	9
D8	2	5	2	5	2	7
D9	5	2	2	2	5	7
D10	2	2	2	2	2	5

Table 6.13: Numbers in the missing cells and % missing in Group D.

The estimated biases and model based standard errors for the parameter estimates are presented in Table 6.14.

Design	$b(\hat{\alpha}_2)$	$se(\hat{\alpha}_2)$	$b(\hat{\beta}_2)$	$se(\hat{\beta}_2)$
D1	-0.0663	0.0040	-0.0581	0.0051
D2	-0.0440	0.0044	-0.0356	0.0053
D3	-0.0345	0.0049	-0.0622	0.0049
D4	-0.0289	0.0048	-0.0427	0.0052
D5	-0.0129	0.0050	-0.0199	0.0054
D6	-0.0127	0.0051	-0.0252	0.0053
D7	-0.0073	0.0051	-0.0140	0.0055
D8	-0.0063	0.0051	-0.0097	0.0055
D9	-0.0037	0.0051	-0.0069	0.0055
D10	-0.0023	0.0051	-0.0060	0.0056

Table 6.14: Estimated biases and standard errors for the designs in Group D.

When there was 7% or less of the total missing, there was no evidence of significant bias in the parameter estimates. The parameter estimate $\hat{\alpha}_2$ also did not appear to be biased when there was 9% missing, but the estimate for β_2 was.

Note that, as the overall percentage missing increased, the standard errors of $\hat{\alpha}_2$ and $\hat{\beta}_2$ decreased. The total sample size was not controlled and, therefore, as more cases were added into the missing cells, the overall total numbers of cases increased for the particular group design. This led to the standard errors decreasing because the parameter estimates were being calculated from more data.

In contrast, as the percentage of missing values increased, so did the magnitude of the estimated bias of both $\hat{\alpha}_2$ and $\hat{\beta}_2$, in general. To illustrate this, the magnitudes of the estimated biases were plotted against the percentage missing for the 10 designs in this group (Figure 6.12).

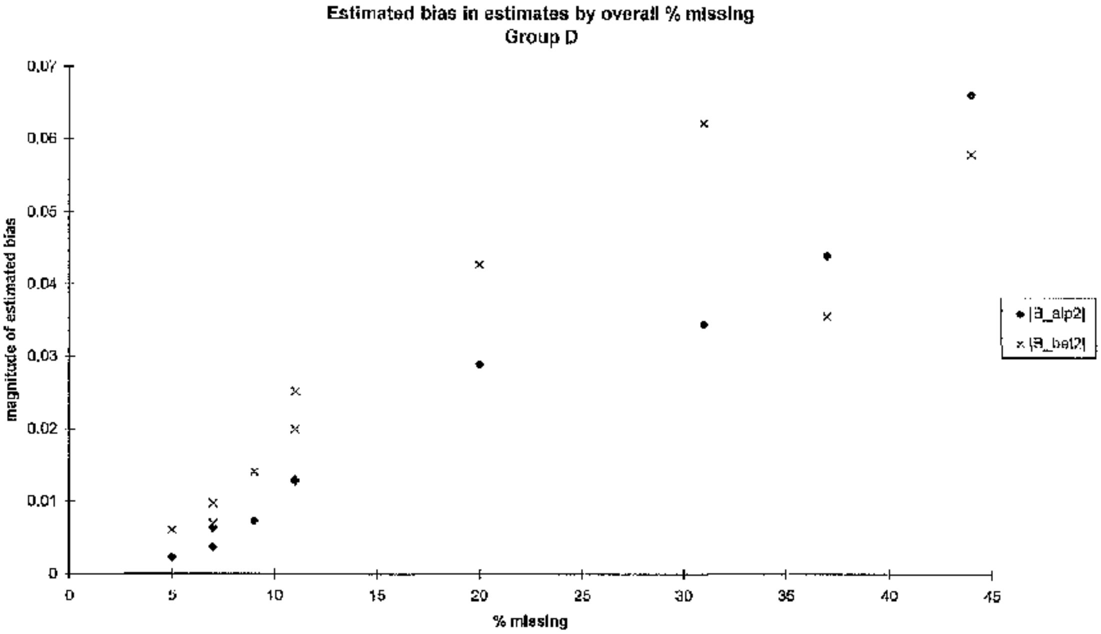


Figure 6.12: Plot of the magnitude of the estimated bias against the overall percentage missing for Group D.

TYPE III -- GROUP E

The numbers in the known cells in this group were very similar to those of type II since the proportions in level 1 of F2 for levels 1 and 2 of F1 were $\frac{1}{6}$ and $\frac{2}{11}$ respectively, with the numbers in the known cells given in Figure 6.13.

	F2	
F1	15	75
	20	90

Figure 6.13: Numbers in the known cells in Group E.

The numbers in the missing cells are shown in Table 6.15 and the estimated biases and standard errors for the parameter estimates from the designs in this group are presented in Table 6.16.

Design	Number in missing cells					% missing
	n_{13}	n_{23}	n_{31}	n_{32}	n_{33}	
E1	150	150	150	150	150	79
E2	125	125	125	125	125	76
E3	125	150	100	75	150	75
E4	100	100	100	100	100	71
E5	75	75	75	75	75	65
E6	50	50	50	50	50	56
E7	40	40	40	40	40	50
E8	20	20	20	20	20	33
E9	5	5	5	5	5	11

Table 6.15: Numbers in the missing cells and % missing in Group E.

Design	$b(\hat{\alpha}_2)$	$se(\hat{\alpha}_2)$	$b(\hat{\beta}_2)$	$se(\hat{\beta}_2)$
E1	0.0005	0.0028	-0.0123	0.0030
E2	0.0014	0.0030	-0.0126	0.0032
E3	0.0009	0.0029	-0.0099	0.0036
E4	0.0017	0.0032	-0.0114	0.0035
E5	0.0052	0.0034	-0.0086	0.0038
E6	0.0039	0.0037	-0.0080	0.0042
E7	0.0054	0.0038	-0.0074	0.0044
E8	0.0058	0.0041	-0.0053	0.0050
E9	0.0085	0.0044	-0.0027	0.0056

Table 6.16: Estimated biases and standard errors for the designs in Group E.

Despite extensive numbers of cases with missing factor levels (even when there was as much as 79% of the total missing), there was no evidence of bias in the parameter estimate for α_2 in any of the designs. However, when there was 65% or more of cases missing, $\hat{\beta}_2$ appeared to be biased, although with 56% or less of the total number of cases missing, there was no evidence of bias for either $\hat{\alpha}_2$ or $\hat{\beta}_2$.

Figure 6.14 shows a systematic pattern of increasing magnitude of bias for $\hat{\beta}_2$ as the percentage of missing values increased. Although there appears to be a decreasing trend for the magnitude of bias of $\hat{\alpha}_2$ with increasing percentage of missing values, there was no evidence that any of the biases for $\hat{\alpha}_2$ were different from zero.

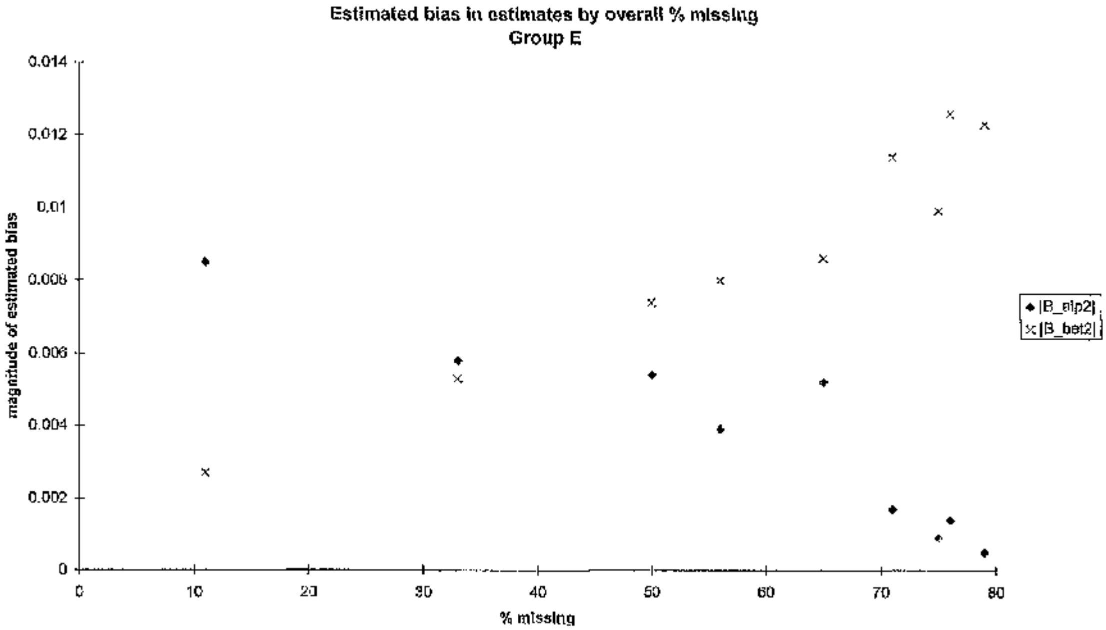


Figure 6.14: Plot of the magnitude of the estimated bias against the overall percentage missing for Group E.

TYPE III -- GROUP F

The numbers in the known and missing cells for Group F are given in Figure 6.15 and Table 6.17 respectively.

	F2	
F1	40	5
	35	120

Figure 6.15: Numbers in the known cells in Group F.

Design	Number in missing cells					% missing
	n_{13}	n_{23}	n_{31}	n_{32}	n_{33}	
F1	20	40	100	35	20	52
F2	40	40	40	40	40	50
F3	20	20	20	20	20	33
F4	15	15	15	15	15	27
F5	10	10	10	10	10	20
F6	10	10	10	5	5	17
F7	10	10	5	10	5	17
F8	10	10	5	5	10	17
F9	10	10	5	5	5	15
F10	10	10	5	2	5	14
F11	5	5	5	5	5	11
F12	5	2	2	10	5	11
F13	10	5	5	2	2	11
F14	2	2	2	2	2	5

Table 6.17: Numbers in the missing cells and % missing in Group F.

Table 6.18 shows the biases and model based standard errors of $\hat{\alpha}_2$ and $\hat{\beta}_2$.

Design	$b(\hat{\alpha}_2)$	$se(\hat{\alpha}_2)$	$b(\hat{\beta}_2)$	$se(\hat{\beta}_2)$
F1	0.0111	0.0050	0.0956	0.0041
F2	0.0477	0.0046	0.0687	0.0042
F3	0.0247	0.0053	0.0413	0.0047
F4	0.0197	0.0055	0.0322	0.0049
F5	0.0143	0.0058	0.0219	0.0051
F6	0.0201	0.0059	0.0127	0.0052
F7	0.0189	0.0058	0.0114	0.0052
F8	0.0229	0.0059	0.0061	0.0053
F9	0.0228	0.0059	0.0061	0.0053
F10	0.0282	0.0059	-0.0016	0.0053
F11	0.0068	0.0061	0.0134	0.0053
F12	0.0036	0.0062	0.0096	0.0054
F13	0.0184	0.0061	0.0024	0.0054
F14	0.0028	0.0064	0.0062	0.0055

Table 6.18: Estimated biases and standard errors for the designs in Group F.

When there was 20% or more of the total missing, both of the parameter estimates were biased in all of the designs. With 17% of the observations were missing, designs F6 and F7 also had biased parameter estimates, but there was no evidence of bias for $\hat{\beta}_2$ for design F8. Thus, it appears that the distribution of the missing values affected the estimation process.

From Figure 6.16, the magnitude of the bias of $\hat{\beta}_2$ increased as the percentage of missing values increased. No obvious relationship was apparent for the bias of $\hat{\alpha}_2$.

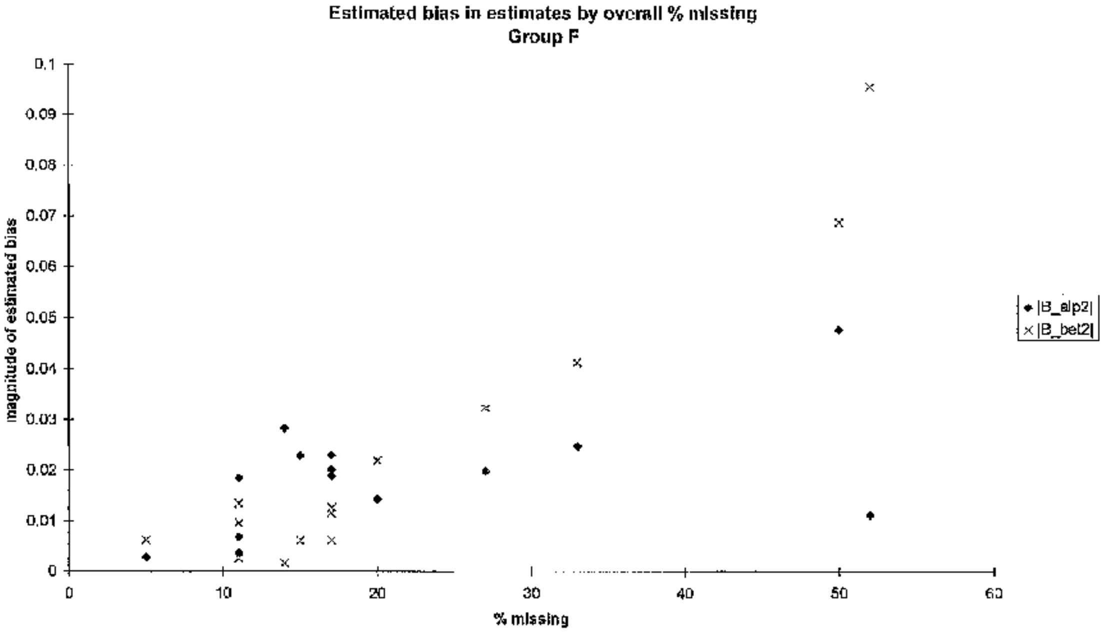


Figure 6.16: Plot of the magnitude of the estimated bias against the overall percentage missing for Group F.

TYPE III -- GROUP G

The numbers in the four known cells in this group are given in Figure 6.17.

	F2	
F1	2	90
	100	8

Figure 6.17: Numbers in the known cells in Group G.

