THE STATISTICAL ANALYSIS OF EXERCISE TEST DATA:

A CRITICAL REVIEW

Alex McConnachie

Submitted to the

Faculty of Information and Mathematical Sciences,
University of Glasgow

for the degree of Ph.D.

July 2003

ABSTRACT

Exercise testing is used for a variety of purposes, in particular for the evaluation of patients with or at high risk of having coronary artery disease. The occurrence of chest pain or electrocardiographic (ECG) changes indicative of restricted coronary blood supply (ischaemia) during exercise are diagnostic for the presence of disease and prognostic of poor coronary outcomes. In a clinical setting the analysis of exercise test data is the responsibility of an experienced physician, based on test outcomes and knowledge of the particular patient.

In clinical trials of anti-anginal therapies, the use of exercise tests in the evaluation of treatment efficacy is required by agencies responsible for authorising the use of new medicinal products. Considerable attention has been paid to the development of standardised exercise protocols that elicit reproducible ischaemic responses, and to alternative methods of analysing exercise test outcomes in order to improve the diagnostic or prognostic value of the test. However, comparatively little research has addressed the problem of analysing exercise test data produced for the assessment of treatment efficacy in clinical trials.

Exercise tests have played a prominent role in the evaluation of therapies currently used for the management of patients with angina, such as nitrates, β-blockers, and calcium antagonists. Such evaluations have shown dramatic improvements in exercise tolerance, most commonly measured by the time spent exercising until the occurrence of anginal pain or ECG signs of ischaemia, and often amongst patients with severe disease. However, the statistical methods used have generally been based on Normal theory, such as the t-test, or non-parametric equivalents, such as the Wilcoxon rank sum test. Such methods make no allowance for the fact that ischaemic endpoints may not occur in all patients, particularly when patients are under active treatment or in patients with less severe symptoms. In the current situation, where there are several therapeutic options of proven clinical effectiveness, new treatments must be evaluated in opposition, or in addition to existing therapies. Thus it is of particular importance that the statistician responsible for an analysis of exercise test data should use appropriate and efficient techniques, since the benefits of new treatments may be small.

Since exercise times may be censored, in that the event of interest need not occur, it has been recognised that methods for the analysis of survival data are the most appropriate for analyses of exercise test data. Using data from the TIBET Study, a large

clinical trial of two anti-anginal therapies administered singly or in combination, this thesis examines in detail the appropriateness of the Cox proportional hazards model, the most popular method for survival regression in the medical literature, to this type of data. It then considers alternatives to this model, and addresses the implications of some common features of exercise test data, in particular the presence of interval censoring and the possibility of multiple exercise tests being conducted on the same patient, using data from the TIBET Study and through simulation studies. Finally, using real data examples, two methods that appear to have received little or no attention with respect to exercise test data are explored, namely competing risks and repeated measures analyses.

It is shown that the Cox model, and potentially other parametric survival regression models perform well with these data, even in the presence of moderate interval censoring. When multiple exercise times are analysed from the same group of patients, however, there is the potential for considerable bias and loss of power if this is not taken into account in the analysis. There is also much potential in the use of more complex statistical models, considering the volume of data routinely collected as part of each exercise test, and these may prove to be useful avenues for further research.

The research contained in this Thesis was carried out by the author during the period 1993-2003, whilst a research student (1993-1997) at:

Robertson Centre for Biostatistics
University of Glasgow Department of Statistics
Boyd Orr Building
University Avenue
Glasgow
G12 8QQ

and whilst employed as Statistician (1996-2003) at:

General Practice and Primary Care
Division of Community Based Sciences
University of Glasgow
4 Lancaster Crescent
Glasgow
G12 0RR

ACKNOWLEDGEMENTS

Ian, for persistence. Sharon, for insistence.
Dayll and Erin, for tolerance. Sophie and Alice, for existence.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 Introduction

1.1 Exercise Testing

The diagnosis and treatment of patients with cardiac conditions begins with their medical history and a physical examination. An important step in determining the aetiology of chest pain, the extent of disease, the response to therapy and the capacity to do work or other activities is to then perform an exercise test. At relatively little cost in terms of both health care resources and patient discomfort, clinically useful information is obtained that can drive the decision-making process, leading to the appropriate implementation of more expensive, invasive and/or dangerous procedures[1].

1.1.1 Preparation and Safety Precautions

The patient's history and physical exam, including resting electrocardiogram and lung function tests, will identify many of the numerous contraindications to performing an exercise test[2], including serious cardiac conditions (such as acute myocardial infarction, serious dysrhythmias or unstable angina), acute or serious non-cardiac disorders and severe physical disability. Other conditions that might contraindicate an exercise test include less serious disorders (cardiac or non-cardiac), drug effects and psychiatric disease.

Patients should not have smoked tobacco, drunk caffeine or had a meal for at least two hours prior to a test. They should wear clothes and shoes appropriate for exercise and be allowed time to familiarise themselves with the apparatus. They must give written informed consent to participate and be aware that they can terminate the test at any time they choose.

Under such conditions, the safety of exercise testing has been established[3,4] though facilities for emergency cardiopulmonary resuscitation should be available, and trained staff should be close at hand in the event of an emergency.

1.1.2 Types of Exercise

Exercise can be classified into two types, isometric and dynamic, though most activities consist of a mixture of the two. Isometric exercise involves constant muscular contraction, whereas dynamic exercise is that which results in movement, usually by the rhythmical contraction of large muscle groups. For the assessment of cardiac patients, a dynamic exercise test is preferred since it results in a greater increase in the total oxygen requirement of the whole body.

As the total body oxygen requirement is increased, cardiac output must increase to ensure that sufficient blood is delivered to the working muscles, where oxygen is used in the energy releasing process, and to the lungs, where oxygen is extracted from the inhaled air, in exchange for the carbon dioxide produced in the muscles. As the result of an exercise load, the metabolic rate can increase by up to 20 times, and cardiac output by up to 6 times relative to the resting rate. The maximal extent of these increases will depend on the age, sex and fitness of the patient, the type of exercise performed, and the presence and extent of heart disease.

During an exercise test, the total workload can be controlled, as can the cardiac output needed to meet the increased oxygen requirements. If necessary, total workload (and hence cardiac output) can be increased to the point where the subject is unable to continue. By studying the patient's haemodynamic and electrocardiographic responses to exercise, particularly near their maximal exercise capacity, the physician may be able to determine the cause or extent of heart disease, evaluate suitability for a surgical procedure or other course of action, or assess the progress of current therapies.

In modern clinical practice, exercise tests are often performed using either a treadmill or bicycle. Bicycle ergometry is cheaper and requires less space, but many patients will suffer muscle fatigue before reaching their cardiorespiratory exercise limit. Treadmills are larger and more costly, but offer greater flexibility in terms of controlling workload. They may also elicit greater effort from patients who are more used to walking than cycling, thereby producing a greater maximal work rate.

1.1.3 Exercise Protocol

Any exercise test should be carried out according to a pre-specified protocol. An early example is the Master step test[5] in which subjects made a certain number of trips on standard sized steps, the number being determined by their age, sex and weight. This was a single stage test in the sense that an individual exercised against a fixed load

throughout the test. More recently, with the use of treadmill and bicycle ergometers, it has become possible to alter the workload during a test without interruption, whilst keeping the patient in a relatively fixed position. This enables measurements to be made throughout the test under increasing levels of stress, giving a more detailed picture of the response to exercise.

A large number of exercise protocols are currently in use, some of which are based on protocols devised more than 30 years ago[6,7,8]. For some time, the most commonly used protocol was that devised by Bruce[6], though the relatively large and unequal increments in workload were found to give biased estimates of exercise capacity[9], and protocols with small or even continuous increments have subsequently been recommended[10,11,12,13].

1.1.4 Responses to Exercise

Cardiac output is defined as the product of stroke volume (the quantity of blood pumped at each beat) and heart rate. When exercise begins, stroke volume increases almost immediately[14], with the magnitude of this increase dependent upon fitness, age and body size[15]. As exercise continues, changes in stroke volume are small, and the necessary increases in cardiac output are predominantly met by increases in heart rate[16].

As workload increases, cardiac output must increase in order for the muscles to perform to the greater work rate. Consequently, myocardial oxygen consumption will increase, as the heart must also perform to a higher work rate. Myocardial oxygen consumption is dependent upon intramyocardial wall tension (the product of left ventricular pressure and end diastolic volume), contractility and heart rate[1]. An accurate measurement of myocardial oxygen consumption requires catheters to be placed in the coronary arteries and coronary venous sinus to measure oxygen content. In most clinical settings this is impractical, but the "double product" of heart rate and systolic blood pressure (SBP) is a good surrogate measure[17]. During progressively increasing exercise, SBP (as well as heart rate) will normally increase[1], as will diastolic blood pressure (DBP), though the increase is less marked.

1.1.5 Measurements

For cardiac patients, the main objective of exercise testing is to assess the response to increased demand of myocardial oxygen supply, which is measured by the double product of heart rate and SBP. Automatic devices for recording blood pressures

during exercise initially suffered from technical difficulties caused by the patient's motion, but have since improved performance[18,19]. SBP, as measured by cuff and auscultation, is a reliable method where an automated device is not available. DBP provides little additional information to the test, and is generally not recorded. If data on SBP are not available, the heart rate alone can be used as a proxy for myocardial oxygen consumption, since it correlates almost as well as the double product[20].

The principal device used to monitor the patient during exercise testing is the electrocardiogram (ECG). This measures the variations in electrical potential across the heart, with a recognisable pattern occurring during each heartbeat. The length of each cycle of this pattern, or an average over several beats, is used to calculate the heart rate of the patient at any stage before, during or after the test.

The level of exercise performed can be measured by several methods. The total workload, or the time spent exercising under a standard protocol, measures the amount of work that a subject is doing, whereas the double product or heart rate estimate the amount of work that the heart is doing. What is often of interest is the level of exercise that a subject can achieve before the blood supply to the heart fails to match its oxygen requirements (ischaemia). When this occurs, a patient will often experience pain to the chest (angina) and/or abnormal ECG changes (see Section 1.3.1). Though the occurrence of anginal pain will be of most interest to the patient, it is a subjective endpoint and may be influenced by motivation. ECG changes offer a more objective measure of ischaemia. However, anginal pain and abnormal ECG changes give alternative but dependent indications of coronary artery disease and increased risk[21,22].

With the necessary equipment, it is possible to measure the gases expired by a subject during exercise. This enables calculation of total body oxygen uptake and thus accurate measurement of workload. It is also useful in assessment of respiratory and physical response to exercise, such as anaerobic threshold. Other physical responses to exercise can be measured by taking small blood samples during exercise. However, in cardiac patients, it is the response to exercise of the heart, rather than the whole body, which is of greatest interest, and these measurements do not add much to a test.

Other methods of detecting myocardial ischaemia are based on the premise that when an area of heart muscle becomes ischaemic, the muscle ceases to move. Such wall motion abnormalities can be detected by nuclear imaging techniques, in which a radioactive substance is injected into the subject, making the heart visible to detectors once the heart becomes filled with blood containing the radionuclides. An alternative

technique is to use an echocardiogram, whereby ultrasound is used to view the heart[1]. These methods allow measurement of wall motion as well as estimation of the volumes of the chambers of the heart. The images can be digitised and recorded for later analysis. Methodological advances have enabled measurements to be made during exercise, when ischaemia is most likely to occur, rather than after the subject has stopped moving, thus increasing the sensitivity of these tests for detecting ischaemic events. However, the equipment required does increase the costs involved in conducting a test, and may be prohibitive for routine tests on a single patient as well as in the clinical trial setting, where large numbers of patients may be tested on several occasions at a number of different sites.

1.2 Angina

Angina pectoris (AP) is a collection of symptoms characterised by discomfort in the chest area associated with a disturbance to the function of the heart[23]. It usually involves pain to the chest, but the pain can radiate to the arms (more often the left), the neck or the back. Sufferers often describe the pain as a sense of tightness, of pressure or an aching or burning feeling across the chest; the name angina comes from the Greek work meaning to choke. It is common, with symptoms becoming more severe with age. In the 1998 Scottish Health Survey[24], the prevalence of angina in the Scottish population was estimated as 3.5% in men and 3.1% in women, based on the Rose Angina Questionnaire[25] definition of angina, Grade I or Grade II. In the oldest age group surveyed (65-74 years) these prevalences were estimated as 8.9% and 5.9% respectively.

Angina can be categorised as stable or unstable. As the term implies, stable angina will have been present for a number of weeks with no recent increase in frequency or severity of attacks. It is brought on by effort or other predictable stressors, and is relieved by rest or sublingual nitrates (see Section 1.2.2). Unstable angina is of recent onset, of increasing frequency or severity, or occurs at rest for no obvious reason. Exercise tests should not be performed on a subject suffering from unstable angina (Section 1.1.1).

1.2.1 Causes

Angina is ischaemic cardiac pain. It occurs when the coronary blood supply cannot meet an increased myocardial demand for oxygen, such as during exercise. It is usually a reflection of coronary artery disease (CAD), the second leading cause of death

after cancer in both men and women[26] accounting for 22.9% of all male deaths and 18.8% of female deaths in Scotland in 2001. CAD is mostly caused by occlusion of the coronary arteries, which reduces the rate at which blood can be supplied to the heart muscle. Any factor which causes an increase in heart rate and thus in the need of the heart muscle for oxygen is likely to precipitate angina. Common causes are exercise, strong emotion (anger or fear), a recent meal, cold temperature, vivid dreams or sex. Combinations of factors increase the risks.

Arterial narrowing is not the only cause of angina. The extent of symptoms does not predict the extent of CAD well; up to 30% of those who have coronary arteriography for the evaluation of chest pain are found to have normal coronary arteries[27] and many of those with CAD are asymptomatic[28]. Several possible mechanisms have been postulated for alternative causes of anginal symptoms, such as coronary vasospasm[29], platelet aggregation[30] and small vessel disease[31].

1.2.2 Management and Treatment

The first phase of management of angina patients involves lifestyle changes, including stopping smoking, improving diet, stress management and possibly some exercise, depending on the patient. Further management of angina is obtained through drug therapy, or in severe cases, surgical intervention. The treatment of stable angina has two aims, namely to prevent coronary endpoints such as myocardial infarction (MI) or death, and secondly to reduce the symptoms of angina.

1.2.3 Drug Therapies to Prevent MI and Death

The primary concern in the management of patients with stable angina is to extend life. Aspirin is antithrombotic and should be used routinely[32], and Angiotensin Converting Enzyme (ACE) Inhibitors are also useful for patients with or at high risk of developing CAD[33]. Statins and antihypertensive agents that are used to control risk factor levels can also prevent severe coronary endpoints and should be prescribed to patients at highest risk[34].

1.2.4 Drug Therapies to Prevent Anginal Symptoms

The four main classes of medical treatment for the relief of anginal symptoms are nitrates, beta-blockers, calcium antagonists and potassium channel openers[32]. Though primarily used to prevent ischaemic symptoms, in some cases they also prevent coronary endpoints such as MI and death.

Nitrates cause vasodilatation, resulting in improved myocardial perfusion and reduced oxygen requirements of the heart. Sublingual nitroglycerin tablets or sprays act within 1-2 minutes and last for half an hour, and are used in the short term to give relief from symptoms or prophylactically, before an activity which is likely to be precipitative. Slow release nitrates, such as isosorbide dinitrate, transdermal nitroglycerin patches and ointments can also be used to prevent the recurrence of angina, but tolerance can become a problem unless sufficient nitrate-free periods are provided. Side effects include headache, syncope, tachycardia and halitosis.

Beta-blockers reduce heart rate, blood pressure and contractility, thereby reducing myocardial oxygen demand. They are effective in most patients, leading to fewer ischaemic events and less need to use sublingual nitrates. They also improve prognosis in terms of survival and the occurrence of stroke and chronic heart failure in patients with a recent infarction[35]. Adversely, they can cause fatigue, insomnia, nightmares and impotence.

Calcium antagonists reduce coronary vascular resistance and arterial pressure, reducing cardiac demand and increasing coronary blood flow. They are therefore of benefit in the treatment of both supply and demand ischaemia. Side effects are due mainly to vasodilatation and include headaches and water retention. They can potentially increase the risk of adverse cardiac events[36] and are used in combination with beta-blockers if the initial treatment is ineffective, or as an alternative when the side effects of initial treatment are unacceptable.

Nicorandil is a potassium channel opener with nitrate-like properties, which dilates coronary blood vessels, increasing blood flow and reducing cardiac preload and afterload[37]. Nicorandil is the only antianginal agent to have been shown to reduce coronary endpoints in patients with stable angina[38]. Evidence regarding side effects is scarce, though there have been some reports of oral[39] and anal[40] ulceration as well as other gastrointestinal events and headache[38].