Table 6.19 provides the numbers in the missing cells for this group, whilst the estimated biases and standard errors for $\hat{\alpha}_2$ and $\hat{\beta}_2$ are shown in Table 6.20.

Design	Number in missing cells					% missing
	n_{13}	n_{23}	n_{31}	n_{32}	n_{33}	
G1	2	2	2	100	100	51
G2	100	50	2	2	2	44
G3	15	15	15	15	15	27
G4	5	5	5	5	5	11
G5	2	2	2	2	2	5
G6	1	1	1	1	1	2

Table 6.19: Numbers in the missing cells and % missing in Group G.

Design	$b(\hat{\alpha}_2)$	$se(\hat{\alpha}_2)$	$b(\hat{\beta}_2)$	$se(\hat{\beta}_2)$
G1	-0.1040	0.0092	-0.1136	0.0091
G2	-0.2213	0.0048	-0.2084	0.0062
G3	-0.2199	0.0069	-0.2346	0.0068
G4	-0.1221	0.0087	-0.1287	0.0087
G5	-0.0634	0.0096	-0.0680	0.0096
G6	-0.0325	0.0100	-0.0366	0.0100

Table 6.20: Estimated biases and standard errors for the designs in Group G.

Despite the fact that design G6 had only 2% of the total number of observations missing, both of the parameter estimates for all of the designs were biased.

A quadratic pattern was seen in the plots of the magnitudes of the biases for the two parameters against the percentages missing for the different designs (Figure 6.17), again suggesting that the distribution of the missing values is important to the estimation process.

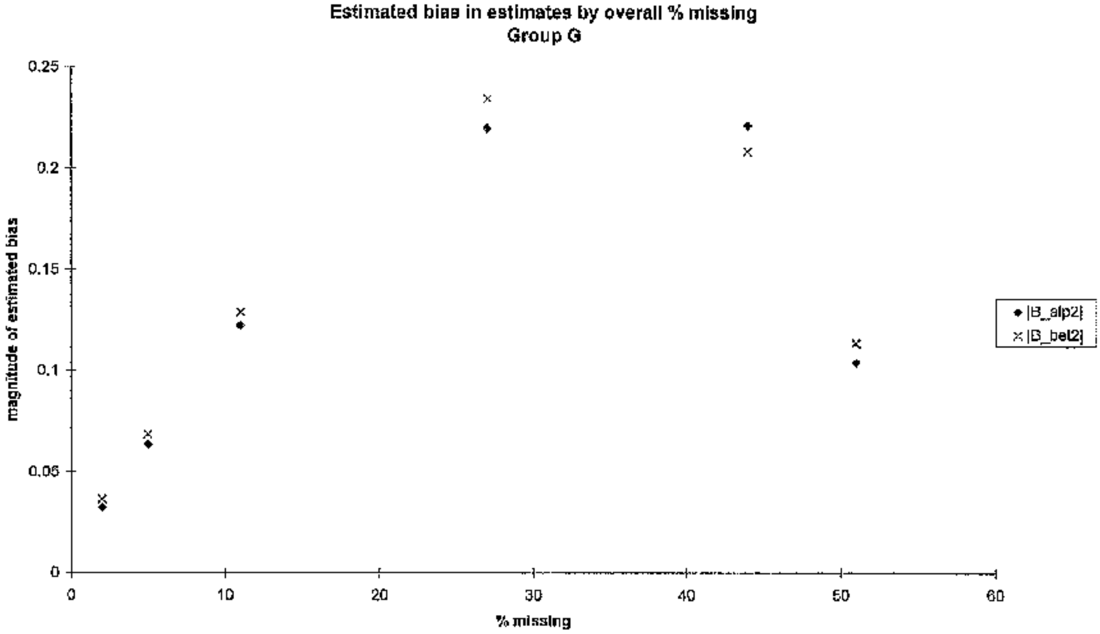


Figure 6.18: Plot of the magnitude of the estimated bias against the overall percentage missing for Group G.

TYPE III -- GROUP H

Figure 6.19 and Table 6.21 gives the numbers in the known and missing cells respectively.

	F2	
F1	20	40
	100	40

Figure 6.19: Numbers in the known cells in Group H.

Design	Number in missing cells					% missing
	n_{13}	n_{23}	n_{31}	n_{32}	n_{33}	
H1	40	40	40	40	40	50
H2	150	2	5	2	15	47
H3	25	25	25	25	25	38
H4	100	2	5	2	15	38
H5	100	2	5	2	10	37
H6	10	2	5	5	50	26
H7	40	10	5	2	10	25
H8	40	2	5	2	15	24
H9	10	10	10	10	10	20
H10	5	10	10	5	5	15
H11	10	5	5	5	10	15
H12	2	10	2	10	10	15
H13	10	5	5	2	10	14
H14	10	10	5	5	2	14
H15	10	2	10	5	5	14
H16	10	5	10	2	5	14
H17	10	5	10	5	2	14
H18	10	10	5	2	2	13
H19	5	5	5	5	5	11
H20	10	5	5	2	2	11
H21	2	2	10	2	2	8
H22	2	2	2	2	2	5

Table 6.21: Numbers in the missing cells and % missing in Group H.

The table of results showing the estimated biases and model based standard errors for Group H for the parameter estimates are given in Table A10.1 in Appendix 10, along with a plot of the magnitudes of the estimated biases against the percentage of missing values (Figure A10.1 in Appendix 10).

There was no obvious pattern between whether or not the parameter estimates were biased and the overall percentage missing. For example, design H2 had 47% missing and yet there was no evidence of bias of $\hat{\beta}_2$, whereas designs H14-H17 had only 14% missing, but both of the parameters were biased for these four designs.

When the estimated biases for both parameter estimates were compared when there were equal numbers in each of the missing cells (designs H1, H3, H9, H19 and H22), the magnitude of the biases of the parameter estimates decreased as the numbers in these cells decreased (Table A10.1, Appendix 10).

TYPE III -- GROUP I

The breakdown of the numbers in the four known cells in this group is shown in Figure 6.20 and Table 6.22 gives the numbers in the missing cells.

	F2	
F1	30	40
	60	20

Figure 6.20: Numbers in the known cells in Group I.

Design	Number in missing cells n_{13}	n_{23}	n_{31}	n_{32}	n_{33}	% missing
I1	10	10	10	10	10	25
I2	5	5	5	5	5	14
I3	5	5	5	5	2	13
I4	5	5	5	2	5	13
I5	5	5	5	2	2	11
I6	5	2	5	5	2	11
I7	2	5	5	5	2	11
I8	2	2	2	2	2	6

Table 6.22: Numbers in the missing cells and % missing in Group I.

The estimated biases and standard errors for $\hat{\alpha}_2$ and $\hat{\beta}_2$ for these eight designs in Group I are given in Appendix 10 (Table A10.2). The plot of the magnitudes of the biases against the overall percentage missing is also presented in Appendix 10 (Figure A10.2). There was a general trend for increasing magnitude of bias in the two parameters with increasing overall percentage of missing values.

CONCLUSION

Except Group E designs, which were similar to type II designs (Groups B and C), all of the type III designs had some combinations of numbers in the missing cells that produced biased parameter estimates when an exponential regression model was fitted to the design.

6.3.6 APPLICATION IN THE CONTEXT OF THE BREAST CANCER AUDIT DATA

INTRODUCTION

Following on from the investigation into the effects of varying the numbers in both the known and missing cells using the 79 designs in the nine groups, six new designs were examined. The new analyses were performed to check the general impressions formed from the nine groups regarding the parameter estimates remained for datasets with a similar missingness structure to the Breast Cancer Audit data. The six designs were based on the distributions of subjects with known and missing information for the pairwise combinations of the four clinical variables: clinical stage, node status, tumour size and ER status. The true numbers in the pairwise combinations for the Breast Cancer Audit are given in Tables A4.5 to A4.10 of Appendix 4. Here, however, to obtain an estimate of bias in an asymptotic context, the numbers in each of the six new designs are scaled up by 1000 in each cell.

The same underlying exponential regression model and methods of simulating data were used as in the previous section. As before, it was assumed there was no censoring and that the true values of the parameters also remained the same.

THE DESIGNS AND THE RESULTS

The six designs are C_E (clinical stage by ER status); C_N (clinical stage by node status); C_T (clinical stage by tumour size); E_N (ER status by node status); E_T (ER status by tumour size) and N_T (node status by tumour size). Table 6.23 shows the numbers, in units of a 1000, in the known cells for these designs.

	Numbers in known cells			
Design	n_{11}	n_{12}	n_{21}	n_{22}
C_E	447	278	64	54
C_N	363	464	120	28
C_T	475	434	30	128
E_N	237	244	158	150
E_T	267	234	131	186
N_T	171	312	269	212

Table 6.23: Numbers in the known cells for the six new designs.

The estimated biases and model based standard errors for the parameter estimates obtained from the complete cases models for the six new designs are given in Table 6.24.

Design	$b(\hat{\alpha}_2)$	$se(\hat{\alpha}_2)$	$b(\hat{\beta}_2)$	$se(\hat{\beta}_2)$
C_E	0.0053	0.0031	-0.0007	0.0022
C_N	0.0004	0.0029	-0.0015	0.0021
C_T	0.0066	0.0028	-0.0026	0.0020
E_N	0.0034	0.0023	0.0016	0.0023
E_T	0.0024	0.0023	0.0014	0.0022
N_T	0.0025	0.0021	0.0022	0.0021

Table 6.24: Estimated biases and standard errors for the complete cases models for the six new designs .

As expected, there was little evidence of bias for the complete cases data with the possible exception of Clinical stage by Tumour size. Here, the estimated bias for $\hat{\alpha}_2$ was 2.34 standard errors.

For the all cases designs, the numbers in the missing cells and the overall percentage missing are shown in Table 6.25. All of the designs had a high percentage of cases missing, ranging from 34% to 51%.

	Number in missing cells					
Design	**n_{13}**	**n_{23}**	**n_{31}**	**n_{32}**	**n_{33}**	**% missing**
C_E	390	69	88	59	170	48
C_N	288	39	100	109	108	40
C_T	206	29	120	100	97	34
E_N	118	83	188	207	234	51
E_T	98	74	227	242	160	49
N_T	100	120	185	138	112	40

Table 6.25: Numbers in the missing cells and % missing.

Based on the finding from the last section, it was anticipated that the model fitted to the data in design E_N would produce estimates with less bias, if any at all, since the proportions $r = \dfrac{n_{11}}{n_{11} + n_{12}}$ and $s = \dfrac{n_{21}}{n_{21} + n_{22}}$ (Appendix 9) were very similar for this design (0.49 and 0.51 respectively). The estimated biases and model based standard errors for the parameter estimates for the all cases models are reported in Table 6.26.

Design	$b(\hat{\alpha}_2)$	$se(\hat{\alpha}_2)$	$b(\hat{\beta}_2)$	$se(\hat{\beta}_2)$
C_E	0.0136	0.0025	0.0010	0.0021
C_N	-0.0274	0.0026	-0.0204	0.0019
C_T	0.0200	0.0026	0.0075	0.0018
E_N	0.0014	0.0021	-0.0027	0.0018
E_T	0.0074	0.0021	0.0184	0.0018
N_T	-0.0147	0.0019	-0.0272	0.0018

Table 6.26: Estimated biases and standard errors for the six designs in the all cases analyses.

The expected result for design E_N was observed. For the remaining designs, there was evidence of bias for both of the parameter estimates, except for $\hat{\beta}_2$ for design C_E. However, for all designs the estimated biases were very small in magnitude, both relatively and absolutely. In fact, the magnitude of $b(\hat{\alpha}_2)$ for the C_N design was the largest. The value of 0.0274 represents 5.5% of the true value, 0.5. By contrast, in the simulated exercise presented in the last section, the largest percentage bias was 78% in G3 for $b(\hat{\beta}_2)$.

SMALL SAMPLE ANALYSES

The simulations so far, both in this section and previously, have been designed to investigate bias in a large sample context. The analyses all relied on the large sample properties of maximum likelihood estimates theory to obtain the standard errors of the parameter estimates. However, in general, there are not usually 100,000 subjects available for analysis. It was, therefore, interesting to fit exponential regression models to the real sample size of the Breast Cancer Audit, with only 1619 subjects in total, to observe the sizes of the biases and standard errors obtained when a relatively small sample size was modelled using an incorrect exponential regression model to investigate the bias due to the proportional hazards assumption being violated.

The results for all these models are based on 20 replicates, with several designs having either 1,000 or 10,000 replicates. The average biases for the estimates based on the different numbers of replicates are given in Table 6.27, along with the sampling standard error for the average biases of the parameter estimates.

Design	Average bias of $\hat{\alpha}_2$	Standard error for the average bias of $\hat{\alpha}_2$	Average bias of $\hat{\beta}_2$	Standard error for the average bias of $\hat{\beta}_2$
C_E				
20 reps.	0.0228	0.0151	-0.0135	0.0146
1000 reps.	0.0159	0.0026	0.0013	0.0020
10,000 reps.	0.0103	0.0008	0.0029	0.0007
C_N				
20 reps.	-0.0363	0.0156	-0.0236	0.0149
10,000 reps.	-0.0361	0.0007	-0.0172	0.0006
C_T				
20 reps.	0.0203	0.0145	0.0063	0.0151
E_N				
20 reps.	-0.0159	0.0168	-0.0082	0.0154
E_T				
20 reps.	-0.0142	0.0183	0.0159	0.0125
1000 reps.	0.0152	0.0018	0.0089	0.0198
10,000 reps.	0.0051	0.0007	0.0197	0.0006
N_T				
20 reps.	-0.0295	0.0115	-0.0366	0.0108
1000 reps.	-0.0147	0.0018	-0.0274	0.0018
10,000 reps.	-0.0155	0.0006	-0.0272	0.0006

Table 6.27: Average biases for $\hat{\alpha}_2$ and $\hat{\beta}_2$ with standard errors for these estimates for the six designs based on the true small sample sizes for varying numbers of replicates (reps.).

These data show the extent of bias with the unknown pattern observed in the pairs of clinical factors in the Breast Cancer Audit data assuming that the data were exponential. As can be seen, there is considerable bias in both parameters when more than 20 replicates are included in the analysis for all of the designs. There is also evidence of bias for both parameter estimates for design N_T when only 20 replicates are included in the analysis. For the design C_N, $\hat{\alpha}_2$ appears to be biased from only 20 replicates.