1.2.5 Surgical Intervention

For patients who are at high risk of death or whose symptoms do not improve after some weeks of medical treatment, surgical intervention will be considered[41]. Depending on the particular circumstances, treatment will be either percutaneous coronary intervention (PCI) or coronary artery bypass grafting (CABG). PCI originally involved the inflation of a balloon within an occluded artery to reduce the level of

stenosis, but more recent techniques have involved the use of rotating blades or lasers to remove lesions, and the insertion of intracoronary stents to prevent restenosis. CABG involves bypassing the damaged section of a coronary artery with a section of another blood vessel.

Both methods increase the blood flow to the myocardium, reducing anginal symptoms and the need for drug therapy, and improving prognosis in terms of subsequent MI or death. However, the chances are high that a severe coronary event will occur or further treatment will be required, even if lifestyle changes are made. In general, CABGs involve longer hospitalisations and recovery periods, but patients show greater improvements in symptoms, whilst PCIs are relatively inexpensive, low risk and repeatable. Mortality and re-infarction rates are similar, except for diabetic patients, for whom CABG offers a better prognosis[42].

1.3 Exercise Testing and Angina

Exercise testing of angina patients is used for a number of reasons. It is used as a diagnostic tool, to detect both the presence and severity of disease; for evaluation of prognosis; to monitor the progress of patients after drug or surgical therapy; and in clinical trials for the comparison of different drug treatments.

1.3.1 Ischaemia

When subjected to increasing levels of exercise, an individual's heart rate and systolic blood pressure will increase, as will the need of the heart muscle for oxygen, and therefore blood. The main factor in controlling supply is the resistance of the vessels that deliver the blood. These are the coronary arteries, which are basically fixed in width and cannot vary their resistance, and arterioles, which can dilate in the presence of increased need and supply greater amounts of blood. The ability of these vessels to dilate is limited; for people with normal coronary arteries this limit is not reached and fatigue or breathlessness is the limiting factor in exercise. For people with obstructed coronary arteries, which have reduced capacity to carry the blood, the arterioles are slightly dilated at rest and under exercise they reach the limit of their ability to increase diameter. The coronary blood supply is then unable to meet the raised requirements of the heart, resulting in myocardial ischaemia.

When ischaemia occurs, it may result in chest pain, though the inability of a subject to exercise may prevent the individual reaching the necessary level of cardiac

work to produce this response. This exercise capacity will depend on both physical fitness and on motivation, which reduces the usefulness of exercise induced chest pain as an outcome. An alternative measure of ischaemia can be defined through ECG changes, usually by inspection of the ST-segment of the ECG curve.

During ischaemia, the ST-segment can become elevated, normalise or become depressed. ST elevation at rest is normal, but during exercise could have one of several causes including variant angina[43], and is more common in those with a prior myocardial infarction. Normalisation of the ST-segment can be thought of as ST elevation in a subject whose ST-segment was depressed at baseline. The most common manifestation of exercise induced myocardial ischaemia is ST-segment depression.

1.3.2 Diagnosis of CAD

Both the occurrence of chest pain and ST-segment depression during exercise can be used as diagnostic indicators for CAD, though the occurrence of either is generally taken as a positive test. A depressed but up-sloping ST-segment is normal; horizontal or down-sloping ST-segment depression indicates ischaemia. Greater levels of ST-segment depression, as well as a more down-sloping ST-segment are related to a greater likelihood of, and greater severity of disease. The level of ST-segment depression used as a sign of disease will influence the sensitivity and specificity of the diagnostic test. A threshold often used to indicate significant ischaemia is ≥1mm of ST-segment depression.

1.3.3 Evaluation of Prognosis

In those with severe CAD, the use of exercise testing to predict outcome is valuable to identify patients for whom surgery will improve prognosis. It is those patients with the poorest prognosis that are most likely to benefit from surgery. Indicators of high risk or improved prognosis with surgery include short exercise time, significant ST-segment depression at low exercise levels that persists late into the recovery period or large amounts of ST-segment depression[44].

1.3.4 Post Myocardial Infarction Testing

Patients may be at greater risk of death or major dysrhythmia during an exercise test after a myocardial infarction (MI). Risks can be reduced, however, by using a submaximal test for these patients. Benefits of post MI exercise testing include optimising the date of discharge, assessment of drug therapy and evaluation of

prognosis. By involvement of the patient's spouse, through watching the test or even performing one themself, the confidence of both patient and spouse can be increased[45,46].

1.3.5 Patient Follow-Up

After starting a course of drug therapy or after surgical treatment, the exercise test can be used to monitor a patient. There is a tendency for an individual to have improved exercise tolerance in terms of total exercise times as more tests are performed, due to improved efficiency or lower levels of anxiety caused by the test[47]. However, endpoints of anginal pain and ≥1mm ST-segment depression will occur at approximately the same double product or heart rate. Thus, changes in the condition of a patient can be seen in changes in the heart rate or double product at ischaemic endpoints during exercise.

1.3.6 Treatment Evaluation

The comparison of different anti-anginal drug treatments in clinical trials is a major application of exercise testing. The European Agency for the Evaluation of Medicinal Products cite exercise testing as the principal method for assessment of efficacy of anti-anginal products in stable angina pectoris[48].

The aim of this thesis is to review methods of analysis of the various data that are produced during exercise tests, with the objective of highlighting differences between treatments. Most trials are performed under a parallel groups design[49], where subjects are randomised to different treatments, and after a period of stabilisation, are given an exercise test.

Since angina is a comparatively stable condition in which treatments are used to control symptoms rather than to cure, it is also sensible to use crossover designs in their evaluation[50]. Here subjects are given all of the treatments under consideration, in a randomised order, allowing a period between the start of each treatment and the exercise test for the effects of the previous treatment to wear off. Such designs should allow more accurate comparisons of treatment effects, since the effects of all treatments are observed in each individual.

1.4 Exercise Test Data

A number of variables are analysed in trials involving exercise tests. Dichotomous variables include the occurrence or not of anginal pain or significant ST-segment

depression. Exercise times include total exercise time and the time until the onset of anginal symptoms or until significant ST depression. Other variables are heart rate or double product at the end of exercise, at the onset of anginal pain or at the occurrence of significant ST-segment depression.

The exercise times or haemodynamic variables at the occurrence of ischaemia are censored, since there is no guarantee that the ischaemic events will occur. The participants in a study may be chosen according to the criterion that at least one of these ischaemic events occurred during an exercise test when no treatment was administered; it may be that when treatment is given to the same subject, an exercise test will not be able to induce these events. Hence in large trials involving patients whose angina is not severe there may be considerable censoring of these variables.

Initially, comparisons of treatments that used exercise testing looked at exercise times, and used methods such as t-tests or non-parametric equivalents to compare treatment groups. The same techniques were used for haemodynamic variables such as heart rate and double product. Censoring of these variables was not a major problem since the patients studied were mostly those with severe CAD, with most suffering ischaemic events during exercise, even when under treatment. Also, treatments were often compared to placebo, and large treatment effect differences were observed, so that the small bias involved in ignoring the censoring of the response did not affect the conclusions of the trials.

More recently, it has been recognised that these response variables are censored and that to ignore this fact will lead to biased estimates of treatment effect differences. This may be important with larger trials, involving healthier subjects for whom treatment effects may not be so drastic. Also, trials that compare new with standard treatments are becoming more common[51] as the benefits of therapy become accepted, so the expected treatment differences are less. Biased estimating procedures may have less power to detect treatment differences. Methods are needed that take account of the censoring to give as much power as possible to find differences between treatments.

For dealing with time to event data, the obvious methods to use for any analysis are those of survival analysis. Simulations[52] have shown that survival methods are more appropriate than other methods in the presence of censoring. For the heart rate or double product at these events, it should be noted that these variables are also, in a sense, censored, since the events do not always occur.

Given that the onset of anginal pain is a subjective response variable, it is desirable to conduct an analysis of the exercise times until significant ST-segment depression occurs. However, though the ECG is monitored throughout the test, it will only be recorded at certain times, usually at regular intervals, for example, every minute. Most ST-segment depression data will thus be interval censored; that is, most events that occur will be detected as having done so at one of these predetermined recording times. In fact, (assuming that ST-segment depression varies monotonically between recording points) we only know that the event occurred during an interval. Analytic methods that take account of this form of censoring could be applied, but the amount to be gained by this in terms of reduced bias or increased power is unclear.

Also, since we know that within any particular individual, the haemodynamic endpoints are highly reproducible, it would seem that within subject variability with these outcomes is much less than between subjects. Models that allow for such heterogeneity in survival data, known as frailty models, can be used to fit a random effect to the hazard of each individual. These methods may be particularly useful for situations where there are repeated exercise tests, either with baseline tests, or tests on different treatments, as in a trial with a crossover design.

1.5 Summary

Exercise tests can be used for a wide range of patients, but are particularly useful in cardiology. They have many applications for angina sufferers, and are a vital part of the process of evaluating new drug therapies. Drug development depends on accurate analysis of exercise test data, some of which is subject to heavy interval censoring and large between subject variability. This thesis will look at the relative merits of various analytical methods, by looking at their application to real and simulated data.

CHAPTER 2 Clinical Trials of Anti-Anginal Therapies

This thesis will outline the various methods that have been used or could be used to analyse data from exercise tests. These methods will be applied to data from exercise tests carried out during the Total Ischaemic Burden European Trial (TIBET). This Chapter describes the TIBET study and other major studies of angina therapies, with particular emphasis on their use of exercise testing.

2.1 Total Ischaemic Burden European Trial (TIBET)

TIBET was a double-blind, parallel-group clinical trial of three anti-anginal therapies, Atenolol (a β-blocker), Nifedipine (a calcium antagonist) and their combination. The main outcomes were cardiovascular morbidity and mortality. Secondary outcomes were exercise test results and 48-hour Holter monitoring.

The study was innovative in that it examined the long-term effects of medical therapy on morbidity and mortality in patients with chronic stable angina. It was a large study, with over 200 patients assigned to each of the three treatments. Each patient underwent four exercise tests, some of which could be used to compare the effects of treatments. Patients were tested by either a bicycle or treadmill exercise test, and were evenly divided between the two types of exercise.

2.1.1 Design

Men and women, aged 40-79 with chronic stable angina were selected for inclusion in the trial (Figure 2.1). Each underwent a two-week active run-in period, when the combination therapy was given to ascertain if the patient was able to tolerate the treatments under study. After a further two weeks of placebo washout, patients underwent their first exercise test of the study, which was also the final inclusion test. Those who did not demonstrate exercise-induced ischaemia defined as ≥1mm ST-segment depression occurring before 10 METS (metabolic equivalents of oxygen

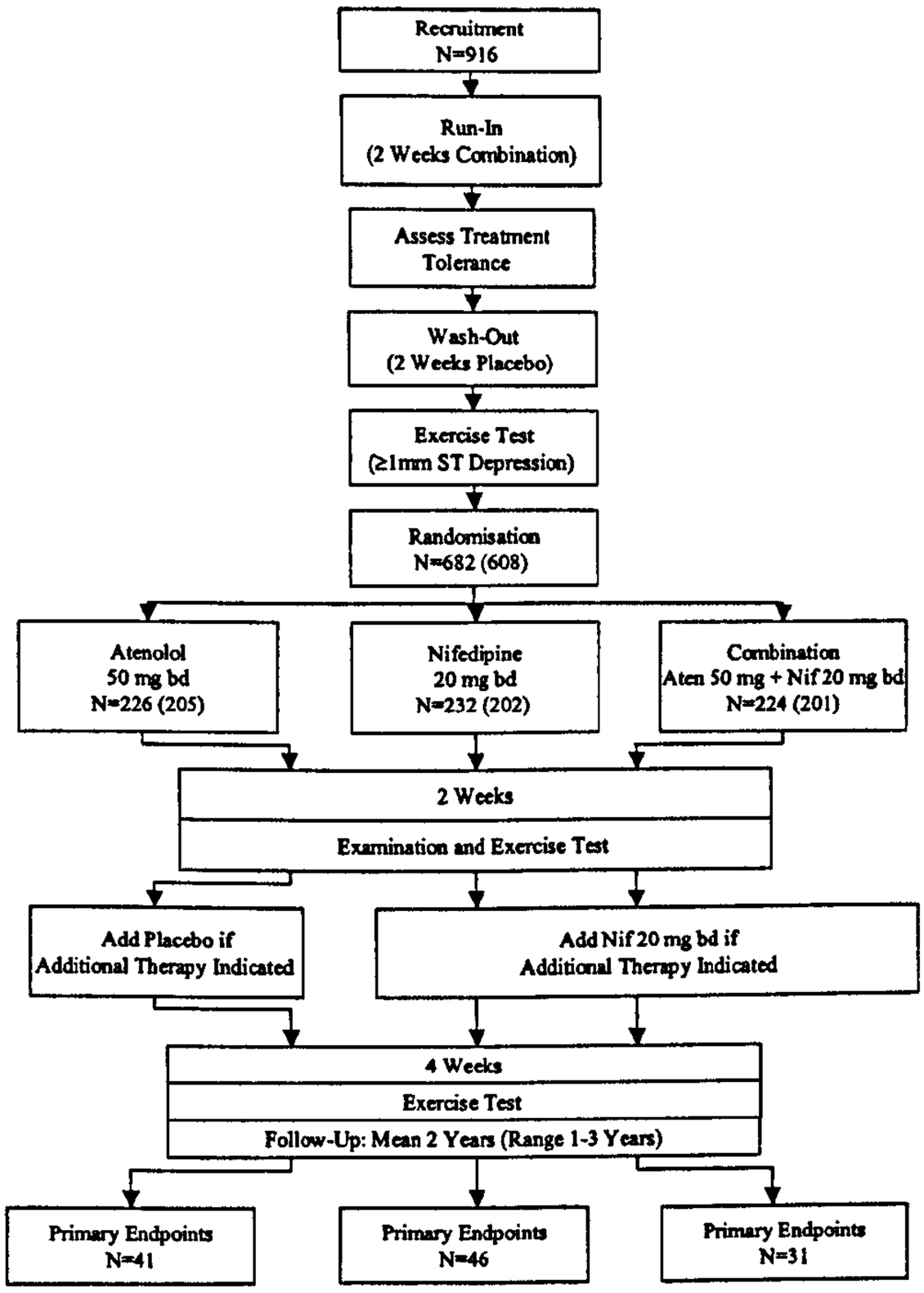


Figure 2.1 Design of the Total Ischaemic Burden European Trial (TIBET)

consumption) were excluded. A more complete description of inclusion and exclusion criteria has been given elsewhere[53].

Those who were selected to take part in the study were then randomised to one of the three treatments; Atenolol (50mg b.i.d.), Nifedipine (20mg b.i.d.) or their combination. Since each treatment had been previously demonstrated to be beneficial, there was no placebo control group. After two weeks of treatment, patients underwent an examination and a second exercise test; those for whom there was clinical indication for increased therapy were given placebo if they were on Atenolol, or a further dose of 20mg Nifedipine if they were on one of the other treatments. A third exercise test took place four weeks later. Patients were followed up with regular visits for an average of 2 years, and were given a final exercise test upon withdrawal from randomised treatment.

2.1.2 Endpoints

Primary endpoints related to morbidity and mortality. Hard endpoints were defined as cardiac mortality, MI and unstable angina. Soft endpoints were defined as CABG, PCTA and treatment failure. Secondary endpoints related to exercise test data and 48 hours of continuous ECG (Holter) monitoring. Endpoints from exercise tests

were time to and exercise capacity at onset of ≥1mm ST-segment depression, onset of anginal pain and termination of exercise. Holter endpoints were number, total duration and circadian distribution of significant ischaemic episodes (defined as ≥1mm ST-segment depression).

2.1.3 Exercise Tests

Approximately half of randomised subjects were given treadmill exercise tests and half used a bicycle. All treadmill tests were conducted according to a standard Bruce protocol[6]. That is, each patient began exercising on a treadmill moving at 1.7mph on a gradient of 10%. Every 3 minutes, both the speed and the gradient of the treadmill were increased: the speed to 2.5, 3.4, 4.2, 5.0, 5.5 and 6.0mph; the gradient increased by 2% at the end of each stage, up to a maximum of 22% at the start of stage 7 (18 minutes). Bicycle tests used a protocol that started at a workrate of 30W and increased by 30W every 3 minutes.

ECG was recorded before exercise commenced, then at the end of every minute of exercise and immediately on stopping; it was also recorded at 1, 3 and 5 minutes after exercise, while the patient was seated and resting. From the ECG recording, the level of ST-segment depression as well as heart rate could be determined. Systolic blood pressure was measured immediately prior to exercise, after every stage (3 minutes) during the exercise test and at the end of exercise.