6.4 RESULTS FROM FITTING COX REGRESSION MODELS TO THE EXPONENTIAL DATASETS

INTRODUCTION

The parameter estimates obtained from fitting an exponential regression model to data which did not satisfy the proportional hazards assumption in the additive main effects model situation were biased for some designs (Sections 6.3.5 and 6.3.6). A main effects Cox proportional hazards regression model is now fitted to four of the nine groups described in Section 6.3.5. The simulated datasets were generated from exponential distributions as previously and the same true values of the parameters used.

When the Cox regression models were fitted to these designs, it was again assumed that there was no censoring. The numbers in the known and missing cells for all of the designs in the four groups remained the same so that, for each design, the results from fitting a Cox model could be compared to the results obtained from fitting the exponential regression model. Note that all of the numbers were in units of 1000 so that the large sample context could be examined.

RESULTS FOR THE COX MODELLING

TYPE I – GROUP A

The numbers in this group for the four known cells are given in Figure 6.21.

	F2	
F1	20	20
	20	20

Figure 6.21: Numbers in the known cells in Group A.

Table 6.28 shows the numbers in the missing cells and the overall percentage missing in each design. The estimated biases and estimated standard errors for $\hat{\alpha}_2$ and $\hat{\beta}_2$ for the two designs that were fitted are presented in Table 6.29.

Design	Number in missing cells					% missing
	n_{13}	n_{23}	n_{31}	n_{32}	n_{33}	
A1	20	20	20	20	20	56
A2	5	10	15	25	35	53

Table 6.28: Numbers in the missing cells and % missing in Group A.

Design	$b(\hat{\alpha}_2)$	$se(\hat{\alpha}_2)$	$b(\hat{\beta}_2)$	$se(\hat{\beta}_2)$
A1	-0.0116	0.0059	0.0101	0.0058
A2	-0.0123	0.0066	-0.0114	0.0059

Table 6.29: Estimated biases and standard errors from fitting the Cox model for the designs in Group A.

There was no evidence of bias in either $\hat{\alpha}_2$ and $\hat{\beta}_2$ for each of the two designs.

TYPE II -- GROUP B

The breakdown of the numbers in the known and missing cells for this group are given in Figure 6.22 and Table 6.30 respectively.

	F2	
F1	20	40
	50	100

Figure 6.22: Numbers in the known cells in Group B.

Design	Number in missing cells					% missing
	n_{13}	n_{23}	n_{31}	n_{32}	n_{33}	
B1	100	5	75	150	2	61
B2	40	40	40	40	40	49
B3	20	20	20	20	20	32
B4	5	10	15	25	35	30
B5	15	15	15	15	15	26
B6	5	5	5	5	5	11

Table 6.30: Numbers in the missing cells and % missing in Group B.

Table 6.31 provides the estimated biases and model based standard errors for the parameter estimates for these six designs.

Design	$b(\hat{\alpha}_2)$	$se(\hat{\alpha}_2)$	$b(\hat{\beta}_2)$	$se(\hat{\beta}_2)$
B1	-0.0113	0.0046	-0.0071	0.0032
B2	-0.0111	0.0040	-0.0035	0.0039
B3	-0.0092	0.0044	0.0001	0.0042
B4	-0.0094	0.0047	-0.0069	0.0043
B5	-0.0078	0.0045	0.0005	0.0043
B6	-0.0033	0.0048	0.0035	0.0046

Table 6.31: Estimated biases and standard errors from the Cox model for the designs in Group B.

For design B1, with 61% of the observations falling in the missing factor level categories, both $\hat{\alpha}_2$ and $\hat{\beta}_2$ were biased. $\hat{\alpha}_2$ was also biased for designs B2 and B3, although $\hat{\beta}_2$ was not. There was no evidence of bias for either of the parameter estimates when there were 30% or fewer of the observations in the missing levels.

TYPE III -- GROUP D

The numbers in the known cells for this group are given in Figure 6.23.

	F2	
F1	10	100
	50	40

Figure 6.23: Numbers in the known cells in Group D.

The numbers in the missing cells are shown in Table 6.32 and the biases and model based standard errors for the parameter estimates presented in Table 6.33.

Design	Number in missing cells					% missing
	n_{13}	n_{23}	n_{31}	n_{32}	n_{33}	
D1	50	50	5	50	5	44
D2	35	25	5	2	50	37
D3	5	10	15	25	35	31
D4	10	10	10	10	10	20
D5	5	5	5	5	5	11
D6	2	5	10	5	2	11
D7	5	2	5	2	5	9
D8	2	5	2	5	2	7
D9	5	2	2	2	5	7
D10	2	2	2	2	2	5

Table 6.32: Numbers in the missing cells and % missing in Group D.

Design	$b(\hat{\alpha}_2)$	$se(\hat{\alpha}_2)$	$b(\hat{\beta}_2)$	$se(\hat{\beta}_2)$
D1	-0.0718	0.0041	-0.0611	0.0051
D2	-0.0493	0.0045	-0.0385	0.0053
D3	-0.0409	0.0049	-0.0654	0.0050
D4	-0.0329	0.0049	-0.0448	0.0052
D5	-0.0156	0.0051	-0.0214	0.0054
D6	-0.0156	0.0052	-0.0268	0.0054
D7	-0.0097	0.0052	-0.0152	0.0055
D8	-0.0084	0.0052	-0.0108	0.0056
D9	-0.0058	0.0052	-0.0080	0.0056
D10	-0.0040	0.0052	-0.0069	0.0056

Table 6.33: Estimated biases and standard errors from the Cox model for the designs in Group D.

For all of the designs with 11% or more of the observations missing, both of the parameter estimates were biased. With 9% missing (D7), there was no evidence of bias for $\hat{\alpha}_2$, although $\hat{\beta}_2$ was biased. With only 7% or fewer of observations in the missing cells, there was no evidence of bias for either of the parameter estimates.

TYPE III – GROUP G

The numbers in the known and missing cells for this group are given in Figure 6.24 and Table 6.34 respectively.

	F2	
F1	2	90
	100	8

Figure 6.24: Numbers in the known cells in Group G.

	Number in missing cells					
Design	n_{13}	n_{23}	n_{31}	n_{32}	n_{33}	**% missing**
G1	2	2	2	100	100	51
G2	100	50	2	2	2	44
G3	15	15	15	15	15	27
G4	5	5	5	5	5	11
G5	2	2	2	2	2	5
G6	1	1	1	1	1	2

Table 6.34: Numbers in the missing cells and % missing in Group G.

The estimated biases and estimated standard errors for $\hat{\alpha}_2$ and $\hat{\beta}_2$ are shown in Table 6.35.

Design	$b(\hat{\alpha}_2)$	$se(\hat{\alpha}_2)$	$b(\hat{\beta}_2)$	$se(\hat{\beta}_2)$
G1	-0.1076	0.0092	-0.1152	0.0091
G2	-0.2222	0.0048	-0.2086	0.0062
G3	-0.2213	0.0069	-0.2348	0.0068
G4	-0.1231	0.0087	-0.1291	0.0087
G5	-0.0638	0.0097	-0.0682	0.0096
G6	-0.0327	0.0100	-0.0366	0.0100

Table 6.35: Estimated biases and standard errors from the Cox model for the designs in Group G.

All of the parameter estimates were biased in all situations.

COMPARISON OF RESULTS FOR THE COX AND EXPONENTIAL REGRESSION MODELS

An interesting observation is that the model based standard errors obtained from fitting the Cox models are virtually identical to those obtained for the exponential regression models (Section 6.3.5) fitted to all of the designs for the four groups. The standard errors for the Cox models have been assumed to be valid, as no checks have been made. That is, unlike for the exponential regression models (Section 6.3.4), there have been no simulations of large numbers of replicates to obtain an estimate of the true sampling variability for the Cox models.

For Group A, neither of the fits from the exponential regression model nor the Cox models produced any biases in the parameter estimates that were significantly different from zero. For Groups B and D, where some of the parameter estimates were biased, the biases were slightly larger for the Cox models than those for the exponential regression models, although it is not possible to determine if these are statistically different. It is interesting that the Cox model produced some evidence of bias in Group B even though the mixing parameters were equal, i.e. $r = s$ (Appendix 9). The parameter estimates were virtually identical for the fits of both the exponential and Cox models for Group G.

DISCUSSION

The general compatibility of the results for determining the parameter estimates from fitting an exponential regression model and a Cox model to sets of data which were generated from exponential distributions is probably because the Cox model is a generalisation of the exponential regression model. Thus, the hazard ratios obtained from these parameter estimates are similar for the two regression models. The standard errors from the two models were also very similar. It appears that there was no benefit gained from fitting the parametric exponential regression model to these data since the partial likelihood method of Cox, using the ranks of the deaths times, gave estimates that were as efficient at using the information as the maximum likelihood estimate technique used by the exponential regression model was.

However, this does not guarantee that there would not have been some benefit gained from fitting an exponential regression model had the survival estimates been examined instead of simply the parameter estimates. This is because the Cox model only models the parameter estimates and not the baseline hazard. This is needed to obtain the survival estimate. A parametric exponential model fitted to data generated from exponential distributions would probably estimate the baseline hazard, and hence the survival estimate, more efficiently.

It is necessary to remember that these results were for the large sample setting and it was assumed that there was no censoring. It is not clear whether similar results would have been observed for a small sample problem or if censoring had been taken into account in the Cox models.

6.5 GENERAL DISCUSSION

It was shown in Section 6.2.2 that the third level in the two factors exponential regression model did not satisfy the proportional hazards assumption when the third levels were assumed to be random mixtures of exponential distributions which were used to generate the observations falling in the known cells.

The general conclusion from Section 6.3.5, when exponential regression models were fitted to each of the designs, was that the bias of the parameter estimates were influenced by the inclusion of the extra levels for the unknown values.

However, there were several scenarios where there was no evidence of significant bias for the parameter estimates based, on the all cases models. These were:

(i) when the relevant proportions of observations in level 1 of factor F2 for both levels of factor F1 (Appendix 9) were very similar;

(ii) when the number of missing values as a percentage of the total numbers of cases was small. This was not true in all circumstances for all of the designs, especially when the proportions mentioned in (i) were very different (see the results for Group G in Section 6.3.5);

(iii) the distribution of the numbers of observations falling in the missing cells sometimes made a difference as to whether or not the parameter estimates were biased.

It would appear that differences among the numbers of observations in the four known cells in the designs, along with the appropriate mixing parameters used to obtain the observations in the missing cells, greatly influenced the results of the estimation procedure. For example, there was no evidence of bias in the parameter estimates for Groups A, B and E for any of the designs where the proportions were the same or very similar. This is in contrast to all of the designs in Group G, which showed evidence of bias for the parameter estimates.

The results may have been easier to interpret if the total number in the samples had been controlled, so that the variance was controlled. Alternatively, it may have been easier to detect patterns if the missing observations had been introduced into only one factor at a time before introducing them into both factors.

It is not clear whether the general conclusions given above based on the large sample sizes, apply in the small sample size context. The main observation from fitting exponential regression models to small sample sizes was that there was bias observed (Section 6.3.6).

From Section 6.4, it appears that in the simulated context examined in this thesis, the findings about the effect on the parameter estimates, in terms of bias, based on the exponential regression model, could be carried through to the Cox model, in general. Here, the data were created to satisfy the exponential regression model for the complete data context, but not for the missing cells.

However, there is uncertainty whether different results would be obtained from the two models if the complete data were generated to be non-exponential, but still with proportional hazards, rather than the constant hazards obtained for the exponential regression models, for the known cells. The uncertainty is because the exponential regression model assumes a particular parametric form for the baseline hazard function whereas the Cox model does not assume any distributional form for this and the only assumption made is that the hazards are proportional between the levels of a factor.

SECTION C :

FINAL THOUGHTS

CHAPTER 7 DISCUSSION AND CONCLUSIONS

7.1 SUMMARY OF KEY FINDINGS

FROM VARIOUS REGISTRY DATASETS

Breast cancer is a major health problem for women throughout the world, with 1 in 12 women getting the disease at some point in their lifetime. The incidence still appears to be rising, although there are some suggestions that the mortality in the UK is beginning to fall. Survival from the disease appears to be rising only slightly in Scotland, although the relative survival figures for Scotland are below average when compared to other European countries.

FROM THE BREAST CANCER AUDIT DATA

Chapter 4 showed that the presence of missing values in the clinical factors were associated with each other, with all two-way interactions, except one, present in the best fitting log-linear model. The conclusion from this model was that a woman was more likely to have a missing value in one of the four variables if she also had a missing value in another of the variables than if she had known information for that variable. The exception was for conditional independence between clinical stage and pathological node status, given the presence of ER status and tumour size. The clinical interpretation of this model appeared to imply that hospitals had agreed protocols, or at least informal practice agreements, for management of women with breast cancer.

Having discovered this pattern amongst these unknown values, different methods for handling missing values in models were discussed. The methods used in survival analyses of several breast cancer studies were then detailed and it was found that only the complete cases and additional categories methods were employed in the studies examined.

In Chapter 5, the results of the initial survival analysis reported by Twelves et al (1998a) were summarised, along with a discussion of the implications of the finding that there were significant different survival chances depending on which Health Board the women were treated in, but that there appeared to be no significant differences amongst the deprivation categories or by surgical case load in the Cox model. These findings are discussed in relation to other relevant literature. The results of Twelves et al (1998a) support the findings from other studies for the need for breast cancer to be managed in the setting of a multidisciplinary team. When it was checked whether any interactions of the clinical factors with the Health Board variable were significant, it was not possible to conclude that any were necessary and, therefore, none were included in the model.

Some model checking on the 'Clinical Full' model revealed that the proportional hazards assumption may be in question for ER status, with the increased hazard of death for women having ER negative tumours appearing to weaken over time. However, it is not entirely clear how to carry out the time-dependent modelling when then covariates are categorical with more than two levels, rather than binary or continuous. No unique method exists for this situation and so the results from the modelling performed have to be interpreted with caution.

The presence of missing values in some of the covariates gave rise to the possibility of drawing different conclusions from the results from fitting Cox regression models depending on whether, and how, these missing values were included in the models. Large absolute differences in survival estimates were observed in some of the tables presented in Section 5.4.2 which could considerably influence the interpretation of the findings.