2.1.4 Baseline Characteristics

Table 2.1 shows some baseline characteristics for the TIBET Study population. There were 682 participants randomised to one of the study treatments, though after subsequent examination of exercise test ECG data, the dataset was reduced to 608 for some analyses. The results given in this thesis are based on the full dataset.

2.1.5 Published Results

The results of the TIBET Study relating to primary[54] and secondary[55] endpoints were published simultaneously in 1996. There were no significant differences between treatment groups in terms of the time to primary endpoint though the trend was towards fewer events with the combination therapy (Atenolol, 47 events; Nifedipine, 46 events; Combination, 31 events; logrank test p=0.14). There were, however, significantly more withdrawals from study medication for those using Nifedipine (Atenolol, 60

		Atenolol	Nifedipine	Combination
N Total		226	232	224
Treadmill	N (%)	119 (52.7%)	119 (51.3%)	116 (51.8%)
Males	N (%)	193 (86.9%)	191 (82.3%)	197 (88.3%)
Age (years)	Mean (SD)	59.3 (7.6)	59.3 (7.5)	59.1 (7.9)
Weight (kg)	Mean (SD)	78.2 (11.0)	75.4 (10.0)	78.7 (11.2)
Previous MI	N	77	71	77
Previous Heart Failure	N	2	4	1
Hypertension	N	52	54	47
Diabetes	N	10	7	18
Previous Angiogram	N	67	62	64
Previous PTCA	N	4	5	5
Previous CABG	N	14	12	9

Table 2.1 Baseline characteristics of TIBET Study population

withdrawals; Nifedipine, 93 withdrawals; Combination 64 withdrawals; logrank test p=0.001).

Though it was reported that survival analysis techniques had been used to analyse exercise times[55], these results were reduced to a single paragraph stating that no significant differences were found. The data were mainly reported in terms of improvement compared to the baseline exercise test of total exercise time, time to pain (or total exercise time, if pain did not occur), time to 1mm ST Segment depression (or total exercise time) and maximum ST segment depression, with data from treadmill and exercise tests reported separately. Also given, for both types of exercise combined, were changes from baseline heart rate, systolic and diastolic blood pressure, both sitting and standing. These data are summarised in Table 2.2.

Atenolol and Combination were associated with reductions in resting heart rate, whilst Nifedipine was found to cause a slight increase. All treatments caused reductions in blood pressure, though these were greater for combination therapy. All treatments

		Atenolol	Nifedipine	Combination
	N	177	175	162
	Sitting HR (bpm)	15.4 (0.8)	-2.9 (0.8)	13.5 (0.8)
	Sitting SBP (mmHg)	12.9 (1.2)	12.5 (1.3)	20.6 (1.3)
	Sitting DBP (mmHg)	7.3 (0.6)	6.7 (0.8)	11.0 (0.8)
	Standing HR (bpm)	17.2 (0.8)	-3.7 (0.9)	15.6 (0.9)
	Standing SBP (mmHg)	12.8 (1.2)	13.6 (1.2)	20.4 (1.2)
	Standing DBP (mmHg)	7.5 (0.7)	6.5 (0.7)	11.4 (0.8)
Treadmill	N	91	87	80
	Total exercise time (s)	91.4 (10.0)	90.5 (11.1)	98.0 (11.7)
	Time to pain (s)	128.0 (11.3)	126.7 (15.0)	144.3 (13.7)
	Time to 1mm ST depression (s)	136.4 (12.4)	131.4 (13.9)	147.3 (11.3)
	Maximum ST depression (mm)	0.43 (0.08)	0.49 (0.09)	0.50 (0.11)
Bicycle	N	86	88	82
	Total exercise time (s)	63.2 (11.0)	63.6 (13.3)	78.5 (17.3)
	Time to pain (s)	106.8 (15.4)	109.4 (15.3)	138.8 (17.9)
	Time to 1mm ST depression (s)	147.4 (14.4)	146.5 (15.6)	162.7 (17.6)
	Maximum ST depression (mm)	0.59 (0.08)	0.74 (0.08)	0.76 (0.10)

Table 2.2 Reductions baseline sitting and standing heart rate (HR), systolic blood pressure (SBP) and diastolic blood pressure (DBP) at rest, and increases from baseline exercise times (total exercise time, time to pain (or total exercise time), time to 1mm ST segment depression (or total exercise time)) and maximum ST segment depression, reported as mean (SE)

were associated with increased exercise times and reduced ST-segment depression during exercise, and though none of these improvements were significantly different between treatment groups, there was a tendency for those on combination therapy to perform better than either single treatment.

2.2 Previous Studies

Clinical studies of the treatment of stable angina concentrate on endpoints such as myocardial infarction (MI) and death (coronary or all cause) to measure the quantity of life, and anginal symptoms and the occurrence of ischaemia to measure the quality of life. Exercise tests are used in the assessment of the latter of these objectives.

2.2.1 Drug Therapies to Prevent MI and Death

Many treatments used in patients with stable angina are targeted at the prevention of MI and death, which is the primary concern in the management of these patients. However, they do not directly influence the myocardial response to exercise and so will be dealt with briefly here.

Aspirin has antithrombotic effects and is recommended as a routine therapy for patients with stable angina[32], since the risk of adverse cardiovascular events has been shown in randomised trials to be reduced by about a third in this way[56]. Angiotensin Converting Enzyme (ACE) Inhibitors are now advocated for patients with or at high risk of developing coronary artery disease (CAD), particularly diabetics[33]; a 20% reduction in the incidence of MI was observed in the HOPE Study[57] which included 5162 (55.5%) sufferers of stable angina. None of these studies reported results of exercise tests, since the reduction of anginal symptoms was not of concern.

It has been observed that fibrinolytic function is associated with subsequent cardiovascular mortality in patients with stable angina[58], and despite evidence of reduced coronary events after treatment with the anticoagulant warfarin in patients without stable angina[59], there is as yet no firm evidence of reduced mortality amongst symptomatic patients. However, low-molecular-weight heparin has been shown to reduce fibrinogen levels and increase exercise time to 1mm ST-segment depression[60]. This study involved 29 patients and analysed exercise times to 1mm ST-segment depression, peak exercise and moderate anginal symptoms, as well as the maximal double product achieved and the double product at 1mm ST-segment depression. Analyses were performed using t-tests and ANOVA methods, though no mention was made of whether all participants achieved the stated endpoints during the on-treatment exercise test. If it occurred, censoring may have been minor, since all participants were required to achieve ischaemic endpoints on two baseline exercise tests

Therapies that aim to reduce risk factors for coronary events have also been shown to reduce subsequent coronary event rates. Lipid lowering with statins has been associated with a reduction in mortality and coronary events of approximately one third[61,62] in large randomised controlled trials that included sufferers of stable angina. Similarly, those with stable angina and hypertension benefit from antihypertensive treatment[63]. These trials have again focused on reductions in morbidity and mortality, without measuring exercise tolerance.

2.2.2 Drug Therapies to Prevent Anginal Symptoms

The main classes of pharmacotherapy for ischaemia and angina, namely β-blockers[64], calcium antagonists[65], nitrates[66] and potassium channel openers[67,68], have all been shown to improve exercise tolerance. The statistical methods used in these studies have been based on normal theory methods, such as t-tests and ANOVA. A variety of

endpoints were used, including time to angina (onset or exercise-limiting), time to ≥1mm ST-segment depression, total exercise time and double product at maximal exercise. One of these studies used individualised exercise protocols for each subject[64], another used log-transformed exercise times[65] and a third analysed changes in exercise times compared to a baseline exercise test[66].

In general, patients with stable angina should be administered sublingual nitroglycerin spray or tablets and be educated in their use for the immediate relief of symptoms and to prevent pain when this can be anticipated[32]. For the long-term prevention of symptoms, the initial treatment choice would be β-blockers[32], though in patients who are intolerant, each of the other classes of treatment may be used as monotherapy, though oral nitrates should be used in a way that avoids nitrate tolerance.

Undesirable side effects may be reduced by combination therapy, and there is evidence of beneficial effects of adding isosorbide mononitrate[69] or a long-acting dihydropiridine[70] (calcium antagonist) to a β-blocker. Studies have compared times to ≥1mm ST-segment depression, levels of ST-segment depression at comparable workload or maximal workload, and maximal ST-segment depression. The method of analysis has been the t-test. Other therapeutic combinations are not recommended, nor is treatment with more that two agents; patients for whom treatment fails to control symptoms may be candidates for surgical intervention.

2.2.3 Surgical Intervention

Surgical interventions for patients with CAD, whether by CABG or PCI, aim to increase life expectancy and relieve symptoms. The survival of patients treated with CABG is greater than that of patients treated pharmacologically[71] at first, particularly for those with the most severe disease, who were at the greatest risk of death without surgery. Approximately 80% of CABG patients are free from angina 5 years after surgery[72]; CABG offers a greater level of symptom relief than initial medical treatment at 5 years and, to a lesser extent at 10 years, though more than 40% of patients initially treated medically will have undergone CABG within that time[71].

In patients with mild symptoms, who are at relatively low risk of coronary death, PCI offers a greater reduction in symptoms than medical treatment, but at a slightly higher rate of future coronary events[73]. Both CABG and PCI provide better symptom relief than medical treatment in patients with more severe, single vessel disease[74], demonstrated using the exercise test endpoint of the occurrence or not of an ischaemic

event. All three methods have similar rates of mortality and MI[74], though CABG offers the lowest risk of future coronary events including surgical interventions. In patients with multivessel disease, CABG and PCI result in similar rates of survival from mortality, MI or stroke, and despite higher rates of additional surgical interventions following PCI, the cost per patient is lower[75]. Only in patients at the highest risk of death does CABG offer lower mortality rates, for example in diabetics[76] and/or those with severe multivessel disease[77].

2.2.4 Literature Review

To assess the statistical methods used to analyse time to event data from exercise tests in situations similar to those seen in the TIBET Study, a brief literature search was performed. Using the PubMed search system[78], a search was carried out using the expressions "exercise test", "clinical trial" and "angina". Searches were limited to English language articles and studies of human subjects. To determine whether there have been any changes in the statistical methods used with these data, the same search was performed for articles published during 1983, 1993 and 2003.

A total of 105 articles were identified (31 from 1983, 52 from 1993 and 22 from 2003). Six were excluded as they were review articles, editorials or meta analyses. On closer inspection, 30 were found not to include any analyses of exercise times. The remaining 69 articles were inspected to determine whether survival methods were used to analyse exercise times, or methods that ignore the censored nature of the data, such as those based on Normal theory (t-tests, ANOVA, regression), or similar non-parametric methods.

Table 2.3 summarises the findings of this literature search. Amongst these articles, the overwhelming majority were found to use methods of analysis suitable for uncensored continuous data, predominantly those based on a Normal distribution. Only two papers[79,80] presented results of survival analysis of exercise times. Equally few used neither method[81,82], reporting comparisons of exercise times in terms of the numbers of participants showing improvement in exercise time under treatment. Though the sample is limited to only three years' publications, they seem to indicate a considerable preference towards the more widely used statistical techniques.

A number of factors might be influencing this apparent lack of use of survival analysis methods with censored time to event data from exercise tests. In medical research, many small-scale studies are carried out by individual or small groups of

Year	Method of Analysis				Total
	Both	Survival	Uncensored	Neither	
2003	0	0	10	0	10
1993	1	1	31	0	33
1983	0	0	24	2	26
Total	1	1	65	2	69

Table 2.3 Numbers of articles published in 2003, 1993 or 1983, analysing exercise times using survival analysis and/or methods for uncensored continuous data (based on the Normal distribution or non-parametric equivalent) methods, by year of publication

clinicians. In such cases, the statistical analyses may be conducted by those without the breadth of statistical training to recognise the need for survival analysis. Furthermore, based on previous studies, it would appear that survival techniques are not the standard method of analysis for such data, and an individual confident in the use of t-test or ANOVA methods would be able to report the findings of their study.

In larger trials, in which more experienced statisticians are employed to perform the analyses, the use of survival methods may be under reported. If methods for both censored and uncensored data are used, it may be the case that the results, whether showing statistically significant differences or not, result in similar conclusions being reached. Since, in general, the readership of such articles will be, or be perceived to be (by the authors of the articles or the editors of the journals) less able to comprehend analyses presented using survival techniques, such analyses may be omitted from articles. It may even be the case that studies analysed using survival analysis methods are less likely to be accepted for publication, particularly when the results are negative.

Finally, in many studies, multiple exercise tests are employed, and the results are presented in terms of the change in exercise time following intervention. Methods for the analysis of repeated survival data are relatively new, particularly with respect to their application using standard statistical software packages. As a result, their use may be limited to those with an interest in current research into statistical methodology, particularly academic statisticians. Statisticians working in medical research may be less aware of developments in methodology. They will be working to more stringent timescales, with the emphasis on producing coherent results on time, with less scope to experiment with alternative methods. Furthermore, they will have to report their results to clinical colleagues in formats that are clearly understood, a constraint that often

precludes the use of more novel approaches, even when these are superior in terms of efficiency or validity.

CHAPTER 3 Estimation of Treatment Effect

Differences I: Cox Proportional Hazards Models

Clinical trials of angina therapies are required to include exercise testing to demonstrate differences in treatment efficacy[48]. In order that test results are comparable between subjects, they should be performed to a standard protocol, or performed to the same protocol on several occasions to demonstrate changes in individual exercise performance. There are numerous response variables that can be recorded, but primary analyses will often be carried out using the time spent exercising until the occurrence of chest pain and/or significant ST-segment depression, or the total exercise duration.

There may be a single exercise test per subject, or a series of tests to measure responses to exercise before and during active treatment. In a randomised trial, subjects will be allocated to receive one of the candidate treatments (in a parallel groups design) or different treatments during different periods of the study (in a crossover design) at random. Since allocations are random, differences in average exercise response between treatment groups can be attributed to differences in treatment effects, subject to the limits of uncertainty due to chance differences between the groups.

It will often be the case that other patient information will be available. In a clinical trial setting, it may be of interest to test whether intervention effects are homogeneous across subgroups of the study population. In clinical epidemiology, it may be necessary to adjust for the potentially confounding effects of factors other than the exposure of interest.

This would generally be achieved through regression models, whereby treatment and covariate effects are estimated simultaneously. It is possible to incorporate covariates that change over time, but this thesis shall only consider models in which covariates are fixed, such as age, sex and weight. Variables that change during exercise, such as heart rate or blood pressure, are undoubtedly connected with the risk of

suffering an ischaemic event, but are also affected by workload, and so to include them in models of the time to an ischaemic event could be misleading.

3.1 Survival Analysis

Survival data consist of survival times, failure indicators and covariate data. That is, $(t_i, \delta_i, \mathbf{z}_i)$ for $i = 1, 2, ..., n$, where $\delta_i = 0$ if the i^{th} subject is censored at t_i and $\delta_i = 1$ if the subject suffers an event at t_i.

3.2 Proportional Hazards Regression Models

Proportional hazards regression models relate the hazard function for an individual to a common baseline hazard function through multiplication by a positive function of that individual's covariates. That is, if $\lambda(t|\mathbf{z})$ is the hazard function for an individual with covariates $\mathbf{z}$, and $\lambda_0(t)$ is the baseline hazard function, then $\lambda(t|\mathbf{z})=\lambda_0(t)\psi(\mathbf{z},\beta)$. The vector β is a set of parameters, at least some of which it may be of interest to estimate.

A common form for the function $\psi(\mathbf{z},\beta)$ is $\exp(\mathbf{z}\beta)$, so that an element of β, β_j say, has the interpretation of being the log hazard ratio associated with a unit increase in the corresponding covariate, z_j, with all other covariates held constant. In particular, if a binary covariate is represented by a {0, 1} indicator variable, then the corresponding parameter is the log hazard ratio between the subgroups. The effects of categorical variables of more than two levels can be estimated by the construction of dummy indicator variables, in exactly the same way as for other linear models.

3.3 Cox Proportional Hazards Regression Model

Different proportional hazards regression models make different assumptions about the baseline hazard function, $\lambda_0(t)$. There are a number of fully parametric formulations for $\lambda_0(t)$, some of which are demonstrated in Section 4.1; however, the Cox proportional hazards model[83], which makes no assumptions about $\lambda_0(t)$, is the most commonly used[84]. It is termed semi-parametric in the sense that, while the baseline hazard is unspecified and is estimated non-parametrically, the covariate effects are modelled parametrically.