FROM THE SIMULATION EXERCISE DATA

In Chapter 6, it was shown that including missing values using the additional categories method in an exponential regression model caused the proportional hazards assumption to be violated, when the missing values comprised of a random mixture of observations from exponential distributions. When this theoretical finding was examined empirically in simulation exercises, some of the parameter estimates obtained appeared to be biased due to fitting an incorrect model. Some of the estimated hazard ratios were different when exponential and Cox regression models were fitted, for the situations with and without the missing data categories (results in Sections 6.3.5 and 6.4). This is analogous to the results observed for the Cox modelling of the Breast Cancer Audit data (Section 5.4.2). There, it appeared that the hazard ratios for some of the Health Boards compared to Greater Glasgow Health Board were different for the complete cases and the all cases models.

7.2 FURTHER RESEARCH POSSIBILITIES

There are several areas of work that it would have been interesting to pursue had there been more time available. These include:

(i) modelling the Breast Cancer Audit data using some of the other methods for handling the missing values.

It would be interesting to examine the results from using other techniques discussed in Section 4.3 to find out whether the interpretations from the models were similar to, or different from, those given by the 'Clinical Full' model obtained from fitting a Cox model to the data using the additional categories method in the initial survival analysis.

(ii) fitting non-proportional hazards models to the Breast Cancer Audit data.

It would be worthwhile to investigate whether the suggested non-proportional hazards result for ER status from the time-dependent modelling exercise was reasonable by fitting some non-proportional hazards models to the data and comparing the results with the 'Clinical Full' model.

(iii) exploring the exponential regression and Cox modelling simulation exercises in more detail.

It would be useful to examine the findings of these simulation exercises if censoring had been incorporated into the design or if the artificial datasets had been created to have proportional hazards which were not generated from exponential distributions.

(iv) investigating other structures than the random mixtures assumed in this thesis for the missing data would also be an interesting exercise to simulate and undertake.

(v) studying a threshold for the amount of missing values acceptable in a Cox model in terms of clinical significance.

The main message taken from Section 5.4.2 when the all cases (ACM) and the complete cases (CCM) models were compared was that the interpretations of the results for these models appeared to be very different, although it was not possible to formally test for statistical differences between the sets of results. A great deal of fluctuation was observed amongst the parameter estimates for the Health Boards obtained for the two cohorts and also for the 5-year % survival estimates for different Health Boards among different prognostic groups.

Here, it was seen that having 64% of cases with some missing information in one of the four main clinical factors caused differences in results to be clinically significant (when compared to magnitudes observed in clinical trials for beneficial treatments; Section 2.3.2). It is not clear, however, exactly how much missing data was needed to observe these clinically significant results, nor whether the results were mainly affected by the introduction of missing values in only one variable in particular, or in any of them. However, from the work given in Chapter 6, it was found that the influence on the bias of the missing values depended on the context, as well as the overall percentage missing and the distribution of the missing values. In some designs, no bias was observed with the introduction of missing values.

As an alternative to the simulations given in Chapter 6, it would be interesting to perform a sensitivity analysis to examine the effect of altering the percentage of missing values in some or all of the factors in the 'Clinical Full' model for the Breast Cancer Audit data. It would be possible to approach this in a couple of ways. The aim in all of the techniques suggested below would be to identify at what point the differences became clinically non-significant, thus highlighting a threshold whereby the inclusion of missing values cease to be important. At this point, it would not matter whether or not the missing values were included in the model. The threshold identified could then be applied to similar data-sets for breast cancer, and the method used applied in survival analyses of other cancers. From Chapter 6, this may not be a straightforward exercise.

USING THE COMPLETE CASES COHORT AS THE BASELINE

Initially the complete cases model (Section 5.4.2) would be fitted and these results used as the baseline for all comparisons. One reason for doing this is that the Cox model is theoretically correct as the proportional hazards assumption holds. Here, it is unimportant that this is a sub-population of the total cohort as this is being used as the baseline. Various different strategies could be employed.

(a) Adding a proportion of cases with missing values, without being concerned about which variables are missing. This could be done by random simulation of cases, with say, 5%, 10%, 25%, 50%, 75% and 100% with missing values in extra cases introduced with the complete cases.

(b) Another approach would be to start with one variable only, say ER status, and firstly examine the complete cases analysis parameter estimates; then keep only cases with ER status known but allow missing values in other variables and examine parameter estimates; then keep only cases where other variables known but allow ER status to be missing; and finally the all cases model. This method is similar to the two partial cases models fitted in Section 5.4.2.

(c) An extension of (b), whereby cases which have complete information on a subset of variables are introduced and the effect of altering the combinations of variables with known and missing information compared.

REFERENCES

LIST OF REFERENCES

Armitage P & Berry G (1994). *Statistical Methods in Medical Research*, 3rd edn. Blackwell Scientific Publications: Oxford.

Barnes DM, Harris WH, Smith P, Millis RR, Rubens RD (1996). Immunohistochemical determination of oestrogen-receptor: comparison of different methods of assessment of staining and correlation with clinical outcome of breast cancer patients. *Br J Cancer* **74:** 1445-1451.

Basnett I, Gill M, Tobias JS (1992). Variations in breast cancer management between a teaching and a non-teaching district. *Eur J Cancer* **28A:** 1945-1950.

Berrino F, Sant M, Verdecchia A, Capocaccia R, Hakulinen T, Esteve J Eds (1995). *Survival of Cancer Patients in Europe: The EUROCARE Study*. IARC Scientific Publications No. 132. IARC: Lyon.

Black RJ, Sharp L, Kendrick SW (1993). *Trends in Cancer Survival in Scotland 1968-1990.* ISD Scotland Publications: Edinburgh.

Blamey RW, Davies CJ, Elston CW, Johnson J, Haybittle JL, Maynard PV (1979). Prognostic factors in breast cancer - the formation of a prognostic index. *Clinical Oncology* **5:** 227-236.

Bohlke K, Spiegelman D, Trichopoulou A, Katsouyanni K, Trichopoulos D (1999). Vitamins A, C and E and the risk of breast cancer: results from a case-control study in Greece. *Br J Cancer* **79:** 23-29.

Boyle P, Leake R (1988). Progress in understanding breast cancer: epidemiological and biological interactions. *Breast Cancer Research and Treatment* **11:** 91-112.

Boyle P, Parkin DM (1991). Statistical Methods for Registries in *Cancer Registration: Principles and Methods* Eds. Jensen OM, Parkin DM, Maclennan R, Muir CS, Skeet RG. IARC Scientific Publications No 95. IARC: Lyon.

Brewster D, Crichton J, Muir C (1994). How accurate are Scottish cancer registration data? *Br J Cancer* **70:** 954-959.

Brewster D, Everington D, Harkness E, Gould A, Warner J, Dewar JA, Arrundale J (1996a). Incidence of and mortality from breast cancer since introduction of screening - Scottish figures show higher incidence and similar mortality (letter). *Br Med J* **312:** 639-640.

Brewster DH, Stroner PL, Young J, on behalf of the Scottish Breast Cancer Focus Group (1996b). Recording of prognostic information in the case notes of women with breast cancer in Scotland: implications for population-based comparative survival analyses (abstract). In: International Association of Cancer Registries 30th Annual Meeting Abstracts. ISD Publications: Edinburgh.

Brewster D, Crichton J, Harvey JC, Dawson G (1997). Completeness of case ascertainment in a Scottish regional cancer registry for the year 1992. *Public Health* **111:** 339-343.

Buell P (1973). Changing incidence of breast cancer in Japanese-American women. *J Natl Cancer Inst* **51:** 1479-1483.

Bundred NJ, Morgan DAL, Dixon JM (1994). Management of regional nodes in breast cancer. *Br Med J* **309:** 1222-1225.

Cancer Research Campaign (1996). Breast cancer - UK (CRC Factsheets Nos 6). Cancer Research Campaign.

Carnon AG, Ssemwogerere A, Lamont DW, Hole DJ, Mallon EA, George WD, Gillis CR (1994). Relation between socio-economic deprivation and pathological prognostic factors in women with breast cancer. *Br Med J* **309:** 1054-1057.

Carstairs V, Morris R (1991). *Deprivation and Health in Scotland.* Aberdeen University Press: Aberdeen.

Carter CL, Allen C, Henson DE (1989). Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases. *Cancer* **63:** 181-187.

Cassidy A (1996). Breast cancer: weighing up the dietary evidence. *The Breast* **5:** 389-397.

Collaborative Group on Hormonal Factors in Breast Cancer (1996). Breast cancer and hormonal contraceptives: collaborative reanalysis of individual data on 53 297 women with breast cancer and 100 239 women without breast cancer from 54 epidemiological studies. *Lancet* **347:** 1713-1727.

Collaborative Group on Hormonal Factors in Breast Cancer (1997). Breast cancer and hormone replacement therapy: collaborative reanalysis of data from 51 epidemiological studies of 52 705 women with breast cancer and 108 411 women without breast cancer. *Lancet* **350:** 1047-1059.

Collett D (1994). *Modelling Survival Data in Medical Research.* Chapman & Hall: London.

Collett K, Skjaerven R, Maehle BO (1998). The prognostic contribution of estrogen and progesterone receptor status to a modified version of the Nottingham Prognostic Index. *Breast Cancer Research and Treatment* **48:** 1-9.

Cox DR (1972). Regression models and life-tables (with discussion). *J R S Soc B* **34:** 187-220.

Cox DR, Oakes D (1984). *Analysis of Survival Data.* Chapman & Hall: London.

Dewar JA, Twelves CJ, Thomson CS (1999). Breast cancer in Scotland: changes in treatment and workload. *Clin Oncol* **11:** 52-54.

Dobson AJ (1990). *An Introduction to Generalized Linear Models,* 2nd edn. Chapman & Hall: London.

Doll R, Payne P, Waterhouse J Eds (1966). *Cancer Incidence in Five Continents, Vol I. A Technical Report.* UICC, Springer-Verlag: Berlin.

Early Breast Cancer Trialists' Collaborative Group (1992). Systemic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy. *Lancet* **339:** 1-15, 71-84.

Early Breast Cancer Trialists' Collaborative Group (1995). Effects of radiotherapy and surgery in early breast cancer: an overview of the randomised trials. *N Engl J Med* **333:** 1444-1455.

Early Breast Cancer Trialists' Collaborative Group (1996). Ovarian ablation in early breast cancer: overview of the randomised trials. *Lancet* **348:** 1189-1196.

Early Breast Cancer Trialists' Collaborative Group (1998a). Tamoxifen for early breast: an overview of the randomised trials. *Lancet* **351:** 1451-1467.

Early Breast Cancer Trialists' Collaborative Group (1998b). Polychemotherapy for early breast cancer: an overview of the randomised trials. *Lancet* **352:** 930-942.

Ederer F, Axtell LM, Cutler SJ (1961). The relative survival rate: a statistical methodology. *Natl Cancer Instit* **6:** 101-121.

Evans DGR, Fentiman IS, McPherson K, Asbury D, Ponder BAJ, Howell A (1994). Familial breast cancer. *Br Med J* **308:** 183-187.

Ewertz M, Gillanders S, Meyer L, Zedeler K (1991). Survival of breast cancer patients in relation to factors which affect the risk of developing breast cancer. *Int J Cancer* **49:** 526-530.

Expert Advisory Group on Cancer (1995). *A Policy Framework for Commissioning Cancer Services.* Department of Health, London.

Ferguson DJP, Anderson TJ (1981). Morphological evaluation of cell turnover in relation to the menstrual cycle in the "resting" human breast. *BrJ Cancer* **44:** 177-181.

Ford I, Norrie J, Ahmadi S (1995). Model inconsistency, illustrated by the Cox proportional hazards model. *Stat Med* **14:** 735-746.

Freedman LS, Edwards DN, McConnell EM, Downham DY (1979). Histological grade and other prognostic factors in relation to survival of patients with breast cancer. *Br J Cancer* **40:** 44-55.

General Registers Office for Scotland (1996). *Annual Report 1995.* Government Statistical Service: Edinburgh.

Gillis CR, Hole DJ (1996). Survival outcome of care by specialist surgeons in breast cancer: a study of 3786 patients in the West of Scotland. *Br Med J* **312:** 145-148.

Giuffrida D, Lupo L, La Porta GA, La Rosa GL, Padova G, Foti E, Marchese V, Belfiore A (1992). Relation between steroid receptor status and body weight in breast cancer patients. *Eur J Cancer* **28:** 112-115.

Gordon NH, Crowe JP, Brumberd DJ, Berger NA (1992). Socioeconomic factors and race in breast cancer recurrence and survival. *Am J Epidemiol* **135:** 609-618.

Gore SM, Pocock SJ, Kerr GR (1984). Regression models and non-proportional hazards in the analysis of breast cancer survival. *Appl Statist* **33:** 176-195.

Greenland S, Finkle WD (1995). A critical look at methods for handling missing covariates in epidemiological regression analyses. *Am J Epidemiol* **142:** 1255-1264.

Gregory WM, Smith P, Richards MA, Twelves CJ, Knight RK, Rubens RD (1993). Chemotherapy of advanced breast cancer: outcome and prognostic factors. *Br J Cancer* **68:** 988-995.

Gregory WM, Bolland K, Whitehead J, Souhami RL (1997). Cautionary tales of survival analysis: conflicting analyses from a clinical trial in breast cancer. *Br J Cancer* **76:** 551-558.

Harris V, Sandridge AL, Black RJ, Brewster DB, Gould A (1998). *Cancer Registration in Scotland 1986-1995.* ISD Scotland Publications: Edinburgh.

Hauck WW & Miike R (1991). A proposal for examining and reporting stepwise regressions. *Stat Med* **10:** 711-715.

Hawkins RA, Tesdale AL, Killen ME, Jack WJL, Chetty U, Dixon JM, Hulme MJ, Prescott RJ, McIntyre MA, Miller WR (1996). Prospective evaluation of prognostic factors in operable breast cancer. *Br J Cancer* **74:** 1469-1478.

Haybittle J, Houghton J, Baum M (1997). Social class and weight as prognostic factors in early breast cancer. *Br J Cancer* **75:** 729-733.

Henderson BE, Ross RK, Judd HL, Krailo MD, Pike MC (1985). Do regular ovulatory cycles increase breast cancer risk? *Cancer* **56:**1206-1208.

Henderson BE, Pike MC, Bernstein L, Ross RK (1996). Breast cancer in *Cancer Epidemiology and Prevention*, 2nd edn. Eds: Schottenfeld D, Fraumeni JF Jr. Oxford University Press: Oxford.

Hoel DG, Wakabayashi T, Pike MC (1983). Secular trends in the distributions of the breast cancer risk factors - menarche, first birth, menopause, and weight - in Hiroshima and Nagasaki, Japan. *Am J Epidemiol* **118:** 78-89.

Howe GR, Hirohata T, Hislop TG, Iscovich JM, Yuan J-M, Katsouyanni K, Lubin F, Marubini E, Modan B, Rohan T, Toniolo P, Shunzhang Y (1990). Dietary factors and risk of breast cancer: combined analysis of 12 case-control studies. *J Natl Cancer Inst* **82:** 561-569.

Hsieh C-C, Trichopoulos D, Katsouyanni K, Yuasa S (1990). Age at menarche, age at menopause, height and obesity as risk factors for breast cancer: associations and interactions in an international case-control study. *Int J Cancer* **46:** 796-800.

Kalbfleisch JD & Prentice RL (1980). *The Statistical Analysis of Failure Time Data.* Wiley: New York.

Karjalainen S, Pukkala E (1990). Social class as a prognostic factor in breast cancer survival. *Cancer* **66:** 819-826.

Kendrick S, Clarke J (1993). The Scottish record linkage system. *Health Bulletin* **51:** 72-79.

Leonard RCF, Rodger A, Dixon JM (1994). Metastatic breast cancer. *Br Med J* **309:** 1501-1504.

Lipworth L (1995). Epidemiology of breast cancer. *Eur J Can Prev* **4:** 7-30.

Little RJA (1992). Regression with missing X's: a review. *Journal of the American Statistical Association* **87:** 1227-1237.

Little RJA, Rubin DB (1987). *Statistical Analysis with Missing Data.* Wiley: New York.

Longnecker MP (1994). Alcoholic beverage consumption in relation to risk of breast cancer: meta-analysis and review. *Cancer Causes and Control* **5:** 73-82.

MacMahon B, Cole P, Lin TM, Lowe CR, Mirra AP, Ravnihar B, Salber EJ, Valaoras VG, Yuasa S (1970). Age at first birth and breast cancer risk. *Bull Wld Hlth Org* **43:** 209-221.

Maudsley, G, Williams, EMI (1993). Death certification by House Officers and General Practioners - practice and performance. *J Pub Health Med* **15:** 192-201.

Michnovicz JJ, Hershcopf RJ, Naganuma H, Bradlow HL, Fishman J (1986). Increased 2-hydroxylation of estradiol as a possible mechanism for the anti-estrogenic effect of cigarette smoking. *N Engl J Med* **315:** 1305-1309.

Miller WR, Ellis IO, Sainsbury JRC, Dixon JM (1994). Prognostic factors. *Br Med J* **309:** 1573-1576.

Newman SC, Lees AW, Jenkins HJ (1997). The effect of body mass index and oestrogen receptor level on survival of breast cancer patients. *Int J Epidemiol* **26:** 484-490.

Office for Population and Census Statistics (1975). *Great Britain. Economic Activity. Part IV (10% sample).* HMSO: London.

Parkin DM, Pisani P, Ferlay J (1993). Estimates of the worldwide incidence of eighteen major cancers in 1985. *Int J Cancer* **54:** 594-606.

Parkin DM, Whelan SL, Ferlay J, Raymond L, Young J Eds (1997). *Cancer Incidence in Five Continents, Vol VII.* IARC Scientific Publications No. 143. IARC: Lyon.

Peto R (1998). Mortality from breast cancer in UK has decreased suddenly (letter). *Br Med J* **317:** 476-477.

Pujol P, Daures J-P, Thezenas S, Guilleux F, Rouanet P, Grenier J (1998). Changing estrogen and progesterone receptor patterns in breast carcinoma during the menstrual cycle and menopause. *Cancer* **83:** 698-705.

Quinn M, Allen E on behalf of the United Kingdom Association of Cancer Registries (1995). Changes in incidence of and mortality from breast cancer in England and Wales since introduction of screening. *Br Med J* **311:** 1391-1395.

Ramirez AJ, Towlson KE, Leaning MS, Richards MA, Rubens RD (1998). Do patients with advanced breast cancer benefit from chemotherapy? *Br J Cancer* **78:** 1488-1494.

Richards MA, Smith IE, Dixon JM (1994). Role of systemic treatment for primary operable breast cancer. *Br Med J* **309:** 1363-1366.

Richards MA, Parrott JC (1996). Tertiary cancer services in Britain: benchmarking study of activity and facilities at 12 specialist centres. *Br Med J* **313:** 347-349.

Richards MA, Wolfe CDA, Tilling K, Barton J, Bourne HM, Gregory WM (1996). Variations in the management and survival of women under 50 years with breast cancer in the South East Thames region. *Br J Cancer* **73:** 751-757.

Richards M, Sainsbury R, Kerr D (1997). Inequalities in breast cancer care and outcome. *Br J Cancer* **76**: 634-638.

Rodger A, Leonard RCF, Dixon JM (1994). Locally advanced breast cancer. *Br Med J* **309:** 1431-1433.

Sainsbury JRC, Anderson TJ, Morgan DAL, Dixon JM (1994). Breast cancer. *Br Med J* **309:** 1150-1153.

Sainsbury R, Haward B, Rider L, Johnston C, Round C (1995a). Influence of clinician workload and patterns of treatment on survival from breast cancer. *Lancet* **345:** 1265-1270

Sainsbury JRC, Rider L, Smith A, McAdam WFA (1995b). Does it matter where you live? Treatment variation for breast cancer in Yorkshire. *Br J Cancer* **71:** 1275-1278.

Sant M, Capocaccia R, Verdecchia A, Gatta G, Micheli A, Coleman MP, Berrino F and the EUROCARE Working Group (1998). Survival of women with breast cancer in Europe: variation with age, year of diagnosis and country. *Int J Cancer* **77:** 679-683.

Schemper M (1992). Cox analysis of survival data with non-proportional hazard functions. *The Statistician* **41:** 455-465.

Schemper M, Heinze G (1997). Probability imputation revisited for prognostic factor studies. *Stat Med* **16:** 73-80.

Schemper M, Smith TL (1990). Efficient evaluation of treatment effects in the presence of missing covariate values. *Stat Med* **9:** 777-784.

Schluchter MD, Jackson KL (1989). Log-linear analysis of censored survival data with partially observed covariates. *Journal of the American Statistical Association* **84:** 42-52.

Schrijvers CTM, Mackenbach JP, Lutz J-M, Quinn MJ, Coleman MP (1995). Deprivation and survival from breast cancer. *Br J Cancer* **72:** 738-743.

Scottish Breast Cancer Focus Group, Scottish Cancer Trials Breast Group, Scottish Cancer Therapy Network (1996). *Scottish Breast Cancer Audit 1987 and 1993*. Scottish Cancer Therapy Network: Edinburgh.

Scottish Breast Screening Programme Central Co-ordinating Unit (1997). *The Scottish Breast Screening Programme Report 1996*. Scottish Breast Screening Programme: Edinburgh.

Scottish Cancer Co-ordinating and Advisory Committee (1996). *Commissioning Cancer Services in Scotland.* The Scottish Office: Edinburgh.

Scottish Intercollegiate Guidelines Network and the Scottish Cancer Therapy Network (1998). *Breast Cancer in Women.* Pilot edition October 1998. Publication No 29. Edinburgh.

Sharp L, Black R, Harkness EF, Finlayson AR, Muir CS (1993). *Cancer Registration Statistics Scotland 1981-1990.* ISD Scotland Publications: Edinburgh.

Shek LLM & Godolphin W (1988). Model for breast cancer survival: relative prognostic roles of axillary nodal status, TNM stage, estrogen receptor concentration, and tumor necrosis. *Cancer Research* **48:** 5565-5569.

Shimizu H, Ross RK, Bernstein L, Yatani R, Henderson BE, Mack TM (1991). Cancers of the prostate and breast among Japanese and white immigrants in Los Angeles County. *Br J Cancer* **63:** 963-966.

Souhami R, Tobias J (1995). *Cancer and Its Management,* 2nd edn. Blackwell Science: Oxford.

Stiller CA (1994). Centralised treatment, entry to clinical trials and survival. *Br J Cancer* **70:** 352-362.

Stoll BA (1996). Diet and exercise regimens to improve breast carcinoma prognosis. *Cancer* **78:** 2465-2470.

Tate HC, Rawlinson JB, Freedman LS (1979). Randomised comparative studies in the treatment of cancer in the United Kingdom: room for improvement? *Lancet* **ii:** 623-625.

Tavani A, Braga C, La Vecchia C, Negri E, Russo A, Franceschi S (1997). Attributable risks for breast cancer in Italy: education, family history and reproductive and hormonal factors. *Int J Cancer* **70:** 159-163.

Titus-Ernstoff L, Longnecker MP, Newcomb PA, Dain B, Greenberg ER, Mittendorf R, Stampfer M, Willett W (1998). Menstrual factors in relation to breast cancer risk. *Cancer Epidemiology, Biomarkers & Prevention* **7:** 783-789.

Tokunaga M, Land CE, Tokuoka S, Nishimori I, Soda M, Akiba S (1994). Incidence of female breast cancer among atomic bomb survivors, 1950-1985. *Radiat. Res.* **138:** 209-223.

Trichopoulos D, MacMahon B, Cole P (1972). Menopause and breast cancer risk. *J Natl Cancer Inst* **48:** 605-613.

Twelves CJ, Thomson CS, Gould A, Dewar JA (1998a). Variation in the survival of women with breast cancer in Scotland. *Br J Cancer* **78:** 566-571.

Twelves CJ, Thomson, CS, Young J, Gould A (1998b). Entry into clinical trials in breast cancer: the importance of specialist teams. *Eur J Cancer* **34:** 1004-1007.

Twelves CJ, Thomson CS, Dewar JA (1999). Social factors affect patterns of referral for breast cancer (letter). *Br Med J* **318:** 326.

Union Internationale Contre Le Cancer (UICC) (1987). *TNM Classification of Malignant Tumours*, 4th edn. Springer-Verlag: Berlin.

Vach W, Blettner M (1991). Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. *Am J Epidemiol* **134:** 895-907.

Vach W, Blettner (1995). Logistic regression with incompletely observed categorical covariates - investigating the sensitivity against violation of the missing at random assumption. *Stat Med* **14:** 1315-1329.

Vach W (1997). Some issues in estimating the effect of prognostic factors from incomplete covariate data. *Stat Med* **16:** 57-72.

Vatten, LJ (1996). Body size and breast cancer risk. *Breast* **5:** 5-9.

World Health Organisation (1976). *International Classification of Diseases for Oncology,* 1st edn. World Health Organisation: Geneva.

World Health Organisation (1977). *Manual of the International Classification of Diseases, Injuries and Causes of Death,* 9th revn. HMSO: London.

World Health Organisation (1996). *World Health Statistics Annual 1995.* World Health Organisation: Geneva.

APPENDICES

APPENDICES

APPENDIX 1 VARIABLES COLLECTED IN THE BREAST CANCER AUDIT

The data collected at each of these stages of management of breast cancer are discussed separately. The list given below provides an idea of the information collected at each stage and is not exhaustive.

Referral Patterns: These included the date the woman saw her GP, the date the woman first saw a surgeon, the hospital of initial referral, the date the woman was first seen by an oncologist and whether the cancer was detected as part of the Screening Programme.

Initial Staging Information: This was collected at the clinic and involved collecting the clinical TNM stage and also the menstrual status of the women. A fine needle aspiration could also be performed in the clinic to help to decide whether the lump was malignant.

Surgical Procedures: Usually some form of surgery was needed to be able to give a definitive diagnosis of breast cancer. This could just be a biopsy to investigate whether the lump was cancerous or could be definitive surgery to remove either just the tumour, in breast conservation, or the whole breast, in a mastectomy. Also given were details relating to whether surgery was performed on the axilla to remove lymph nodes and whether the ovaries were surgically removed. The surgeon performing the operation and at which hospital the operation took place were recorded.

Other Forms of Treatment: Details were given relating to the hormone treatment administered, including any ovarian suppression, along with any chemotherapy regimens the woman was started on. The site as well as dates were given for any radiotherapy the women received.

Pathology Details: This information was normally included as a separate report in the case notes, having been sent from the pathology laboratory. The size of the tumour excised was given, when measured in the laboratory, along with details of any tumour involvement in the margins of the tissue removed. The number of nodes that were found in the sample or clearance of the axilla was given together with the number of nodes with tumour involvement. Histological grading information was also reported, as was the ER status which was determined through several different techniques, whose continuous scores cannot be combined. Thus, either positive or negative was also given for ER status, as well as the score from the particular assay.

Follow-up and Outcome Information: Dates and sites of the first local, regional and distant recurrences were recorded. Also given were details of any clinical trials into which the woman had been entered. The status (alive or dead) was noted, with either the date of death or the date last seen recorded.

APPENDIX 2 KEY TO THE HEALTH BOARD CODES

Table A2.1 gives the codes that are usually given for the Health Boards in most NHSiS documents. Due to the small numbers of women treated in Orkney, Shetland and Western Isles, these Health Boards are represented here by the 'Health Board', 'the Islands', to represent off-mainland treatment.

Health Board Label	Health Board
A	Ayrshire & Arran
B	Borders
C	Argyll & Clyde
F	Fife
G	Greater Glasgow
H	Highland
I	Islands
L	Lanarkshire
N	Grampian
S	Lothian
T	Tayside
V	Forth Valley
Y	Dumfries & Galloway

Table A2.1: Key to the Health Boards.

APPENDIX 3 BREAKDOWNS OF THE VARIABLES USED IN THE ANALYSIS OF THE BREAST CANCER AUDIT

In all of the tables, 'No.' stands for number.