3.4 Parameter Estimation

The likelihood of a set of survival data can be written as

$$L(\theta,\beta|\{t_i,\delta_i,\mathbf{z}_i : i=1,2,...,n\}) = \prod_{i=1}^{n} S(t_i|\theta,\beta,\mathbf{z}_i)\lambda(t_i|\theta,\beta,\mathbf{z}_i)^{\delta_i}$$

$$= \prod_{i=1}^{n} \exp(-\Lambda(t_i|\theta,\beta,\mathbf{z}_i))\lambda(t_i|\theta,\beta,\mathbf{z}_i)^{\delta_i}$$

$$= \prod_{i=1}^{n} \exp(-\Lambda_0(t_i|\theta)\psi(\mathbf{z}_i,\beta))\{\lambda_0(t_i|\theta)\psi(\mathbf{z}_i,\beta)\}^{\delta_i}$$

where θ is the set of parameters for the baseline hazard function. Under the Cox proportional hazards model, this is

$$L(\theta,\beta|\{t_i,\delta_i,\mathbf{z}_i : i=1,2,...,n\}) = \prod_{i=1}^{n} \exp(-\Lambda_0(t_i|\theta)\exp(\mathbf{z}_i\beta))\{\lambda_0(t_i|\theta)\exp(\mathbf{z}_i\beta)\}^{\delta_i}$$

so that the log likelihood can be written as

$$l(\theta,\beta|\{t_i,\delta_i,\mathbf{z}_i : i=1,2,...,n\}) = \sum_{i=1}^{n}\{-\Lambda_0(t_i|\theta)\exp(\mathbf{z}_i\beta) + \delta_i(\log\lambda_0(t_i|\theta) + \mathbf{z}_i\beta)\}.$$

3.4.1 Partial Likelihood

When fitting the Cox proportional hazards model, the log likelihood cannot be maximised directly, since the functional form of the baseline hazard function is unknown, and an alternative method must be adopted. The partial likelihood function is defined as

$$L_p(\beta|\mathbf{z}) = \prod_{i=1}^{k} \frac{\exp(\mathbf{z}_{(i)}\beta)}{\sum_{j\in R(t_{(i)})}\exp(\mathbf{z}_{(j)}\beta)} \qquad \textbf{(Eq. 3.1)}$$

where the $t_{(i)}$, i = 1, 2, ..., k, are the ordered failure times and $R(t_{(i)})$ denotes the risk set at time $t_{(i)}$, or those subjects whose survival times are at least as large as $t_{(i)}$. The term $\frac{\exp(\mathbf{z}_{(i)}\beta)}{\sum_{j\in R(t_{(i)})}\exp(\mathbf{z}_{(j)}\beta)}$ is equivalent to $\frac{\lambda(t_{(i)}|\mathbf{z}_{(i)})}{\sum_{j\in R(t_{(i)})}\lambda(t_{(j)}|\mathbf{z}_{(j)})}$, and can be interpreted as the conditional probability that subject (i) fails at time $t_{(i)}$ given that one of those under observation at $t_{(i)}$ fails at that time. Maximum partial likelihood estimates have similar properties to normal maximum likelihood estimates[83], and asymptotic variances of the parameter estimates are derived from the second derivative of the log partial likelihood function.

The partial likelihood function as defined by (Eq. 3.1) is valid only when all failure times are distinct. In practice, failure times are not unique, since times are often rounded to convenient values. For example, total exercise time, or the time to occurrence of anginal pain, might be recorded to the nearest 10 seconds. What is more, the time to occurrence of significant ST-segment depression can only be one of those times when an ECG trace was recorded, which might be every minute.

In situations where failure times are tied, the partial likelihood is defined as

$$L_t(\beta|z)=\prod_{i=1}^{k}\frac{\exp(s_{(i)}\beta)}{\sum_{R\in R(t_{(i)},d_{(i)})}\sum_{j\in R}\exp(z_{(j)}\beta)} \qquad \textbf{(Eq. 3.2)}$$

where $d_{(i)}$ is the number of events occurring at time $t_{(i)}$, $s_{(i)}$ is the sum of covariate vectors of individuals observed to fail at time $t_{(i)}$ and $R(t_{(i)}, d_{(i)})$ denotes the set of all sets of size $d_{(i)}$ drawn from $R(t_{(i)})$. An individual term in this product can be interpreted as a conditional probability, that it was those subjects who were observed to fail at that time who did so, given that $d_{(i)}$ individuals fail at time $t_{(i)}$.

With large datasets or large numbers of ties, the number of calculations required to evaluate (Eq. 3.2) can make its maximisation slow, and approximations may be used to speed up the calculations. When there are few ties at each failure time, the partial likelihood is well approximated (Breslow[83]) by

$$L_b(\beta|z)=\prod_{i=1}^{k}\frac{\exp(s_{(i)}\beta)}{\left\{\sum_{j\in R(t_{(i)})}\exp(z_{(j)}\beta)\right\}^{d_{(i)}}}. \qquad \textbf{(Eq. 3.3)}$$

However, when the number of ties at any failure time is large, (Eq. 3.3) can produce biased estimates, and a better approximation (Efron[85]) is given by

$$L_e(\beta|z)=\prod_{i=1}^{k}\left\{\exp(s_{(i)}\beta)\Big/\prod_{r=1}^{d_{(i)}}\left[\sum_{j\in R(t_{(i)})}\exp(z_{(j)}\beta)-\frac{(r-1)}{d_{(i)}}\sum_{j\in D(t_{(i)})}\exp(z_{(j)}\beta)\right]\right\}. \qquad \textbf{(Eq. 3.4)}$$

Example 3.1 TIBET Study, Cox Proportional Hazards Models for Time to Anginal Pain

Table 3.1 shows the treatment effect estimates obtained by fitting proportional hazards regression models to the time until anginal pain under exercise (with 95% confidence intervals and p-values) from the TIBET study. For each model, the estimates were obtained using the Efron approximation to the partial likelihood; almost exactly the same values were achieved using the Breslow approximation, in terms of effect

	Unadjusted Model			Adjusted Model		
	Haz. Ratio	95% CI	p	Haz. Ratio	95% CI	p
Nifedipine–Atenolol	1.36	(0.96, 1.91)	0.081	1.32	(0.94, 1.87)	0.11
Nifedipine–Combination	1.06	(0.74, 1.52)	0.75	1.06	(0.74, 1.52)	0.76
Age (/10 years)				1.19	(0.97, 1.45)	0.093
Gender (Female–Male)				1.19	(0.75, 1.91)	0.46
Weight (/10 kg)				0.96	(0.83, 1.12)	0.60
-2 log likelihood	1943.2			1939.1		

Table 3.1 Treatment and covariate effect estimates from Cox proportional hazards models for the time to anginal pain

estimates (and CIs) and the values of the log likelihoods of the fitted models. Each model allows different baseline hazard functions for the different types of exercise, and effect estimates are shown from models that do not adjust for other covariates as well as models adjusting for gender, age (as a linear effect) and body weight (linear effect). The changes in –2 log likelihood after inclusion of covariate effects (not shown) are small, thus suggesting that none of these variables improves the fit of the model.

3.5 Goodness-of-Fit

When fitting any regression model, a number of assumptions are made, and the extent to which the data deviate from these assumptions should be checked. Serious lack of fit invalidates any inferences drawn and would indicate that the model needs modification, if not complete respecification. However, study reports often do not refer to any assessment of goodness-of-fit.

Under a proportional hazards regression model, it is particularly important to check the appropriateness of the linear predictor, $\mathbf{z}\beta$, and the validity of the proportional hazards assumption. Whilst the Cox model in particular and proportional hazards models in general are readily available as part of many statistical packages, methods for checking the goodness-of-fit are more limited. However, some have been implemented within standard software, and by manipulation of model outputs it is possible to carry out a number of checks of model assumptions.

3.5.1 Linear Predictor

An implicit assumption under the proportional hazards regression model in its usual form, $\lambda(t|\mathbf{z})=\lambda_0(t)\exp(\mathbf{z}\beta)$, is that equal increments in a continuous covariate, z_j, are associated with equal proportional increases in hazard.

This assumption can easily be checked by a simple graphical technique. The covariate is divided into categories, for example into quartiles or quintiles, and the model is refitted including the categorical variable in place of the continuous variable. When the effect estimates for the levels of the new variable are plotted against the median value of the continuous variable within each category, any severe departures from the assumption of linearity will become apparent.

Alternatively, the assumption may be checked statistically. Including a quadratic term in the covariate may be used to detect simple departures from linearity, though in the same way as with linear regression models, parameter estimates are more stable and more interpretable if covariates are centered by subtraction of a suitable value, such as the mean or median. A more flexible method of determining non-linearity of covariate effects is through the use of cubic splines[86], in which a smooth function of the covariate is estimated to represent the association between the covariate and the hazard. A likelihood ratio test can then be used to test whether there is any statistically significant departure from non-linearity for that covariate. These methods have been incorporated into standard software[87].

3.5.2 Proportional Hazards Assumption

The model assumption that receives the greatest attention in proportional hazards models is that of proportional hazards, and many methods have been proposed for checking this assumption. Some of these are applicable under any proportional hazards model, whilst some are exclusive to the Cox proportional hazards model.

Many of the global goodness-of-fit tests for the Cox proportional hazards model are related to each other. The methods presented here are some of the more frequently recommended, ranging from the simplest graphical checks to more formal significance tests.

3.5.2.1 Log Cumulative Hazards Plots

This method is applicable to any proportional hazards model for survival data. If the proportional hazards assumption holds, subjects in the separate treatment groups

will have hazard functions that are proportional to the same baseline hazard function. If there are J treatment groups,

$$\lambda_j(t) = \gamma_j \lambda_0(t)$$

where j=1, 2, ..., J, and the γ_j are constants, so that

$$\begin{aligned} \log \Lambda_j(t) &= \log \int_0^t \lambda_j(u)\,du \\ &= \log\left\{ \gamma_j \int_0^t \lambda_0(u)\,du \right\} \\ &= \log \gamma_j + \log \Lambda_0(t). \end{aligned} \qquad \textbf{(Eq. 3.5)}$$

and the curves of $\Lambda_j(t)$ against t are parallel. A check of proportional hazards is given by a plot of $\log(\hat{\Lambda}_j(t))$ against t for each treatment group; severely non-parallel curves indicate non-proportional hazards.

$\Lambda_j(t)$ is estimated by first generating Kaplan Meier estimates of the survivor functions, $\hat{S}_j(t)$, for each treatment group. $\hat{\Lambda}_j(t)$ is then evaluated as $-\log \hat{S}_j(t)$. It can be useful to smooth the log cumulative hazard curves as an aid to assessing parallelism. However, this assessment can be difficult even after smoothing, and by plotting the differences between pairs of curves, $\log \hat{\Lambda}_{j_1}(t) - \log \hat{\Lambda}_{j_2}(t)$, against time, a clearer impression is given since these curves should be constant if the proportional hazards assumption holds, as

$$\log \hat{\Lambda}_{j_1}(t) - \log \hat{\Lambda}_{j_2}(t) = \log \frac{\gamma_1}{\gamma_2}.$$

These methods can be employed with many statistical packages that generate Kaplan Meier curves, but there are difficulties in assessing parallelism of curves, or constancy of differences between curves. It is not obvious how much non-parallelism will occur due to random variations in survival or in covariates that affect survival between groups.

Example 3.2 Logged Cumulative Hazards Plots for Cox PH Model of TIBET Time to Anginal Pain

Figure 3.1 shows log cumulative hazards plots for the time until the onset of anginal pain for patients exercising on a bicycle ergometer. Plot (a) is the estimated log

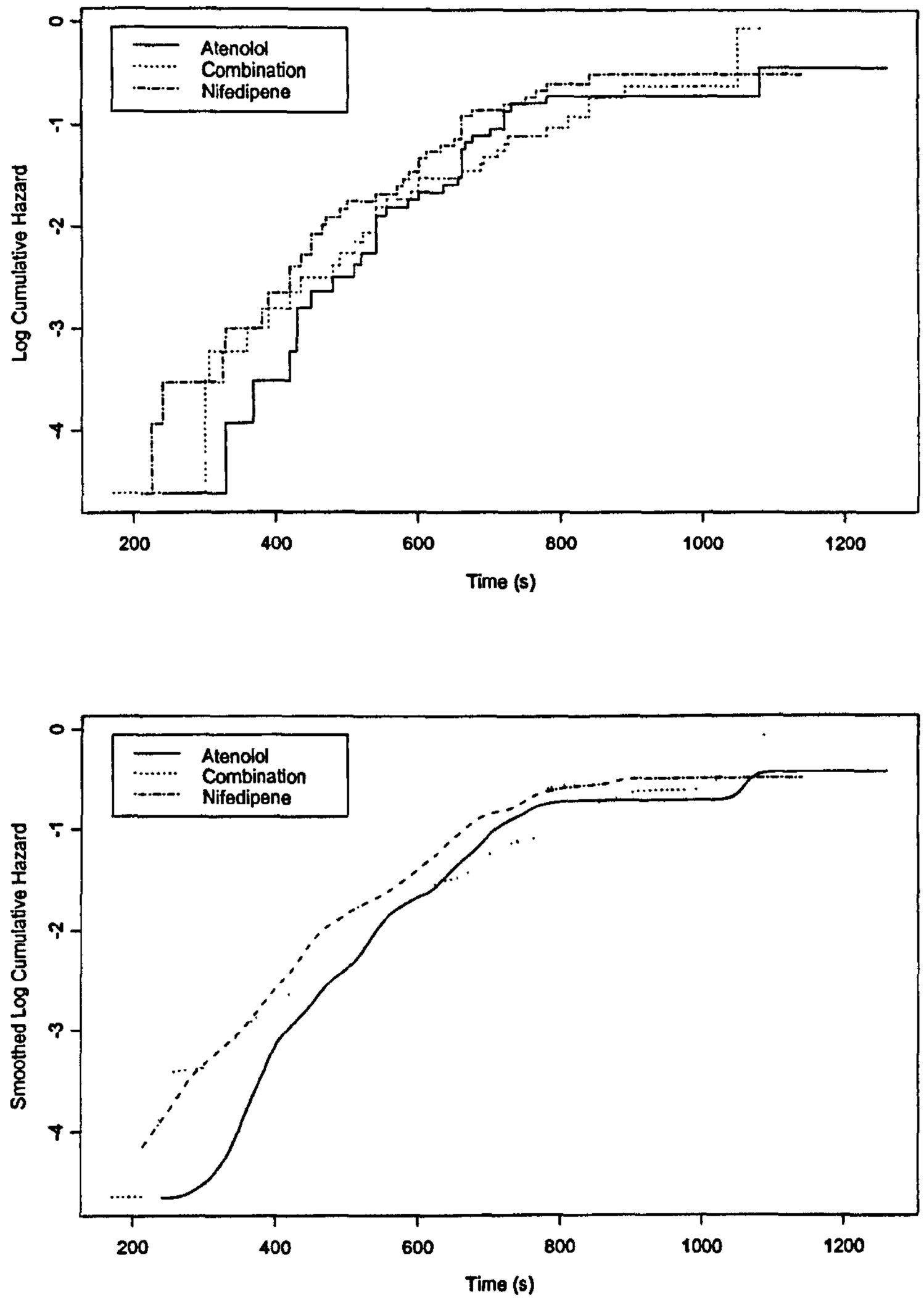
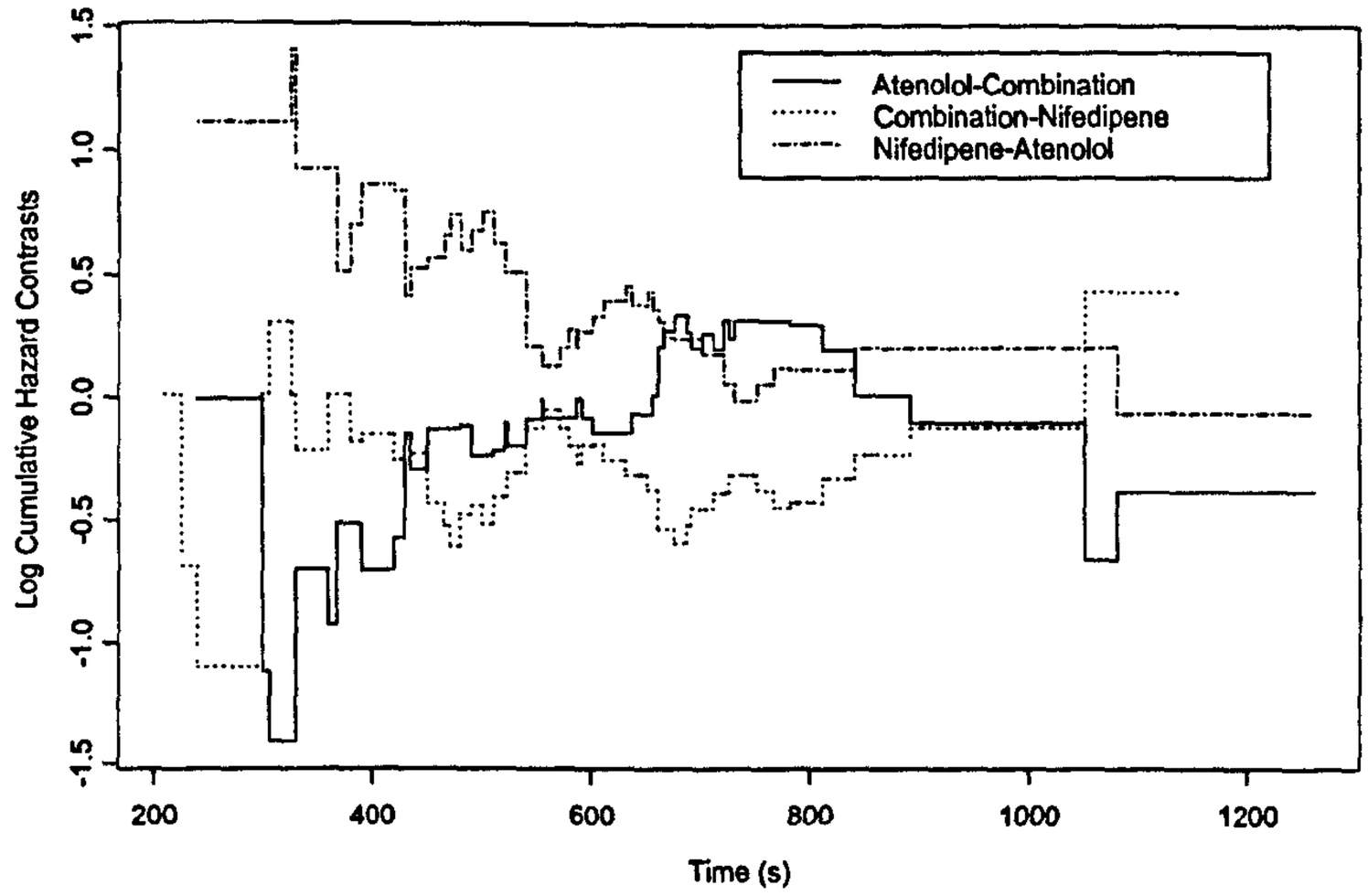


Figure 3.1 Log cumulative hazards plots for time to anginal pain during bicycle exercise: (a) raw estimates; (b) smoothed estimates.

cumulative hazard, $\log \hat{\Lambda}(t)$, plotted against t for each treatment. If the hazards in the three groups are proportional, these curves would be parallel, due to (Eq. 3.5); this is not easy to assess, since the curves are step functions. In (b), a smooth version of plot (a) is shown, where a normal kernel smoother with a bandwidth of 60 seconds[88] has been used. The question of parallelism is still difficult to judge, however, since the lines are curved and are close together relative to the range of the y-axis.

Figure 3.2 shows differences in log cumulative hazards against t for each pairwise treatment comparison. The steps in plot (a) confuse the issue, and a smooth of this plot is given in (b). If the hazards are proportional, the lines should be horizontal, and this



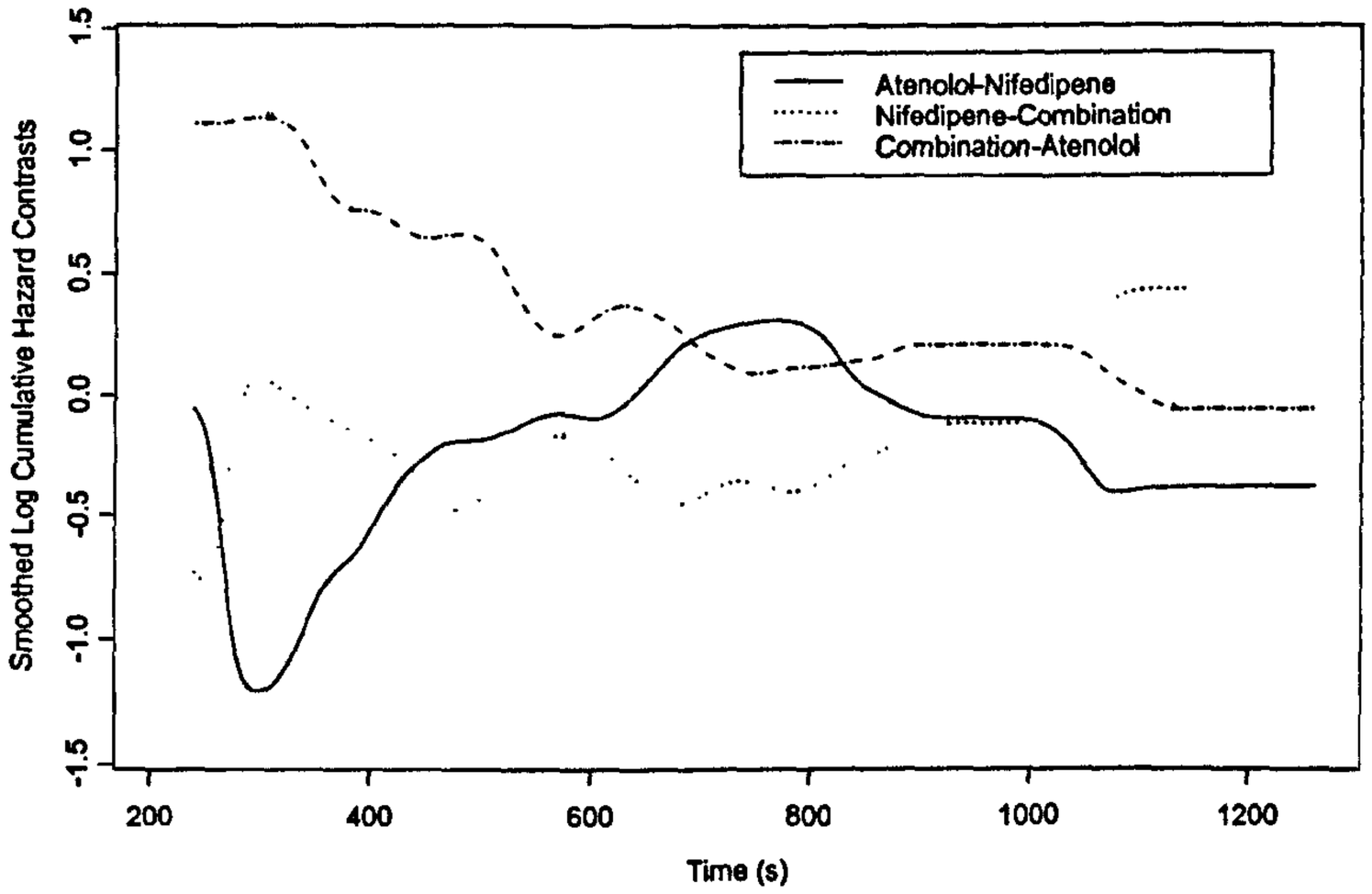


Figure 3.2 Log cumulative hazards contrast plots for time to anginal pain during bicycle exercise: (a) raw estimates; (b) smoothed estimates.

plot is perhaps the most informative; the same graph is shown in Figure 3.3 for the time until anginal pain for those who exercised on a treadmill.

The main difficulty in interpreting these plots is in deciding the extent of any departures from horizontal lines. In general, a plot would to be interpreted as showing non-proportionality if there was a clear and consistent trend over the time axis. For the bicycle data, Figure 3.2(b) gives some cause for concern but is not equivocal, and further investigation will be required; the treadmill data shown in Figure 3.3 do not indicate non-proportional hazards, with the lines departing from horizontal during the initial part of the time axis only.

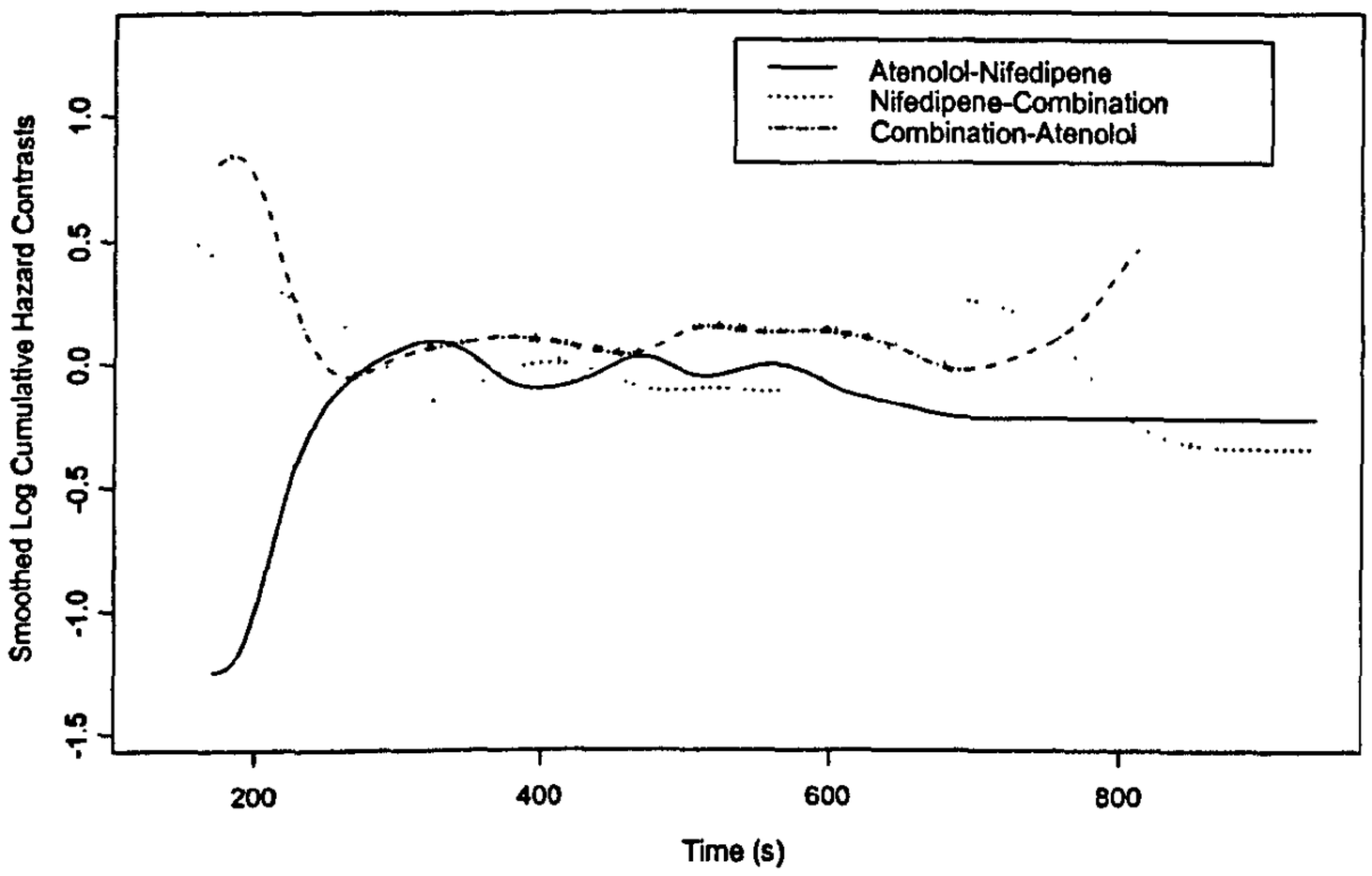


Figure 3.3 Smoothed log cumulative hazard contrast plots for time to anginal pain during treadmill exercise

3.5.2.2 Time Dependent Covariates

Implicit in the proportional hazards assumption is that hazard ratios between groups are constant over time. A test of the proportional hazards assumption is achieved by adding time dependent covariates to the model and testing if this significantly improves the fit.

To illustrate, consider a trial with two treatments, A and B, represented by an indicator variable z that takes the values 0 and 1 according to whether treatment A or B is administered. By fitting the model

$$\lambda(t|z) = \lambda_0(t)\exp\{\beta_1 z + \beta_2 z(\log(t) - \gamma)\}$$

(where γ = mean log failure time), the hazard ratio between treatment groups B and A under this model will be $\exp(\beta_1 - \beta_2\gamma)t^{\beta_1}$. Since $\exp(\beta_1-\beta_2\gamma)>0$, this ratio will increase over time if $\beta_2>0$ or decrease if $\beta_2<0$. A similar test of the proportional hazards assumption can be achieved by including any covariate of the form zf(t), where f(t) is a monotonic function of time.

This method has the advantage that it involves a formal significance test for a monotonic trend in hazard ratio, though its power to detect non-proportional hazards will be dependent upon the choice of functional form of the time dependent covariate. In particular, it will be less effective if the hazard ratio between groups does not follow

Exercise Type	-2 log likelihood		Difference ($\sim\chi^2_2$)	p-value
	Without time dependent covariates	With time dependent covariates		
Bicycle	933.88	933.13	0.75	0.69
Treadmill	1455.03	1454.56	0.47	0.79

Table 3.2 Results of adding time dependent covariates to Cox models for the time to anginal pain

a monotonic trend; in such instances a further time-dependent covariate, including a quadratic function of time, might be considered.

Example 3.3 Time Dependent Covariates in Cox PH Model for TIBET Time to Anginal Pain

The TIBET study involved the comparison of three treatment regimes. The proportional hazards model for the time until the onset of anginal pain, allowing for the effects of treatments can be written as

$$\lambda(t|\mathbf{z}) = \lambda_0(t)\exp\{\beta_1 z_1 + \beta_2 z_2\}, \qquad \textbf{(Eq. 3.6)}$$

where z_1 and z_2 are indicator variables for whether the patient is receiving the first or second treatments. The effects of the third treatment are absorbed into the baseline hazard, $\lambda_0(t)$, and the parameters β_1 and β_2 estimate effects of the first and second treatments relative to that of the third. Assuming that z_1 identifies treatment with atenolol and z_2 nifedipine, $\lambda_0(t)$ will be the hazard function of an individual on the combination therapy.

Adding time dependent covariates involves fitting the model

$$\lambda(t|\mathbf{z}) = \lambda'_0(t)\exp\{\beta'_1 z_1 + \beta'_2 z_1(\log(t)-\gamma) + \beta'_3 z_2 + \beta'_4 z_2(\log(t)-\gamma)\} \qquad \textbf{(Eq. 3.7)}$$

where γ is the mean log failure time. Evidence that either β'_2 or β'_4 are non-zero would indicate that there are time trends in the hazard ratios between atenolol and combination or between nifedipene and combination respectively. The change in -2(log likelihood) between the models represented by (Eq. 3.6) and (Eq. 3.7) will follow a χ^2_2 distribution if the hypothesis of proportional hazards is true.

Table 3.2 shows the improvement in fit as measured by the change in -2(log likelihood) caused by including time dependent covariates into the Cox models. There is

no sign that any of the more complex models gives a better fit, so there is no evidence to suggest that the proportional hazards assumption is incorrect.

3.5.2.3 Distinct Time Epochs

This method can be used to give a global test of the proportional hazards assumption, as well as providing a graphical representation of changing hazard ratios over time. Briefly, the time axis is divided into distinct intervals, or epochs, and the model is fitted to the data in each epoch. As a global test of the proportional hazards assumption, the individual models are compared to the model applied to the full dataset. Graphically, the regression coefficients for each model parameter can be plotted against time, to look for patterns in hazard ratios over time. The times that divide the epochs would normally be chosen to keep roughly equal numbers of events in each dataset, though in some situations, particularly for the purposes of presentation, natural dividers might be more convenient, such as every year in a long-term follow-up study.

Splitting the time axis into two epochs, the cut-point might be the median failure time, $T_{0.5}$, so that approximately half of the observed failures fall into each epoch. Let the full dataset be denoted by $\{T_i, \delta_i: i=1, 2, \ldots, n\}$.

The set of data corresponding to the first epoch will contain all subjects, with survival times right-censored at $T_{0.5}$. That is, if $T_i \leq T_{0.5}$, then the survival time is T_i and the failure indicator is δ_i. If, on the other hand, $T_i > T_{0.5}$, then the survival time is $T_{0.5}$ and the failure indicator is 0. The set of data corresponding to the second epoch will consist of the full data set left-truncated at $T_{0.5}$. That is, it will only contain observations for those subjects where $T_i > T_{0.5}$; the failure indicators will be unchanged.

In large studies where many failures are observed, data sets can be created based on more than two epochs. For k epochs the cut-points will be $\{t_0, t_1, \ldots, t_k\}$, where $t_{j-1} < t_j$, $t_0 = 0$ and $t_k = \infty$. Left truncating at t_{j-1} and right censoring at t_j creates the data set for the j^{th} epoch.