Clinical Variables:

Age group	No.	%		Clinical stage	No.	%		ER status	No.	%
< 50 years	476	29.4		I	302	18.7		Positive	599	37.0
50 - 64	591	36.5		II	813	50.2		Negative	391	24.2
65 - 79	480	29.6		III	187	11.6		Not known	629	38.9
≥ 80 years	72	4.4		Not known	317	19.6				

Table A3.1: Numbers and percentages for each of the clinical variables.

Node status	No.	%		Tumour size	No.	%
Positive	583	36.0		≤ 2 cm	625	38.6
Negative	601	37.1		> 2 cm	662	40.9
Not known	435	26.9		Not known	332	20.5

Table A3.1 cont: Numbers and percentages for each of the clinical variables.

Treatment Variables:

Adjuvant endocrine therapy	No.	%		Adjuvant chemotherapy	No.	%		Adjuvant radiotherapy	No.	%
Given	1052	65.0		Given	123	7.6		Given	660	40.8
Not given	567	35.0		Not given	1496	92.4		Not given	959	59.2

Table A3.2: Numbers and percentages for each of the treatment variables.

Type of surgery	No.	%		Adjuvant chemotherapy or endocrine therapy	No.	%
Mastectomy	976	60.3		Given	1138	70.3
Conservation	643	39.7		Not given	481	29.7

Table A3.2 cont: Numbers and percentages for each of the treatment variables.

Service Variables:

Deprivation group	No.	%		Referral to oncologist	No.	%		Surgeon case load	No.	%
Least deprived	404	25.0		No	738	45.6		1 - 9 cases	278	17.2
Intermediate	982	60.7		Yes	852	52.6		10 - 29	683	42.2
Most deprived	233	14.4		Not known	29	1.8		Team / 30+	647	40.0
								Not known	11	0.7

Table A3.3: Numbers and percentages for each of the service variables. Note that 'no' for referral to an oncologist included those women who saw an oncologist after three months of diagnosis. Note that least deprived is the first quintile, intermediate deprivation group includes quintiles II, III and IV and most deprived is the last quintile.

Health Board	No.	%
A	126	7.8
B	22	1.4
C	107	6.6
F	91	5.6
G	343	21.2
H	72	4.4
I	25	1.5
L	135	8.3
N	186	11.5
S	235	14.5
T	148	9.1
V	68	4.2
Y	61	3.8

Table A3.3 cont: Numbers and percentages for Health Board.

APPENDIX 4 CROSS-TABULATIONS OF THE PAIRWISE CLINICAL VARIABLES AND THE CLINICAL VARIABLES BY SURGEON CASE LOAD

In all of the tables, 'NK' stands for unknown.

Cross-tabulations of the pairwise clinical variables.

	C I	C II	CIII	C NK	Total	Total Number
A <50	25.2	49.8	7.1	17.9	100	476
A 50-64	18.1	51.4	10.8	19.6	100	591
A 65-79	13.3	50.4	16.0	20.2	100	480
A ≥ 80	15.3	41.7	16.7	28.4	100	72
Total	18.7	50.2	11.6	19.6	100	1619

Table A4.1: Percentages of clinical stage (C) by age (A).

	E +	E -	E NK	Total	Total Number
A <50	34.5	30.9	34.7	100	476
A 50-64	40.8	25.2	34.0	100	591
A 65-79	36.9	19.4	43.8	100	480
A ≥ 80	23.6	2.8	73.6	100	72
Total	37.0	24.2	38.9	100	1619

Table A4.2: Percentages of ER status (E) by age (A).

	N +	N -	N NK	Total	Total Number
A <50	37.2	43.1	19.7	100	476
A 50-64	39.1	40.1	20.8	100	591
A 65-79	34.0	31.3	34.8	100	480
A ≥ 80	16.7	12.5	70.8	100	72
Total	36.0	37.1	26.9	100	1619

Table A4.3: Percentages of node status (N) by age (A).

	$T \leq 2$	$T > 2$	T NK	Total	Total Number
A <50	39.1	34.2	26.7	100	476
A 50-64	40.1	41.6	18.3	100	591
A 65-79	36.9	45.6	17.5	100	480
$A \geq 80$	34.7	47.2	18.1	100	72
Total	38.6	40.9	20.5	100	1619

Table A4.4: Percentages of tumour size (T) by age (A).

	E +	E -	E NK	Total	Total Number
C I	39.1	20.2	40.7	100	302
C II	40.5	26.7	32.8	100	813
C III	34.2	28.9	36.9	100	187
C NK	27.8	18.6	53.6	100	317
Total	37.0	24.2	38.9	100	1619

Table A4.5: Percentage of ER status (E) by clinical stage (C).

	N +	N -	N NK	Total	Total Number
C I	17.2	47.4	35.4	100	302
C II	38.3	39.5	22.3	100	813
C III	64.2	15.0	20.9	100	187
C NK	31.5	34.4	34.1	100	317
Total	36.0	37.1	26.9	100	1619

Table A4.6: Percentages of node status (N) by clinical stage (C).

	$T \leq 2$	$T > 2$	T NK	Total	Total Number
C I	59.9	17.5	22.5	100	302
C II	36.2	46.9	17.0	100	813
C III	16.0	68.4	15.5	100	187
C NK	37.9	31.5	30.6	100	317
Total	38.6	40.9	20.5	100	1619

Table A4.7: Percentages of tumour size (T) by clinical stage (C).

	N +	N -	N NK	Total	Total Number
E +	39.6	40.7	19.7	100	599
E -	40.4	38.4	21.2	100	391
E NK	29.9	32.9	37.2	100	629
Total	36.0	37.1	26.9	100	1619

Table A4.8: Percentages of node status (N) by ER status (E).

	T ≤ 2	T > 2	T NK	Total	Total Number
E +	44.6	39.1	16.4	100	599
E -	33.5	47.6	18.9	100	391
E NK	36.1	38.5	25.4	100	629
Total	38.6	40.9	20.5	100	1619

Table A4.9: Percentages of tumour size (T) by ER status (E).

	T ≤ 2	T > 2	T NK	Total	Total Number
N +	29.3	53.5	17.2	100	583
N -	44.8	35.3	20.0	100	601
N NK	42.5	31.7	25.7	100	435
Total	38.6	40.9	20.5	100	1619

Table A4.10: Percentages of tumour size (T) by node status (N).

Cross-tabulations of the clinical variables with surgeon case load.

* Note (for Tables A4.11 to A4.15) that for 11 women, the surgeon performing the operation was not recorded in the case notes. Therefore, the case load of the surgeon was unknown and not included in all analyses.

	A <50	A 50-64	A 65-79	A ≥ 80	Total Number
S 1-9	28.8	33.8	30.2	7.2	278
S 10-29	23.9	35.1	35.6	5.4	683
S Team /30+	35.5	39.4	22.9	2.2	647
Total	29.4	36.6	29.5	4.4	1608*

Table A4.11: Percentages of age (A) by surgeon case load (S).
The P value for the test of association was <0.001.

	C I	C II	C III	C NK	Total Number
S 1-9	16.9	46.4	10.4	26.3	278
S 10-29	19.0	45.4	11.9	23.7	683
S Team /30+	19.2	57.0	11.7	12.1	647
Total	18.7	50.2	11.6	19.5	1608*

Table A4.12: Percentage of clinical stage (C) by surgeon case load (S). The P value for the test of association was <0.001.

	E +	E -	E NK	Total Number
S 1-9	25.9	15.8	58.3	278
S 10-29	28.8	22.7	48.5	683
S Team /30+	50.7	29.2	20.1	647
Total	37.1	24.1	38.7	1608*

Table A4.13: Percentages of ER status (E) by surgeon case load (S). The P value for the test of association was <0.001.

	N +	N -	N NK	Total Number
S 1-9	37.8	32.0	30.2	278
S 10-29	31.5	31.5	37.0	683
S Team /30+	40.3	45.7	13.9	647
Total	36.1	37.3	26.6	1608*

Table A4.14: Percentages of node status (N) by surgeon case load (S). The P value for the test of association was <0.001.

	$T \leq 2$	$T > 2$	T NK	Total Number
S 1-9	34.9	42.1	23.0	278
S 10-29	38.7	42.2	19.2	683
S Team /30+	40.6	39.3	20.1	647
Total	38.8	41.0	20.2	1608*

Table A4.15: Percentages of tumour size (T) by surgeon case load (S). The P value for the test of association was 0.42.

APPENDIX 5 STANDARD ERROR FOR THE SURVIVAL ESTIMATE FROM COX REGRESSION USING THE SPSS STATISTICS PACKAGE

This apparent problem was identified using Version 9.0 of the SPSS statistics package during this research when the standard errors for some of the HB, prognostic factors combinations were obtained and confidence intervals (CIs) given for the 5-year survival estimates (Section 5.4.2). The first two examples illustrate why there appears to be some uncertainty regarding the estimate of the standard error.

Example (i): From Table 5.20b in Section 5.4.2, the 5-year % survival estimates for the group E+, N+, T>2 for Health Boards G and T were 76.9% and 61.1% respectively. The corresponding standard errors were 2.03% and 3.03% respectively. Thus, the 95% CIs for survival are for G: (72.9%, 80.9%) and for T: (55.2%, 67.0%). These CIs do not overlap and so informally it appears that there are significant differences between these two Health Boards.

However, the hazard ratio for HB T vs G is 1.87 (95% CI: 0.90, 3.89), which implies that HB T is not significantly worse than HB G.

Example (ii): From Table 5.20b, the 5-year % survival estimates for the group E-, N+, $T \leq 2$ for Health Boards G and Y were 46.0% and 87.1% respectively. The corresponding standard errors were 3.59% and 1.21% respectively. Thus, the 95% CIs are for G: (39.0%, 53.0%) and for Y: (79.3%, 89.5%). These CIs do not overlap by a wide margin, and so informally it appears that there are significant differences between these two Health Boards.

However, the hazard ratio for HB Y vs G is 0.18 (95% CI: 0.02, 1.29), which implies that HB Y is not significantly worse than HB G.

To investigate this apparent inconsistency, SAS Version 6.12 was used to compare the results. However, due to the difficulties of fitting categorical factors in SAS (especially with HB having 13 levels and an interaction being present in the 'Clinical

Full' model), a much simpler situation was considered where only binary variables were modelled in two examples. The product-limit method was used in SAS to compute the survivor function estimates. The parameter estimates using this method exactly matched those obtained from SPSS. There is only one option available in SPSS.

Example (iii): The binary variable (Y=1, G=0) compared HB Y with HB G with only these two Health Boards included in the fit. The corresponding results obtained from SAS and SPSS were as follows (Table A5.1):

	SAS	SPSS
Log hazard ratio for Y vs G	-1.311103	-1.3111
Standard error for log hazard ratio	1.00937	1.0094
5-yr survival estimate for Y	0.91848	0.9185
Standard error for 5-yr survival for Y	0.07807	0.0121
5-yr survival estimate for G	0.72941	0.7294
Standard error for 5-yr survival for G	0.03456	0.0356

Table A5.1: Results for the Cox models fitted by the two statistical packages for Example (iii).

The number of decimal places for each figure reflect those given by default in the output from the two packages exactly.

Thus, showing very different results for the standard errors for the 5-yr survival estimates for HB Y between SAS and SPSS. (The values for G were also different).

For both packages: the hazard ratio for Y vs G was 0.270 with 95% CI (0.037, 1.949)

NOT DIFFERENT

For SAS: the 5-yr survival estimate for Y was 0.9185 with 95% CI (0.7655, 1)
for G was 0.7294 with 95% CI (0.6617, 0.7921)

NOT DIFFERENT

For SPSS: the 5-yr survival estimate for Y was 0.9185 with 95% CI (0.8948, 0.9422)
for G was 0.7294 with 95% CI (0.6596, 0.7992)

DIFFERENT

Thus, the SAS set of figures produced consistent interpretations from the hazard ratio and the survival estimates, whereas the SPSS figures did not.

Example (iv): The binary variable (S=1, G=0) compared HB S with HB G with only cases for these two Health Boards included in the fit. The corresponding results obtained from SAS and SPSS were as follows (Table A5.2):

	SAS	SPSS
Log hazard ratio for S vs G	-0.206406	-0.2064
Standard error for log hazard ratio	0.20589	0.2059
5-yr survival estimate for S	0.78096	0.7810
Standard error for 5-yr survival for S	0.03142	0.0222
5-yr survival estimate for G	0.73792	0.7379
Standard error for 5-yr survival for G	0.03276	0.0258

Table A5.2: Results for the Cox models fitted by the two statistical packages for Example (iv).

Thus, showing the discrepancies between the two estimates for the standard errors from the two packages, but these were not as large as for HB Y vs HB G.

This finding was discussed with members of the Robertson Centre for Biostatistics (part of Glasgow University) and, independently, these inconsistencies were replicated on a different (much larger) dataset.

CONCLUSION

Due to these findings, it was decided not to use the standard errors until the apparent discrepancies had been resolved. Discussions with SPSS Inc. are still on-going and the issue remains unresolved.

APPENDIX 6 HAZARD RATIOS AND SURVIVAL ESTIMATES FOR VARIOUS SURVIVAL ANALYSES

Variable	Hazard Ratio (95% CI)
Age	
< 50 years	1
50 - 64	1.07 (0.85, 1.34)
65 - 79	1.29 (1.02, 1.63)
≥ 80 years	1.91 (1.28, 2.84)
Clinical Stage	
Stage I	1
II	1.33 (0.99, 1.77)
III	1.90 (1.33, 2.71)
Not known	1.34 (0.96, 1.87)
ER Status	
Positive	1
Negative	2.14 (1.72, 2.67)
Not known	1.43 (1.13, 1.81)
Node Status by Tumour Size	
N NK, T ≤ 2cm	2.55 (1.63, 3.99)
N NK, T > 2	4.10 (2.61, 6.44)
N NK, T NK	3.50 (2.19, 5.62)
N +ve, T ≤ 2cm	4.28 (2.79, 6.57)
N +ve, T > 2	4.45 (2.97, 6.66)
N +ve, T NK	4.92 (3.11, 7.78)
N -ve, T ≤ 2cm	1
N -ve, T > 2	2.82 (1.82, 4.37)
N -ve, T NK	1.57 (0.91, 2.72)
Health Board	
A	1.53 (1.11, 2.11)
C	1.50 (1.07, 2.12)
F	1.53 (1.04, 2.26)
G	1
L	1.21 (0.87, 1.68)
N	0.95 (0.69, 1.32)
S	0.89 (0.65, 1.21)
T	1.35 (0.96, 1.90)
Y	1.09 (0.69, 1.72)

Table A6.1: Hazard ratios with 95% CIs for the model based on 1432 cases, with Health Boards: B, H, I, V dropped from the ACM. Note that N and T stand for node status and tumour size respectively. Also, NK stands for not known, +ve for positive and -ve for negative.

Variable	All Cases 5-year % survival	Complete Cases 5-year % survival
Age		
< 50	73.3	69.9
50-64	73.8	75.6
65-79	68.1	70.6
≥ 80	48.6	50.0
Clinical Stage		
I	83.8	85.2
II	71.1	71.1
III	56.7	62.0
Node Status		
Positive	58.8	60.6
Negative	84.5	83.7
Tumour Size		
Size ≤ 2 cm	80.2	82.0
Size > 2 cm	62.2	64.5
ER Status		
Positive	80.0	82.7
Negative	60.9	56.1

Table A6.2: Kaplan-Meier % survival estimates at five years without CIs for the two analyses. Note that N and T stand for node status and tumour size respectively.

Variable	All Cases 5-year % survival	Complete Cases 5-year % survival
Health Board		
A	63.5	75.0
C	63.9	58.4
F	68.1	56.1
G	73.8	72.8
L	67.4	71.9
N	75.8	73.3
S	78.3	78.3
T	66.2	65.2
Y	68.9	90.0
Deprivation Category		
Least deprived	73.8	74.5
Intermediate	70.9	73.5
Most deprived	65.7	64.0
Surgical Case load		
1 - 9 cases	66.6	61.1
10 - 29 cases	69.1	73.2
Team or ≥ 30	74.7	73.6
Seen by an Oncologist		
Yes	70.4	70.6
No	71.5	74.9
Type of Surgery		
Mastectomy	68.0	68.2
Conservation	75.1	79.1
Adjuvant Radiotherapy		
Given	70.2	70.0
Not given	71.3	74.3
Adjuvant Chemotherapy		
Given	61.0	55.9
Not given	71.7	74.7
Adjuvant Endocrine Therapy		
Given	70.4	77.4
Not given	71.6	64.0
Adjuvant Chemotherapy or Endocrine Therapy		
Given	69.9	74.5
Not given	73.2	67.0

Table A6.2 cont: Kaplan-Meier % survival estimates at five years without CIs for the two analyses.

	(i) Node status and tumour size known		(ii) Three pathological factors known	
Variable	**Hazard Ratio**	**95% CI for Hazard Ratio**	**Hazard Ratio**	**95% CI for Hazard Ratio**
Age				
< 50	1	*	1	*
50-64	1.05	(0.80, 1.37)	0.93	(0.67, 1.28)
65-79	1.05	(0.78, 1.40)	1.00	(0.70, 1.43)
≥ 80	2.20	(1.23, 3.95)	2.10	(0.50, 8.78)
Clinical Stage				
I	1	*	1	*
II	1.64	(1.09, 2.48)	1.68	(1.01, 2.78)
III	2.08	(1.28, 3.38)	1.80	(0.98, 3.30)
ER Status				
Positive	1	*	1	*
Negative	2.51	(1.91, 3.31)	2.72	(2.05, 3.62)
Node Status by Tumour Size				
N+ T≤2	3.95	(2.63, 5.91)	4.62	(2.82, 7.57)
N+ T >2	4.52	(3.08, 6.61)	4.40	(2.74, 7.07)
N- T≤2	1	*	1	*
N- T >2	2.72	(1.81, 4.09)	2.82	(1.70, 4.65)
Health Board				
A	1.13	(0.70, 1.81)	1.24	(0.75, 2.05)
C	1.91	(1.24, 2.94)	2.33	(1.40, 3.86)
F	2.04	(1.27, 3.28)	2.50	(1.47, 4.25)
G	1	*	1	*
L	1.11	(0.71, 1.75)	0.99	(0.54, 1.81)
N	1.18	(0.80, 1.73)	1.21	(0.77, 1.88)
S	0.83	(0.57, 1.20)	0.84	(0.56, 1.24)
T	1.30	(0.78, 2.17)	1.97	(1.02, 3.82)
Y	0.34	(0.11, 1.08)	0.18	(0.02, 1.29)

Table A6.3: Hazard ratios (HR) with 95% CIs for the further two analyses. Note that N and T stand for node status and tumour size respectively.

APPENDIX 7 DERIVATION OF THE NORMAL EQUATIONS FOR TWO FACTORS IN AN EXPONENTIAL REGRESSION MODEL WITH COMPLETE CASES ONLY

DESIGN WITHOUT MISSING VALUES

Here, the two factors have only two levels with observations which are taken to satisfy an exponential regression model. It is assumed that there is no interaction between the factors in an additive model. Suppose there are n_{kl} observations falling in cell(k,l), where $k = 1,\ 2$ and $l = 1,\ 2$ represent the levels of factors F1 and F2 respectively. Let these n_{kl} observations be denoted by $y_{klm_{kl}}$ with $m_{kl} = 1,\ \ldots,\ n_{kl}$.

Then the basic design can be represented by Figure A7.1.

		F2	
		1	**2**
F1	**1**	n_{11}	n_{12}
	2	n_{21}	n_{22}

Figure A7.1: Diagram to represent the design without missing values in the two factors.

The pdf for the n_{kl} observations falling in cell(k,l) is given by

$$f\left(y_{klm_{kl}}\right) = \lambda_{kl} \exp\left(-\lambda_{kl} y_{klm_{kl}}\right),$$

for $k = 1,\ 2$ and $l = 1,\ 2$, where

$$\lambda_{kl} = \exp\left[\mu_{11} + \alpha_k + \beta_l\right]$$

and the constraints $\alpha_1 = 0$ and $\beta_1 = 0$ are imposed.

The α_k parameters are main effects related to factor F1 and the β_l parameters are main effects related to factor F2.

By the independence assumed between observations falling in each of the separate cells, the joint pdf for the n_{kl} observations falling in cell (k,l) is given by

$$\prod_{m_{kl}=1}^{n_{kl}} f\left(y_{klm_{kl}}\right).$$

By the independence between the cells, the overall joint pdf for the four cells is therefore given by

$$f_C\left(\underline{y};\underline{\lambda}\right)=\prod_{k,l}\left[\left(\lambda_{kl}\right)^{n_{kl}}\exp\left(-\lambda_{kl}\sum_{m_{kl}=1}^{n_{kl}} y_{klm_{kl}}\right)\right],$$

for $k=1,\ 2$, $l=1,\ 2$ and $m_{kl}=1,\ \ldots,\ n_{kl}$.

The subscript C on $f_C\left(\underline{y};\underline{\lambda}\right)$ is given here to demonstrate that only the complete cases (i.e. the known values) are included.

The likelihood function can be obtained from the joint pdf and taking logs gives the log-likelihood function as

$$l_C\left(\underline{\lambda}\right)=\sum_{k,l} n_{kl}\log\lambda_{kl}+\sum_{k,l}\left(-\lambda_{kl}\sum_{1}^{n_{kl}} y_{klm_{kl}}\right)+d,$$

where d does not depend on $\underline{\lambda}$.

The aim here is to obtain the parameter estimates $\hat{\mu}_{11}$, $\hat{\alpha}_2$ and $\hat{\beta}_2$. Therefore, the λ_{kl} are replaced by their values given above. Thus, the log-likelihood function becomes

$$\begin{aligned}l_C\left(\underline{\lambda}\right)&=n_{11}\left(\mu_{11}\right)+n_{12}\left(\mu_{11}+\beta_2\right)+n_{21}\left(\mu_{11}+\alpha_2\right)+n_{22}\left(\mu_{11}+\alpha_2+\beta_2\right)\\&\quad-\left[\exp\left(\mu_{11}\right)\right]\sum_{1}^{n_{11}} y_{11m_{11}}-\left[\exp\left(\mu_{11}+\beta_2\right)\right]\sum_{1}^{n_{12}} y_{12m_{12}}\\&\quad-\left[\exp\left(\mu_{11}+\alpha_2\right)\right]\sum_{1}^{n_{21}} y_{21m_{21}}-\left[\exp\left(\mu_{11}+\alpha_2+\beta_2\right)\right]\sum_{1}^{n_{22}} y_{22m_{22}}+d\end{aligned}$$

The three normal equations obtained are:

$$\frac{\partial l_C(\underline{\lambda})}{\partial \hat{\mu}_{11}} = n_{11} + n_{12} + n_{21} + n_{22} - \left[\exp\left(\hat{\mu}_{11}\right)\right]\sum_{1}^{n_{11}} y_{11m_{11}} - \left[\exp\left(\hat{\mu}_{11} + \hat{\beta}_2\right)\right]\sum_{1}^{n_{12}} y_{12m_{12}}$$

$$- \left[\exp\left(\hat{\mu}_{11} + \hat{\alpha}_2\right)\right]\sum_{1}^{n_{21}} y_{21m_{21}} - \left[\exp\left(\hat{\mu}_{11} + \hat{\alpha}_2 + \hat{\beta}_2\right)\right]\sum_{1}^{n_{22}} y_{22m_{22}}$$

$$= 0$$

$$\frac{\partial l_C(\underline{\lambda})}{\partial \hat{\alpha}_2} = n_{21} + n_{22} - \left[\exp\left(\hat{\mu}_{11} + \hat{\alpha}_2\right)\right]\sum_{1}^{n_{21}} y_{21m_{21}} - \left[\exp\left(\hat{\mu}_{11} + \hat{\alpha}_2 + \hat{\beta}_2\right)\right]\sum_{1}^{n_{22}} y_{22m_{22}}$$

$$= 0$$

and $$\frac{\partial l_C(\underline{\lambda})}{\partial \hat{\beta}_2} = n_{12} + n_{22} - \left[\exp\left(\hat{\mu}_{11} + \hat{\beta}_2\right)\right]\sum_{1}^{n_{12}} y_{12m_{12}} - \left[\exp\left(\hat{\mu}_{11} + \hat{\alpha}_2 + \hat{\beta}_2\right)\right]\sum_{1}^{n_{22}} y_{22m_{22}} .$$

$$= 0$$

APPENDIX 8 DERIVATION OF THE NORMAL EQUATIONS FOR TWO FACTORS IN AN EXPONENTIAL REGRESSION MODEL FOR ALL CASES

DESIGN WITH MISSING VALUES

When the missing values are also included in the exponential regression model, with the missing values being incorrectly taken to satisfy the model, the derivation of the normal equations proceeds in the same manner as for the context without the missing values (Appendix 7). The assumption of independence in an additive model is again made. Here there are now n_{kl} observations falling in cell (k,l), where $k = 1,\ 2,\ 3$ and $l = 1,\ 2,\ 3$. The new design can be represented diagrammatically by Figure A8.1.

		F2		
		1	**2**	**3**
	1	n_{11}	n_{12}	n_{13}
F1	**2**	n_{21}	n_{22}	n_{23}
	3	n_{31}	n_{32}	n_{33}

Figure A8.1: Diagram to represent the design for the two factors when missing values are included.

Since the observations falling into cells $(1,3)$, $(2,3)$, $(3,1)$, $(3,2)$ and $(3,3)$ are taken here to be exponential then the pdfs for all nine cells is given by

$$f\left(y_{klm_{kl}}\right) = \lambda_{kl} \exp\left(-\lambda_{kl} y_{klm_{kl}}\right),$$

for $k = 1,\ 2,\ 3$ and $l = 1,\ 2,\ 3$, where

$$\lambda_{kl} = \exp\left[\mu_{11} + \alpha_k + \beta_l\right]$$

and λ_{kl} represent the hazard functions for the nine cells with the constraints $\alpha_1 = 0$ and $\beta_1 = 0$ still imposed.

Again, by independence assumed between observations falling in the same cell, the joint pdf for the n_{kl} observations in cell (k,l) is still given by

$$\prod_{m_{kl}=1}^{n_{kl}} f\left(y_{klm_{kl}}\right).$$

Therefore, by independence between observations in different cells, the overall joint pdf for the nine cells is given by

$$f_A\left(\underline{y};\underline{\lambda}\right) = \prod_{k,l}\left[\left(\lambda_{kl}\right)^{n_{kl}} \exp\left(-\lambda_{kl}\sum_{m_{kl}=1}^{n_{kl}} y_{klm_{kl}}\right)\right],$$

with $k = 1,\ 2,\ 3$, $l = 1,\ 2,\ 3$ and $m_{kl} = 1,\ \ldots,\ n_{kl}$.

The subscript A here is given to show that all of the cases (i.e. known and missing values) have now been included. This is in contrast to the C used before for the complete cases in Appendix 7.

Therefore, the log-likelihood function for all cases is given by

$$l_A\left(\underline{\lambda}\right) = \sum_{k,l} n_{kl} \log\lambda_{kl} + \sum_{k,l}\left(-\lambda_{kl}\sum_{1}^{n_{kl}} y_{klm_{kl}}\right) + e,$$

where e does not depend on $\underline{\lambda}$.