A significance test of the proportional hazards assumption is achieved by comparing the log likelihood of the full model to those of the model applied to the k epochs. The statistic for the global test will be

$$-2 \times \log \text{lik}(\text{full model}) - \left\{ -2 \sum_{j=1}^{k} \log \text{lik}(j^{th} \text{ epoch model}) \right\}$$

Exercise Type	Number of Epochs	$-2\sum$ log lik	Difference	df	p-value
Bicycle	1	933.88	-	-	-
	2	933.86	0.01	2	0.995
	4	924.40	9.48	6	0.15
Treadmill	1	1455.03	-	-	-
	2	1454.71	0.32	2	0.85
	4	1451.44	3.59	6	0.73

Table 3.3 Results of fitting Cox models when the time axis is divided into distinct epochs

which, under the null hypothesis of proportional hazards, will follow a χ^2 distribution with q(k-1) degrees of freedom, where q is the number of regression parameters estimated in each model.

This method has the advantage that non-monotonic changes in hazard ratio can be detected. Any departures from the proportional hazards assumption can be evaluated by plotting the regression parameters for each covariate against time. The time value against which to plot effect estimates is arbitrary, though the median or mean failure time of those failing during each epoch would seem the most sensible options.

Example 3.4 Distinct Time Epochs in Cox PH Model for TIBET Time to Anginal Pain

Table 3.3 shows the values of -2(log likelihood) for the model fitted to all the data (1 epoch), as well as the sum of -2(log likelihood) for models fitted over partitioned time axes (2 or 4 epochs) and the change in -2(log likelihood) compared to the full model. The degrees of freedom (df) for the χ^2 distribution of this difference under the hypothesis of proportional hazards and the corresponding p value are also shown. These figures give no evidence against the proportional hazards assumption.

3.5.2.4 Score Process

The methods presented thus far can be used for any proportional hazards regression model. The following two methods are specifically designed for use with the Cox model. They are most readily described by adopting the notation of the Anderson-Gill generalisation of this model[89], which will be described briefly here.

Each subject is associated with a counting process, $N_i(t)$, which increases by 1 at each event time. The standard survival situation, where each subject may experience at most one event (for example, death), $N_i(0) = 0$ for all i; if subject i experiences the event at time t_i then $N_i(t) = 0$ for all $t < t_i$ and $N_i(t) = 1$ for $t \geq t_i$. $N_i(t)$ therefore "counts" the number of events experienced by subject i up to and including time t. Also associated with each subject is the risk indicator function, $Y_i(t)$, which indicates if subject i is at risk and under observation immediately prior to time t; thus if subject i has a survival time of t_i, at which point they either suffer an event or are censored, then $Y_i(t) = 1$ for all $t \leq t_i$, and $Y_i(t) = 0$ for all $t > t_i$. The proportional hazards regression model is then written as $\lambda(t|\mathbf{z}) = \lambda_0(t)Y_i(t)\exp(\mathbf{z}\beta)$.

Within the framework of the Anderson-Gill generalisation of the Cox proportional hazards model, the partial likelihood is written as

$$L_p(\beta|\mathbf{z}) = \prod_{i=1}^{n} \prod_{s\geq 0} \left\{ \frac{Y_i(s)\exp(\mathbf{z}_i\beta)}{\sum_{j=i}^{n} Y_j(s)\exp(\mathbf{z}_j\beta)} \right\}^{dN_i(s)}.$$

Notice that $dN_i(s) = 1$ only at the instant that subject i fails, if at all; otherwise $dN_i(s) = 0$. Consequently, contributions are made to the likelihood only at those instants where subjects are observed to fail, and the product over $s\geq 0$ need only be calculated at the unique failure times. The derivatives of the log partial likelihood with respect to β are

$$\frac{\partial}{\partial \beta_j} \log L_p = \sum_{i=1}^{n} \int_0^{\infty} \left\{ z_{ij} - \bar{z}_j(\beta, s) \right\} dN_i(s)$$

where

$$\bar{z}_j(\beta, s) = \frac{\sum_{i=1}^{n} Y_i(s) z_{ij} \exp(\mathbf{z}_i\beta)}{\sum_{i=1}^{n} Y_i(s)\exp(\mathbf{z}_i\beta)}.$$

The interpretation of $\bar{z}_j(\beta, t_{(k)})$ is as a weighted mean of the j^{th} covariate amongst those still at risk just before time $t_{(k)}$, with weights equal to the hazard for each individual.

The j^{th} score process, $S_j(t)$, is defined as

$$S_j(t) = \sum_{i=1}^{n} \int_0^{\infty} \left\{ z_{ij} - \bar{z}_j(\hat{\beta}, s) \right\} dN_i(s),$$

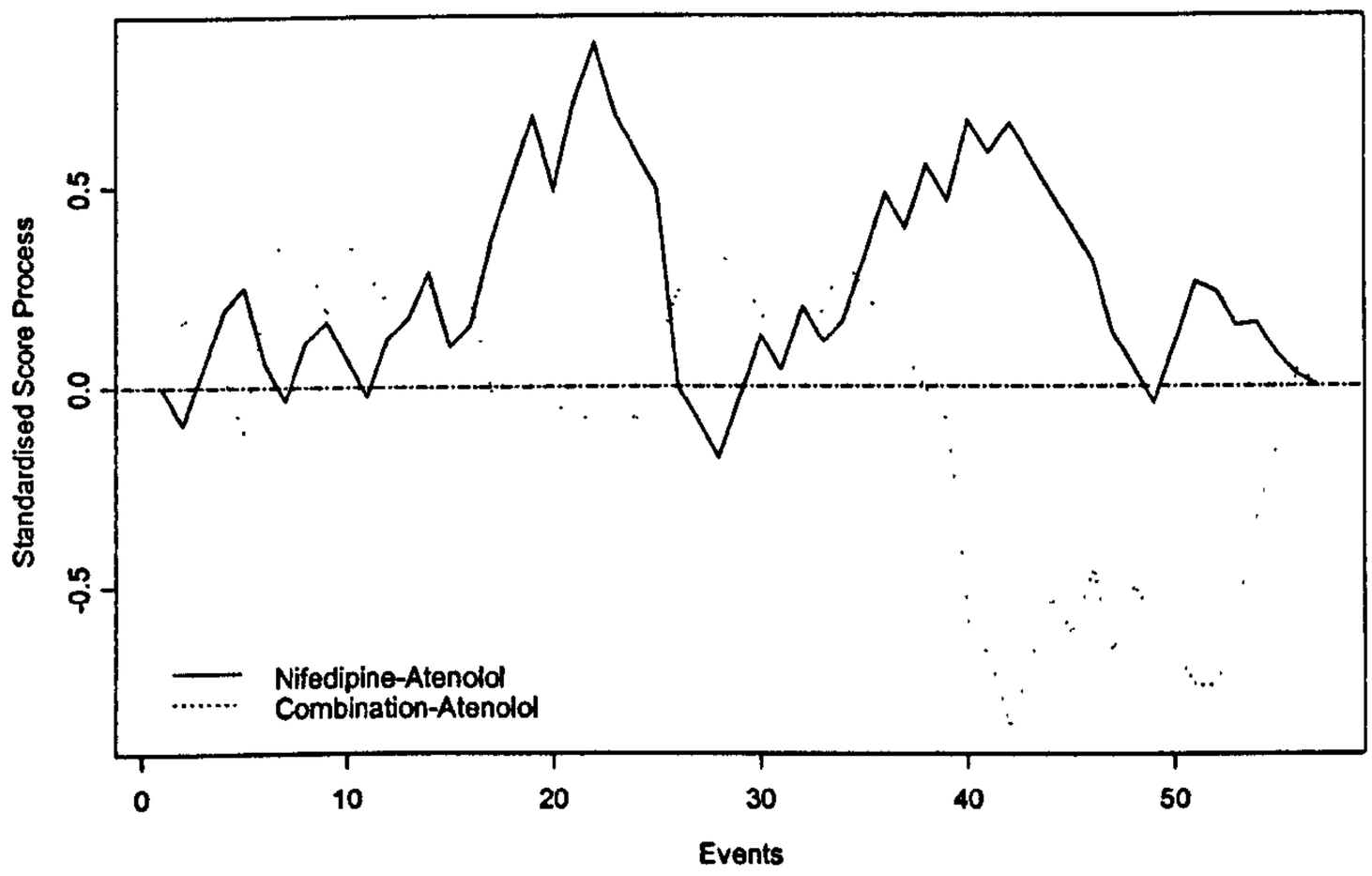
where $\hat{\beta}$ is the maximum partial likelihood estimate of β from the Cox proportional hazards model. The standardised score process, $S_j^*(t)$, is defined as

$$S_j^*(t) = V(\hat{\beta}_j)^{1/2} S_j(t),$$

where $V(\hat{\beta}_j)$ is the estimated variance of the j^{th} parameter estimate. $S_j(0) = S_j(\infty) = 0$, and it can be shown[90] that $S_j^*(t)$ converges asymptotically to a tied down Brownian motion process, $W_0(t)$. This result can be used to derive a test of the fit of the j^{th} covariate in the model; a possible statistic for such a test could be the maximum deviation of the standardised score process, $\sup_{0 \le t < \infty} |S_j^*(t)|$. Critical values for this statistic can be found by referring to the distribution of $\sup_{0 \le t < 1} |W_0(t)|$, which has been tabulated elsewhere[91], but it is not known how well the statistic fits this distribution with moderate sample sizes, heavy censoring or a large number of tied failure times.

The increments of the unstandardised score process, $\int_0^{\infty} \left\{ z_{ij} - \bar{z}_j(\hat{\beta}, s) \right\} dN_i(s)$, are the Schoenfeld residuals[92] of a fitted Cox proportional hazards regression model and readily accessible with many statistical computer packages. As an alternative to comparing $\sup_{0 \le t < \infty} |S_j^*(t)|$ to the distribution of $\sup_{0 \le t < 1} |W_0(t)|$, a randomised permutation test of the observed statistic could be used as a test of the proportional hazards assumption for the j^{th} covariate. This requires the calculation of the test statistic using a large number of random reorderings of the Schoenfeld residuals; the distribution of values obtained is taken as the null distribution to which the observed value is compared to obtain a p-value. Since deviations from the proportional hazards assumption are evident from large values of the statistic, the p-value will be one-sided. Whichever method is used, the test statistic should be sensitive to departures from the proportional hazards assumption where the effect of a covariate is increasing or decreasing over time, but not to situations where effects vary in a more complex manner.

(a)



(b)

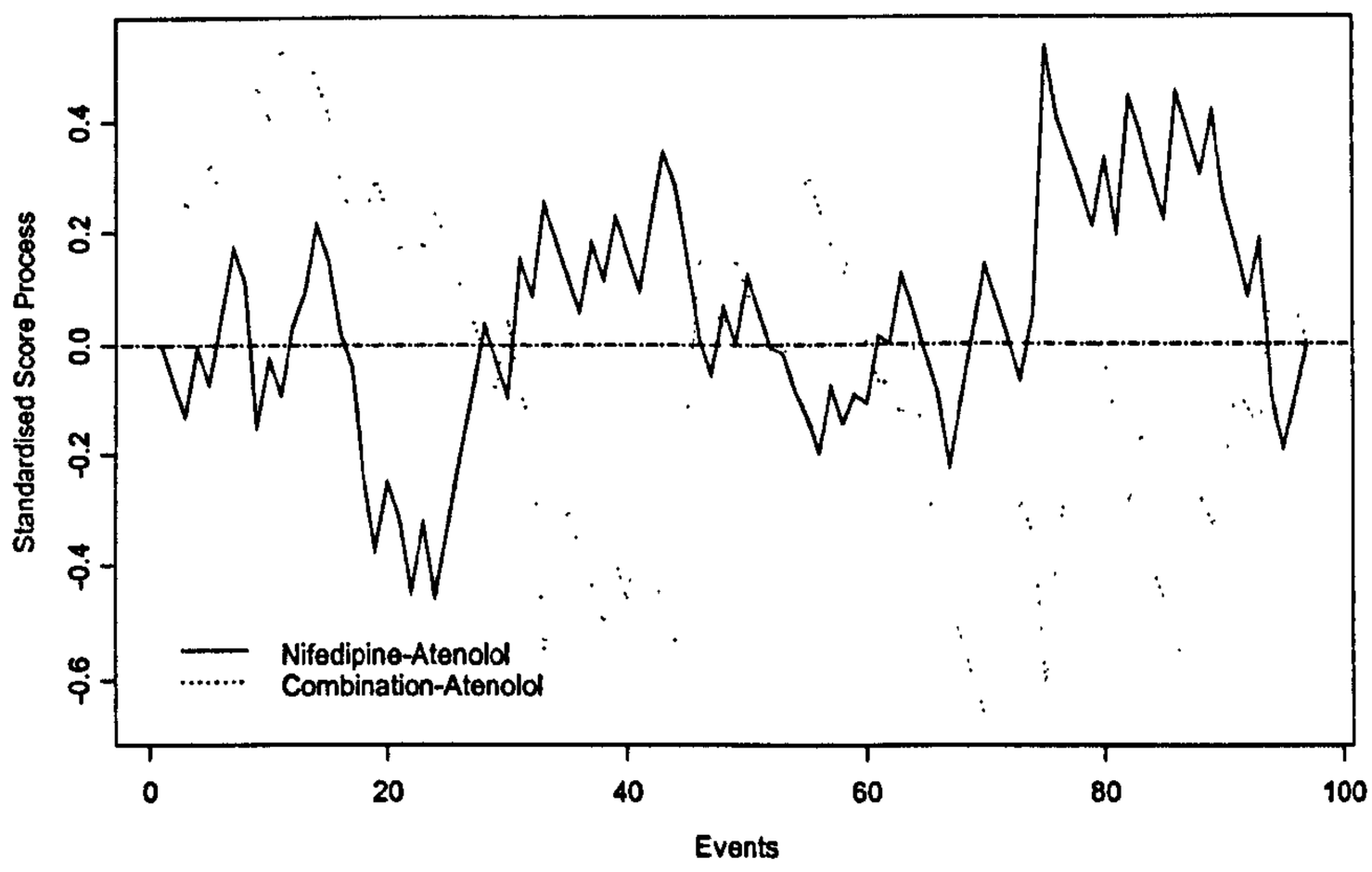


Figure 3.4 Standardised score processes for time to anginal pain during (a) bicycle and (b) treadmill exercise

Example 3.5 Score Process for Cox PH Model of TIBET Time to Anginal Pain

Figure 3.4 shows the standardised score processes for each treatment contrast, relative to Atenolol therapy, for the data from both exercise types. The critical value[91] of $\sup_{0\le t<\infty}\left|S_j^*(t)\right|$ at the 5% significance level is 1.4802, and it is clear that none of the processes shown approach this value. This is confirmed by the results given in Table 3.4, which gives the maximum observed deviation of the processes, the 5% critical values based on 100,000 random permutations of the Schoenfeld residuals, and an approximate p-value for the observed data based on the randomised permutation test. Again there is no evidence against the proportional hazards assumption.

Exercise Type	Treatment Contrast	Observed maximum	Threshold	p-value
Bicycle	Nifedipine-Atenolol	0.85	1.33	0.42
	Combination-Atenolol	0.86	1.43	0.51
Treadmill	Nifedipine-Atenolol	0.54	1.54	0.97
	Combination-Atenolol	0.67	1.61	0.89

Table 3.4 Observed maximum absolute values of score processes, with approximate thresholds and p-values from randomised permutation tests (sample size = 100,000)

3.5.2.5 Time Varying Coefficients

The following method for checking the proportional hazards assumption of the Cox model is based on considering a more general model and then testing if the new model would offer a significant improvement in fit. It provides a global test of the proportional hazards assumption, and for each covariate it gives a statistical test of the assumption as well as a graphical representation of how its effect exhibits non-proportionality.

Under the assumption of proportional hazards, the effect of a covariate is constant over time. The time varying coefficients (TVC) model[92,93] generalises the standard proportional hazards model by allowing the coefficients for each covariate to be a function of time. Using the Andersen-Gill notation, the TVC model can be written as

$$\lambda(t|\mathbf{z}_i) = \lambda_0(t) Y_i(t) \exp(\mathbf{z}_i \beta(t)),$$

so that the Cox model is expressed as the special case, $\beta(t) = \beta$. To test the proportional hazards assumption, $\beta(t)$ is written as $\beta + \mathbf{G}(t)\theta$, where $\mathbf{G}(t)$ is a diagonal p×p matrix (p being the number of covariates in the model), with elements $G_{jj}(t) = g_j(t)$ that vary about 0. $\mathbf{G}(t)$ is chosen in advance according to the type of non-proportional hazards that it is desired to detect. Under this formulation, a test of the proportional hazards assumption is a test of the hypothesis $\theta = 0$.

Let there be d observed events, occurring at times t_k, where $k = 1, 2, \ldots d$, and $0 \leq t_1 \leq t_2 \leq \ldots \leq t_d \leq \infty$. Under the proportional hazards model, the conditional weighted mean of the covariates amongst those still at risk at time s is

$$\bar{z}(\beta,s)=\frac{\sum_{i=1}^{n}Y_i(s)z_i\exp(z_i\beta)}{\sum_{i=1}^{n}Y_i(s)\exp(z_i\beta)}$$

and the conditional weighted variance-covariance matrix of the covariates is given by

$$V_Z(\beta,s)=\frac{\sum_{i=1}^{n}Y_i(s)z_iz_i'\exp(z_i\beta)}{\sum_{i=1}^{n}Y_i(s)\exp(z_i\beta)}-\bar{z}(\beta,s)\bar{z}(\beta,s)'.$$

If $\hat{\beta}$ is the maximum partial likelihood estimate of β under the proportional hazards model, the observed Schoenfeld residuals[92] under this model are defined to be

$$\begin{aligned}r_k(\hat{\beta})&=z_{(k)}-\bar{z}(\hat{\beta},t_k)\\&=\{z_{(k)}-\bar{z}(\hat{\beta}(t_k),t_k)\}+\{\bar{z}(\hat{\beta}(t_k),t_k)-\bar{z}(\hat{\beta},t_k)\}\\&=r_k(\hat{\beta}(t_k))+\{\bar{z}(\hat{\beta}(t_k),t_k)-\bar{z}(\hat{\beta},t_k)\}\end{aligned}$$

where $\bar{z}(\beta(s),s)$ is the conditional weighted mean of the covariates amongst those still at risk at time s under the TVC model. If the TVC model is correct, then the true Schoenfeld residuals under this model, $r_k(\beta(t_k))$, have zero mean. By expanding $\bar{z}(\beta(t_k),t_k)$ as a Taylor series about $\beta(t_k)=\hat{\beta}$ and ignoring 2nd and higher order terms,

$$\begin{aligned}\bar{z}(\hat{\beta}(t_k),t_k)-\bar{z}(\hat{\beta},t_k)&\approx\frac{\partial}{\partial\beta}\bar{z}(\beta,t_k)\Big|_{\beta=\hat{\beta}}(\beta(t_k)-\hat{\beta})\\&=V_Z(\hat{\beta},t_k)(\beta(t_k)-\hat{\beta})\end{aligned}$$

so that

$$\begin{aligned}E(V_Z^{-1}(\hat{\beta},t_k)r_k(\hat{\beta}))&=E(\beta(t_k)-\hat{\beta})\\&=\beta(t_k)-\beta\\&=G(t_k)\theta.\end{aligned}$$

That is, $V_Z^{-1}(\hat{\beta},t_k)r_k(\hat{\beta})$ has an expectation of zero under the proportional hazards assumption, but an expectation of $G(t_k)\theta$ under the TVC model.

If there are proportional hazards, a plot of $V_Z^{-1}(\hat{\beta}_j,t_k)r_k(\hat{\beta}_j)$ against $g_j(t_k)$ should show no linear association. If the specified TVC model is more appropriate it will have

slope θ_j. $\hat{\theta}$ can be estimated by ordinary least squares, but since the observed residuals $r_k(\hat{\beta})$ are correlated, it is more correct to use a weighted least squares estimate[93], namely

$$\hat{\theta} = \mathbf{D}^{-1} \sum_k \mathbf{G}(t_k) r_k(\hat{\beta})$$

where

$$\mathbf{D} = \sum_k \mathbf{G}(t_k) \mathbf{V}_Z(\hat{\beta}, t_k) \mathbf{G}(t_k) \left[\sum_k \mathbf{G}(t_k) \mathbf{V}_Z(\hat{\beta}, t_k) \right] \left[\sum_k \mathbf{V}_Z(\hat{\beta}, t_k) \right]^{-1} \left[\sum_k \mathbf{G}(t_k) \mathbf{V}_Z(\hat{\beta}, t_k) \right]',$$

and a global test of the proportional hazards assumption is given by the statistic

$$\left[\sum_k \mathbf{G}(t_k) r_k(\hat{\beta}) \right]' \mathbf{D}^{-1} \left[\sum_k \mathbf{G}(t_k) r_k(\hat{\beta}) \right]$$

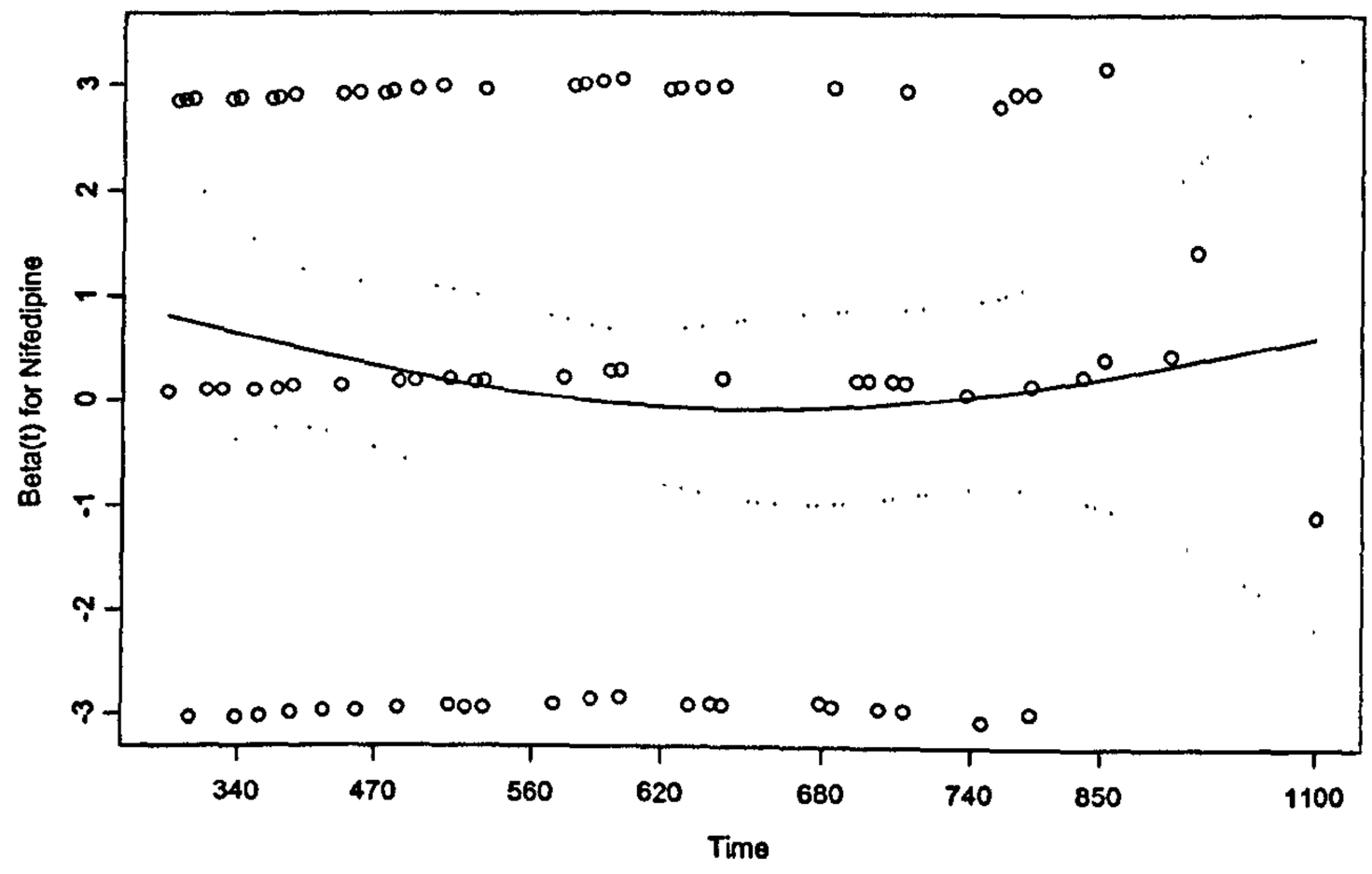
which is asymptotically χ^2_p distributed. In practice, $\mathbf{V}_Z(\hat{\beta}, t_k)$ can be substituted by the average value I/d, where $I \left(= \sum_k \mathbf{V}_Z(\hat{\beta}, t_k) \right)$ is minus the second derivative of the log partial likelihood from the Cox model.

The advantage of this method is its versatility, since it provides a formal test of deviations from the proportional hazards assumption, and a graphical representation of how the effect of each covariate changes over time. Different choices of **G**(t) lead to different alternative hypotheses to proportional hazards, and many global tests that have been proposed can be seen as special cases of this approach[93].

Example 3.6 Time Varying Coefficients in Cox PH Model for TIBET Time to Anginal Pain

Figure 3.5 shows plots of $\mathbf{V}_Z^{-1}(\hat{\beta}_j, t_k) r_k(\hat{\beta}_j) + \hat{\beta}_j$ versus t, where j subscripts the two treatment variables in the Cox model for the time to anginal pain during bicycle exercise. Each plot shows a smooth of the data with confidence bands at ±2 standard errors. This shows the functional form of $\beta_j(t)$, so that a horizontal line is consistent with the assumption of proportional hazards, whereas a trend in the graph would imply an increasing or decreasing treatment effect. These and similar figures for exercise using a treadmill give no indication of non-proportional hazards.

(a)



(b)

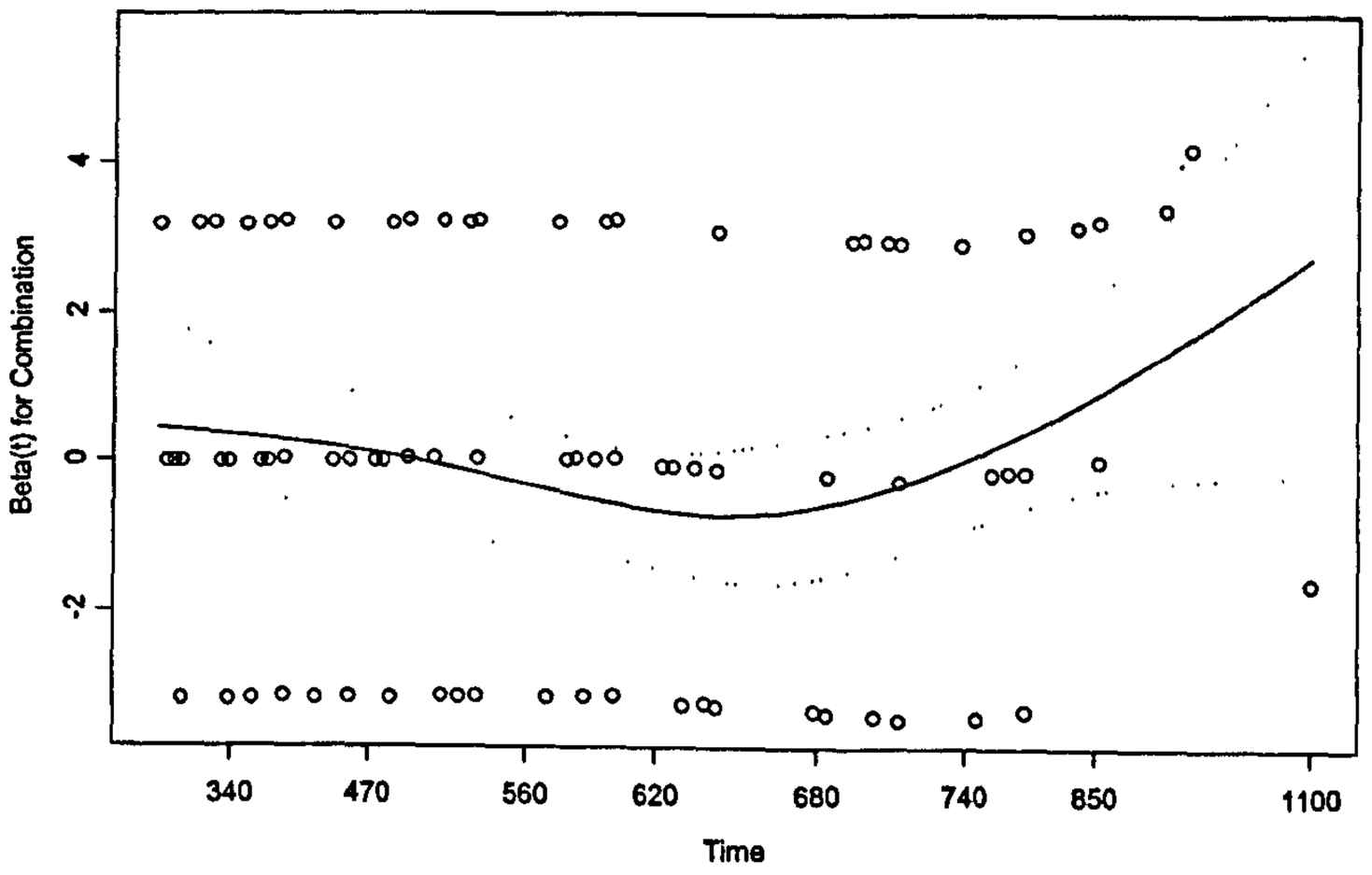


Figure 3.5 Time varying coefficients plots for (a) Nifedipine-Atenolol and (b) Combination-Atenolol treatment contrasts under Cox model for time to anginal pain during bicycle exercise.

Table 3.5 gives the results of the formal tests of proportional hazards. For each outcome, the test results for each treatment variable and for a global test of both variables are listed. None of these tests indicate non-proportional hazards.

Exercise Type	Contrast	χ^2	df	p-value
Bicycle	Nifedipine-Atenolol	0.46	1	0.50
	Combination-Atenolol	0.02	1	0.90
	Global	0.77	2	0.68
Treadmill	Nifedipine-Atenolol	0.003	1	0.95
	Combination-Atenolol	0.05	1	0.83
	Global	0.08	2	0.96

Table 3.5 Results of time varying coefficients test for non-proportional hazards in Cox models for treatment effects

CHAPTER 4 Estimation of Treatment Effect Differences II: Other Methods

This Chapter will outline a number of alternative methods that have been or could be used to analyse exercise test data from clinical trials. Some alternative methods of analysing censored survival data will be considered, followed by some standard regression methods, such as analysis of variance, that have been used with such data, despite their inherently censored nature. Throughout, methods will be illustrated using data from the TIBET Study.

4.1 Parametric Distributions of Survival Time

The Cox proportional hazards model described in Section 3.3 makes no assumptions regarding the functional form of the baseline hazard function, $\lambda_0(t)$. The partial likelihood method of estimating the regression coefficients effectively treats the baseline hazard as a nuisance. The popularity of this model lies in its flexibility; by allowing $\lambda_0(t)$ to take any shape, there is no need to assess the fit of the data in terms of assumptions about the baseline hazard function.

There are, however, a number of candidate functions for the distribution of survival data for which there is an extensive literature[94,95]. Some of these are outlined in Table 4.1, which shows their hazard functions, $\lambda(t)$ and the survivor functions, $S(t)$.

The simplest is the exponential distribution, in which the hazard is constant. It is often of limited applicability, since it has only one parameter and is therefore sensitive to moderate lack of fit, particularly in the tail. It is, however, the most important distribution in survival analysis, since many of the other distributions shown in Table 4.1 contain the exponential distribution as a special case.

The Weibull distribution is equivalent to the exponential when the shape parameter, γ is 1. When $\gamma<1$, the hazard at $t=0$ is infinite, and decreases with time; when $\gamma>1$, the hazard at $t=0$ is zero, and increases with time. The linear exponential and

Name	$\lambda(t)$	$S(t)$
Exponential	λ	$\exp(-\lambda t)$
Weibull	$\lambda\gamma(\lambda t)^{\gamma-1}$	$\exp\left[-(\lambda t)^{\gamma}\right]$
Linear exponential	$\lambda+\gamma t$	$\exp\left[-\left(\lambda t+\frac{1}{2}\gamma t^2\right)\right]$
Gompertz	$\exp(\lambda+\gamma t)$	$\exp\left[-\frac{\exp(\lambda)}{\gamma}\{\exp(\gamma t)-1\}\right]$
Gamma	$\frac{(\lambda t)^{\gamma-1}\exp(-\lambda t)}{\int_t^{\infty}(\lambda x)^{\gamma-1}\exp(-\lambda x)dx}$	$\frac{\lambda}{\Gamma(\gamma)}\int_t^{\infty}(\lambda x)^{\gamma-1}\exp(-\lambda x)dx$
Log logistic	$\frac{1}{1+(t\lambda)^{\gamma}}$	$\frac{\gamma t^{\gamma-1}\lambda^{\gamma}}{1+(t\lambda)^{\gamma}}$

Table 4.1 Some common distributions of survival times

Gompertz distributions both revert to the exponential when $\gamma=0$, and otherwise have monotonically increasing (if $\gamma>0$) or decreasing ($\gamma<0$) hazard functions. Regardless of γ, the hazard at $t=0$ is λ with the linear exponential distribution, or $\exp(\lambda)$ with the Gompertz distribution. The gamma distribution contains the exponential as a special case when $\gamma=1$. For $\gamma<1$ then hazard decreases monotonically from infinity and when $\gamma>1$ it increases monotonically from zero; in either case the hazard approaches λ as time progresses.