To obtain the normal equations, it is necessary to replace the λ_{kl} by the linear combinations of the parameter estimates of interest. Differentiation of $l_A\left(\underline{\lambda}\right)$ with respect to the parameter estimates $\hat{\mu}_{11}$, $\hat{\alpha}_2$, $\hat{\beta}_2$, $\hat{\alpha}_3$ and $\hat{\beta}_3$ yields five normal equations which need to be solved simultaneously once they are all set equal to zero. The five normal equations for the design with missing values in the two factors case are:

$$\frac{\partial\, l_A(\underline{\lambda})}{\partial\hat{\mu}_{11}} = n_{11} + n_{12} + n_{13} + n_{21} + n_{22} + n_{23} + n_{31} + n_{32} + n_{33} - \left[\exp\left(\hat{\mu}_{11}\right)\right]\sum_{1}^{n_{11}} y_{11m_{11}}$$

$$- \left[\exp\left(\hat{\mu}_{11} + \hat{\beta}_2\right)\right]\sum_{1}^{n_{12}} y_{12m_{12}} - \left[\exp\left(\hat{\mu}_{11} + \hat{\beta}_3\right)\right]\sum_{1}^{n_{13}} y_{13m_{13}}$$

$$- \left[\exp\left(\hat{\mu}_{11} + \hat{\alpha}_2\right)\right]\sum_{1}^{n_{21}} y_{21m_{21}} - \left[\exp\left(\hat{\mu}_{11} + \hat{\alpha}_2 + \hat{\beta}_2\right)\right]\sum_{1}^{n_{22}} y_{22m_{22}}$$

$$- \left[\exp\left(\hat{\mu}_{11} + \hat{\alpha}_2 + \hat{\beta}_3\right)\right]\sum_{1}^{n_{23}} y_{23m_{23}} - \left[\exp\left(\hat{\mu}_{11} + \hat{\alpha}_3\right)\right]\sum_{1}^{n_{31}} y_{31m_{31}}$$

$$- \left[\exp\left(\hat{\mu}_{11} + \hat{\alpha}_3 + \hat{\beta}_2\right)\right]\sum_{1}^{n_{32}} y_{32m_{32}} - \left[\exp\left(\hat{\mu}_{11} + \hat{\alpha}_3 + \hat{\beta}_3\right)\right]\sum_{1}^{n_{33}} y_{33m_{33}}$$

$$= 0$$

$$\frac{\partial\, l_A(\underline{\lambda})}{\partial\hat{\alpha}_2} = n_{21} + n_{22} + n_{23} - \left[\exp\left(\hat{\mu}_{11} + \hat{\alpha}_2\right)\right]\sum_{1}^{n_{21}} y_{21m_{21}} - \left[\exp\left(\hat{\mu}_{11} + \hat{\alpha}_2 + \hat{\beta}_2\right)\right]\sum_{1}^{n_{22}} y_{22m_{22}}$$

$$- \left[\exp\left(\hat{\mu}_{11} + \hat{\alpha}_2 + \hat{\beta}_3\right)\right]\sum_{1}^{n_{23}} y_{23m_{23}}$$

$$= 0$$

$$\frac{\partial\, l_A(\underline{\lambda})}{\partial\hat{\beta}_2} = n_{12} + n_{22} + n_{32} - \left[\exp\left(\hat{\mu}_{11} + \hat{\beta}_2\right)\right]\sum_{1}^{n_{12}} y_{12m_{12}} - \left[\exp\left(\hat{\mu}_{11} + \hat{\alpha}_2 + \hat{\beta}_2\right)\right]\sum_{1}^{n_{22}} y_{22m_{22}}$$

$$- \left[\exp\left(\hat{\mu}_{11} + \hat{\alpha}_3 + \hat{\beta}_2\right)\right]\sum_{1}^{n_{32}} y_{32m_{32}}$$

$$= 0$$

$$\frac{\partial\, l_A(\underline{\lambda})}{\partial\hat{\alpha}_3} = n_{31} + n_{32} + n_{33} - \left[\exp\left(\hat{\mu}_{11} + \hat{\alpha}_3\right)\right]\sum_{1}^{n_{31}} y_{31m_{31}} - \left[\exp\left(\hat{\mu}_{11} + \hat{\alpha}_3 + \hat{\beta}_2\right)\right]\sum_{1}^{n_{32}} y_{32m_{32}}$$

$$- \left[\exp\left(\hat{\mu}_{11} + \hat{\alpha}_3 + \hat{\beta}_3\right)\right]\sum_{1}^{n_{33}} y_{33m_{33}}$$

$$= 0$$

and

$$\frac{\partial\, l_A(\underline{\lambda})}{\partial\hat{\beta}_3} = n_{13} + n_{23} + n_{33} - \left[\exp\left(\hat{\mu}_{11} + \hat{\beta}_3\right)\right]\sum_{1}^{n_{13}} y_{13m_{13}} - \left[\exp\left(\hat{\mu}_{11} + \hat{\alpha}_2 + \hat{\beta}_3\right)\right]\sum_{1}^{n_{23}} y_{23m_{23}}$$

$$- \left[\exp\left(\hat{\mu}_{11} + \hat{\alpha}_3 + \hat{\beta}_3\right)\right]\sum_{1}^{n_{33}} y_{33m_{33}}$$

$$= 0$$

APPENDIX 9 DETERMINING THE MIXING PARAMETERS FOR GENERATING THE MISSING VALUES FOR THE TWO FACTORS DESIGN

A sequence of randomly generated uniform numbers was utilised to decide which exponential distribution to use to generate the observations falling in the missing cells. The appropriate probabilities were chosen based on the relative frequencies of the cases in the known levels. The missing value was then generated from the appropriate exponential distribution.

The following probabilities for the five missing cells are:

For cell $(1,3)$, the mixing parameter was $r = \dfrac{n_{11}}{n_{11} + n_{12}}$, with an observation coming from an $Ex\left(\exp\left(\mu_{11}\right)\right)$ distribution when the attached random uniform number had a value $\leq r$; otherwise the observation was generated from an $Ex\left(\exp\left(\mu_{11} + \beta_2\right)\right)$ distribution. Similarly, for cell $(2,3)$, the mixing parameter was $s = \dfrac{n_{21}}{n_{21} + n_{22}}$, with a similar use made of the attached uniform numbers. For cells $(3,1)$ and $(3,2)$, the cut-offs were $t = \dfrac{n_{11}}{n_{11} + n_{21}}$ and $u = \dfrac{n_{12}}{n_{12} + n_{22}}$ respectively, again with similar use of the attached uniform random numbers.

For cell $(3,3)$, however, three probabilities were needed. These were at $a = \dfrac{n_{11}}{n_{11} + n_{12} + n_{21} + n_{22}}$, $b = \dfrac{n_{11} + n_{12}}{n_{11} + n_{12} + n_{21} + n_{22}}$ and $c = \dfrac{n_{11} + n_{12} + n_{21}}{n_{11} + n_{12} + n_{21} + n_{22}}$. The observations in this cell were generated from the four possible distributions depending on the values of the uniform random numbers attached to each observation. For a uniform number of value $\leq a$, the observation in cell $(3,3)$ was generated from an $Ex\left(\exp\left(\mu_{11}\right)\right)$ distribution. When the number lay in the interval $(a,b]$ or $(b,c]$, the distribution used was an $Ex\left(\exp\left(\mu_{11} + \beta_2\right)\right)$ or $Ex\left(\exp\left(\mu_{11} + \alpha_2\right)\right)$ respectively.

Otherwise, the observation was generated from an $Ex(\exp(\mu_{11} + \alpha_2 + \beta_2))$ distribution.

APPENDIX 10 FURTHER RESULTS FROM FITTING AN EXPONENTIAL REGRESSION MODEL TO DIFFERENT DESIGNS WITH TWO FACTORS AND MISSING VALUES PRESENT

From Section 6.3.5:

GROUP H

Table A10.1 gives the estimated biases and standard errors for the parameter estimates for Group H.

Design	$b(\hat{\alpha}_2)$	$se(\hat{\alpha}_2)$	$b(\hat{\beta}_2)$	$se(\hat{\beta}_2)$
H1	-0.0519	0.0041	-0.0631	0.0040
H2	-0.0160	0.0049	-0.0088	0.0048
H3	-0.0373	0.0044	-0.0431	0.0042
H4	-0.0158	0.0049	-0.0087	0.0048
H5	-0.0145	0.0049	-0.0082	0.0048
H6	-0.0114	0.0050	-0.0080	0.0047
H7	-0.0317	0.0046	-0.0136	0.0047
H8	-0.0148	0.0049	-0.0083	0.0048
H9	-0.0179	0.0049	-0.0199	0.0046
H10	-0.0145	0.0050	-0.0127	0.0047
H11	-0.0090	0.0050	-0.0134	0.0047
H12	-0.0098	0.0051	-0.0091	0.0047
H13	-0.0155	0.0050	-0.0082	0.0048
H14	-0.0197	0.0049	-0.0117	0.0047
H15	-0.0116	0.0050	-0.0119	0.0047
H16	-0.0162	0.0049	-0.0101	0.0047
H17	-0.0162	0.0050	-0.0133	0.0047
H18	-0.0137	0.0049	-0.0073	0.0048
H19	-0.0082	0.0050	-0.0081	0.0047
H20	-0.0102	0.0050	-0.0062	0.0048
H21	-0.0048	0.0051	-0.0058	0.0048
H22	-0.0020	0.0051	-0.0016	0.0048

Table A10.1: Estimated biases and standard errors for the designs in Group H.

No clear pattern could be determined from a plot of magnitude of bias and percentage of missing values (Figure A10.1).

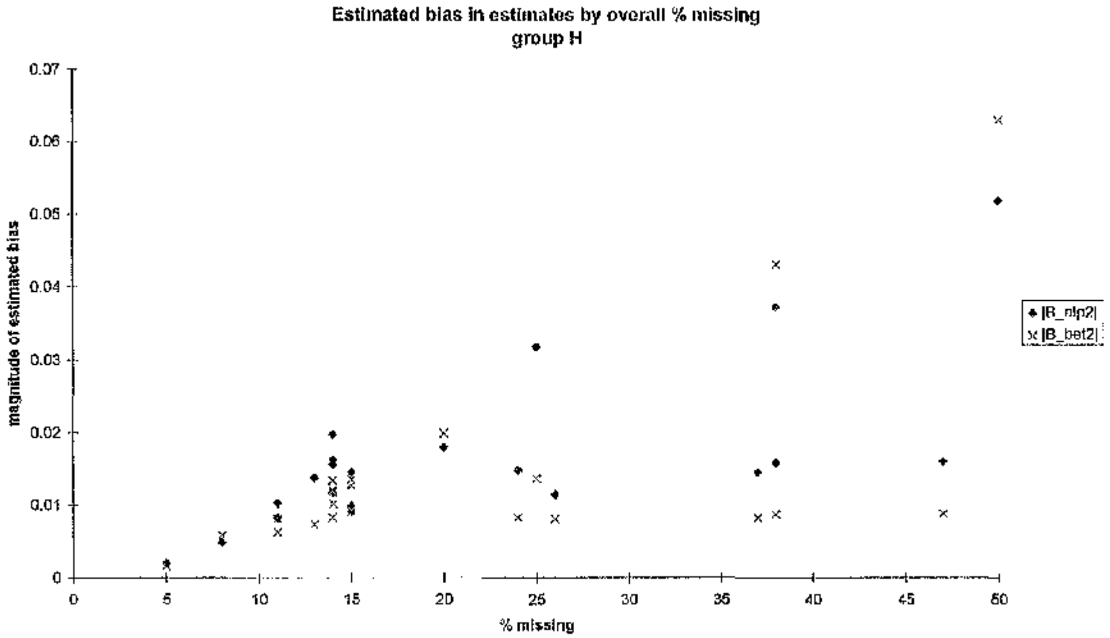


Figure A10.1: Plot of the magnitude of the estimated bias against the overall percentage missing for Group H.

GROUP I

Table A10.2 gives the estimated biases and standard errors for the parameter estimates for Group I.

Design	$b(\hat{\alpha}_2)$	$se(\hat{\alpha}_2)$	$b(\hat{\beta}_2)$	$se(\hat{\beta}_2)$
I1	-0.0202	0.0051	-0.0261	0.0052
I2	-0.0116	0.0053	-0.0129	0.0054
I3	-0.0116	0.0053	-0.0128	0.0054
I4	-0.0106	0.0053	-0.0095	0.0054
I5	-0.0106	0.0053	-0.0095	0.0054
I6	-0.0076	0.0053	-0.0115	0.0054
I7	-0.0103	0.0053	-0.0125	0.0054
I8	-0.0060	0.0054	-0.0060	0.0055

Table A10.2: Estimated biases and standard errors for the designs in Group I.

The amount of missing values seemed to affect whether or not the parameter estimates were biased, with 14% or more missing leading to biased estimates (Figure A10.2). Models I3 and I4 both had 13% missing with model I3 having biased parameter estimates, whereas there was no evidence of bias of $\hat{\beta}_2$ for model I4, although $\hat{\alpha}_2$ was biased.

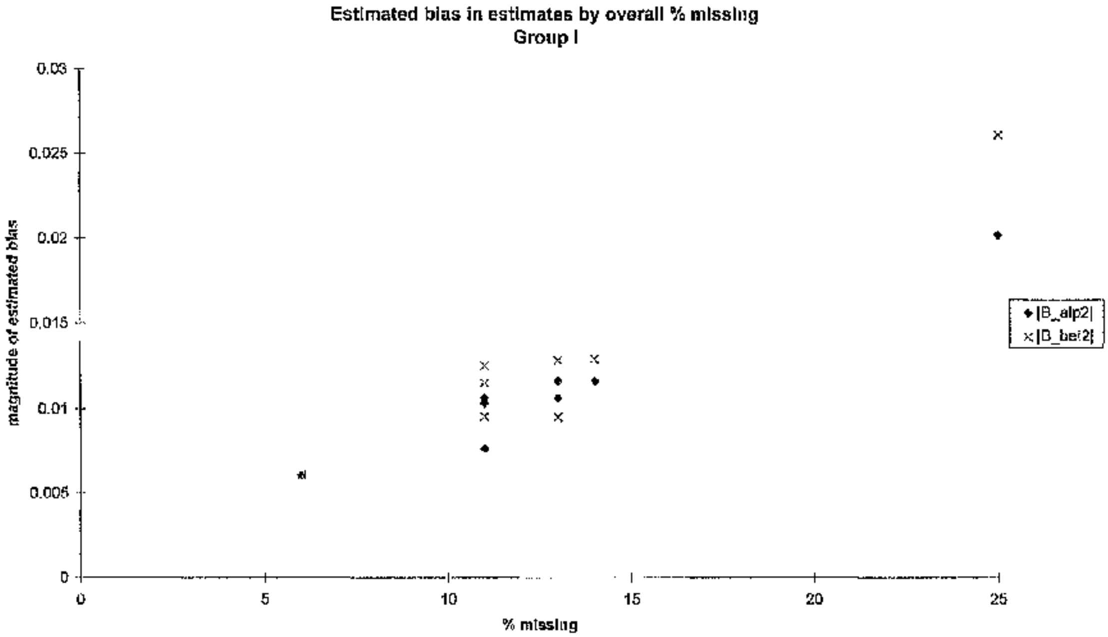


Figure A10.2: Plot of the magnitude of the estimated bias against the overall percentage missing for Group I.