The log logistic distribution does not include the exponential as a special case. Indeed, when $\gamma>1$, the hazard function is not monotonic, having a single maximum value, though when $\gamma<1$, the hazard is monotonic decreasing.

Example 4.1 Parametric Model Selection with TIBET Time to Anginal Pain

When considering which survival distribution would be most suitable as a model for a particular set of data, graphical techniques based on the cumulative hazard function, $\Lambda(t)$, can be used. The Kaplan-Meier estimate of the survivor function, $\hat{S}(t)$, is used to estimate the cumulative hazard function, since $\hat{\Lambda}(t)=-\log\hat{S}(t)$. For example,

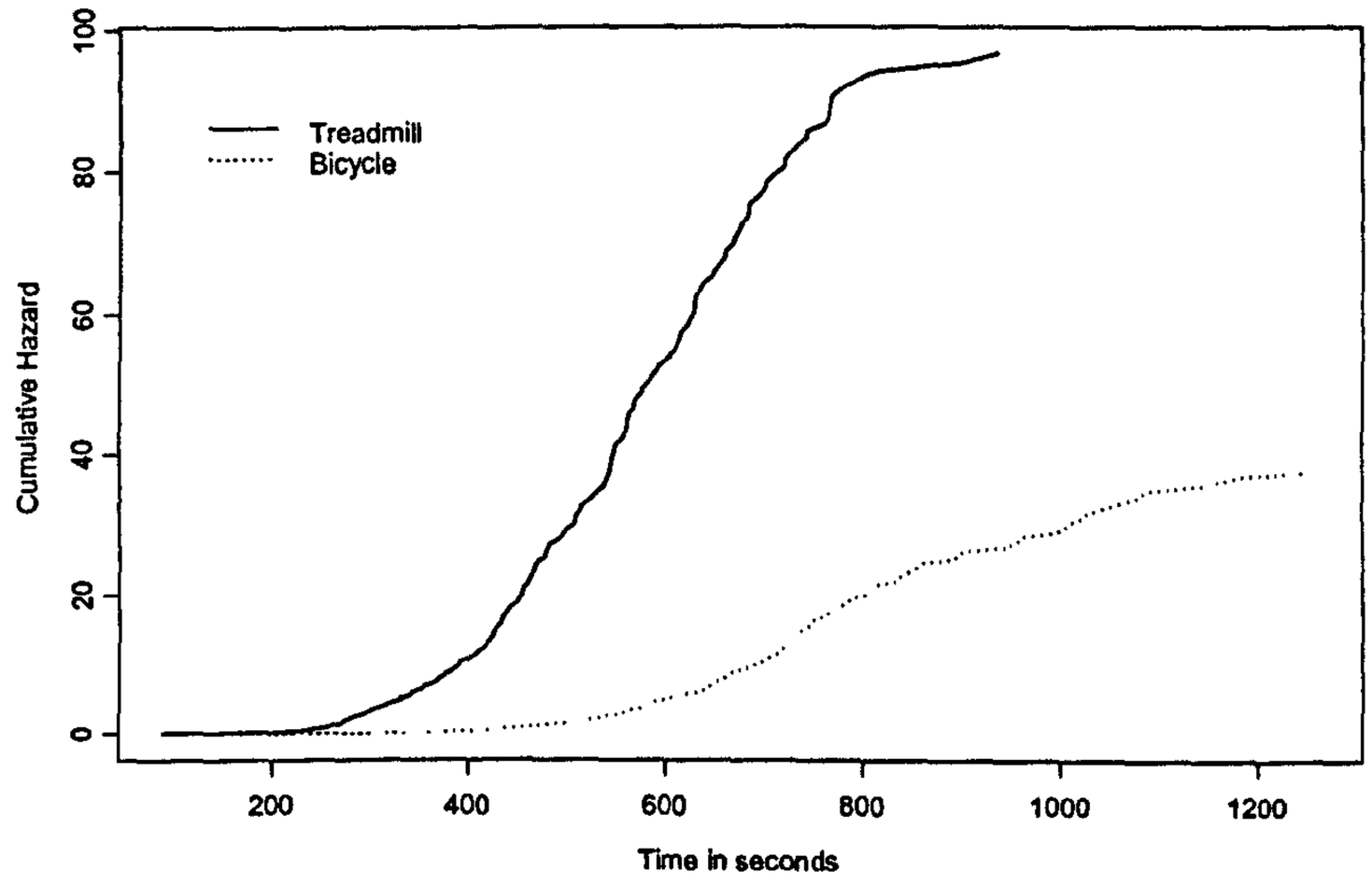


Figure 4.1 Cumulative hazard plots for the time to anginal pain, by exercise type

using the time to anginal pain data from the TIBET Study, Figure 4.1 shows $\hat{\Lambda}(t)$ plotted against t for those exercising on a treadmill and on a bicycle separately. The simplest survival distribution, the exponential distribution, with a constant hazard would have a cumulative hazard function that is linear in t. Similarly, if the data followed a linear exponential distribution, the cumulative hazard function would be quadratic in t. These distributions are clearly not supported by these data, though the fact that the two curves are of a similar shape would suggest at this stage that the time to anginal pain could follow similar distributions under the two types of exercise.

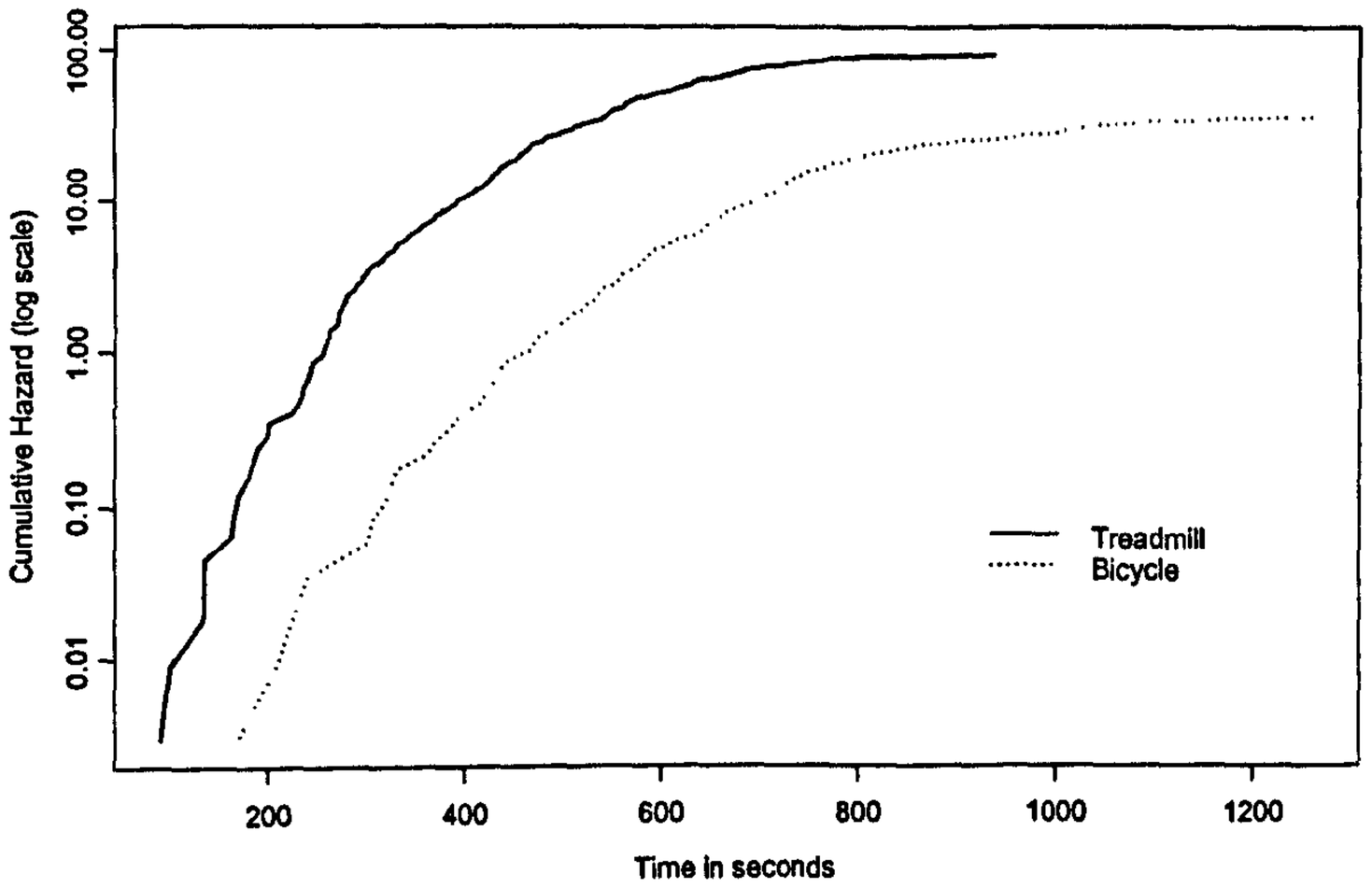


Figure 4.2 Cumulative hazard plots for the time to anginal pain with y-axis log-transformed, by exercise type

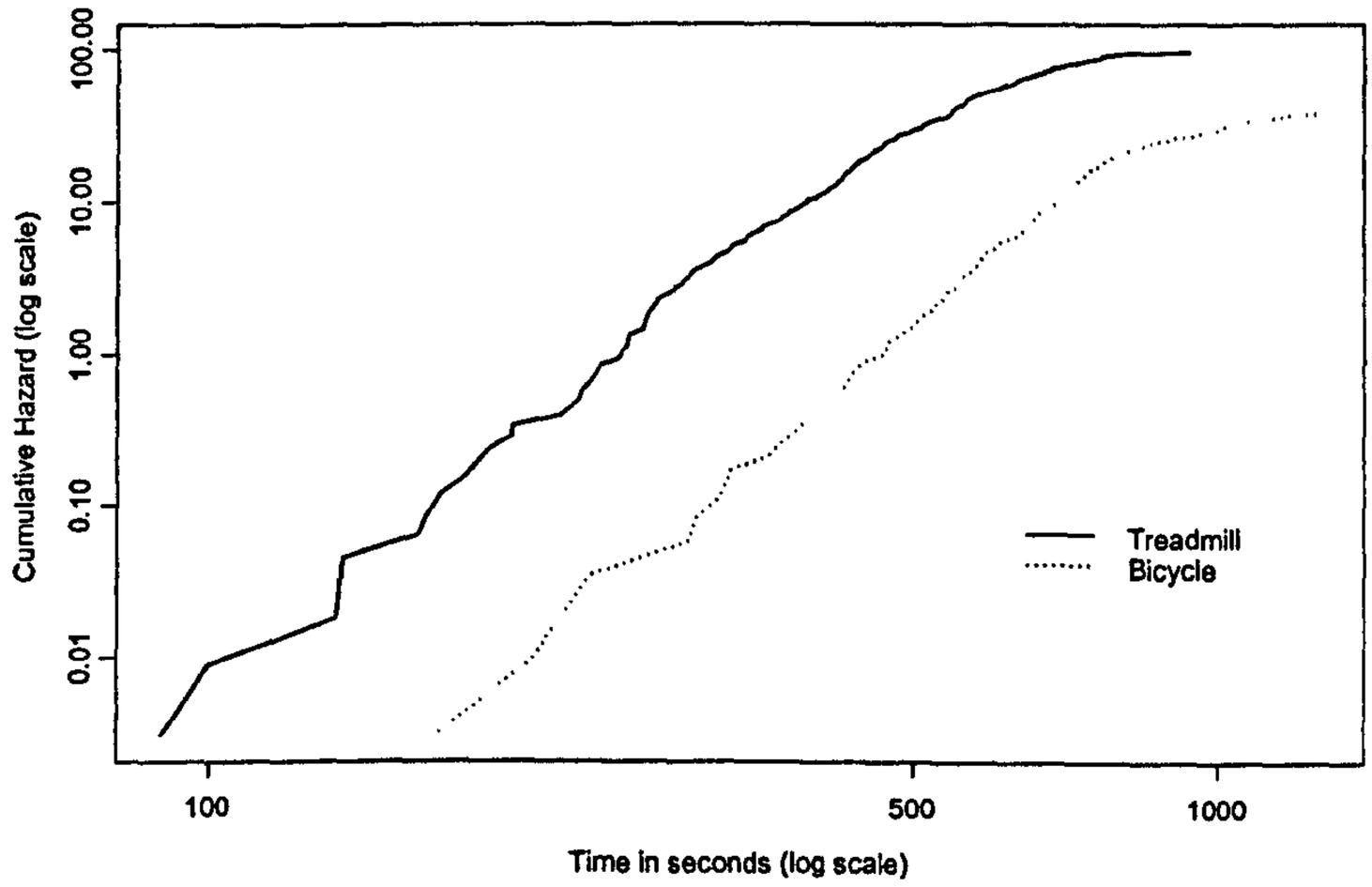


Figure 4.3 Cumulative hazard plots for the time to anginal pain with x- and y-axes log-transformed, by exercise type

By showing $\hat{\Lambda}(t)$ on a log scale, a linear relationship would suggest a Gompertz distribution, since

$$\log(\Lambda(t)) \propto \log(\exp\{\lambda + \gamma t\} - 1)$$
$$\approx \lambda + \gamma t$$

which is also clearly not the case, as shown by Figure 4.2. However, if $\log\hat{\Lambda}(t)$ were linear in log(t), this would suggest a Weibull distribution, since

$$\log(\Lambda(t)) = \gamma[\log(\lambda) + \log(t)].$$

Figure 4.3 shows this to be a possibility with these data. There appears to be a predominantly linear relationship over most of the time axis, though there is some evidence of non-linearity for large exercise times. The variance of the estimator of $\Lambda(t)$ is greatest towards the right of this graph, so this apparent non-linearity may be within the limits of chance. However, the patterns are similar for both types of exercise, which might be taken to indicate a true deviation from the Weibull distribution. For the purposes for further examples, it shall be assumed that these data are Weibull distributed, though the goodness-of-fit of any models should be examined for large values of t in particular.

4.2 Parameter Estimation

Whilst many standard statistical packages will fit parametric survival models for some of the distributions shown in Table 4.1, particularly the Weibull distribution (and therefore the exponential distribution), it is often not possible to compare the fits of all of the distributions shown. However, by the application of maximum likelihood, it is, in general, possible to fit any specified distribution to any set of survival data. If $\{t_i,\delta_i: i=1,2,...,n\}$ are an observed set of survival times and failure indicators, and $\lambda(t|\theta)$, $\Lambda(t|\theta)$ and $S(t|\theta)$ are the proposed hazard, cumulative hazard and survivor functions, dependent upon parameters θ, then the likelihood can be written as

$$L(\theta|\{t_i,\delta_i : i=1,2,...,n\}) = \prod_{i=1}^{n} S(t_i|\theta)\lambda(t_i|\theta)^{\delta_i}$$

and the log likelihood as

$$l(\theta|\{t_i,\delta_i : i=1,2,...,n\}) = \sum_{i=1}^{n} \{-\Lambda(t_i|\theta) + \delta_i \log\lambda(t_i|\theta)\}$$

so that as long as the hazard and cumulative hazard functions can be expressed algebraically, standard maximum likelihood methods can be used to estimate θ. Inferences can be drawn about and confidence intervals constructed for parameter estimates either by using likelihood ratio tests or by using the second derivative of minus the log likelihood to estimate the variance-covariance matrix for the estimates.

Example 4.2 Weibull Parameter Estimation for TIBET Time to Anginal Pain

Table 4.2 shows the maximum likelihood estimates obtained for the Weibull distribution parameters λ and γ, using all data as well as data from subgroups defined by exercise type and treatment group. To evaluate whether the times to anginal pain follow different distributions under the two types of exercise, the log likelihood of the model fitted to all data is compared to those obtained from the data separated by exercise type. Twice the difference in log likelihood is compared to a χ^2 distribution with 2 degrees of freedom (since an additional 2 parameters are estimated), yielding a highly significant ($p<0.0001$) result and providing very strong evidence that the distributions are not the same under treadmill and bicycle exercise.

Data		Estimates			Test		
		Scale λ	Shape γ	2×Log Lik		2×ΔLog Lik	p
All	A	0.064	2.03	-2061.3			
Treadmill	T	0.090	2.28	-1132.5	T=B	106.7	<0.0001
Bicycle	B	0.053	2.60	-822.2			
Treadmill (Atenolol)	T_A	0.088	2.37	-363.6			
Treadmill (Nifedipine)	T_N	0.090	2.21	-403.5	$T_A=T_N=T_C$	0.6	0.97
Treadmill (Combination)	T_C	0.092	2.27	-364.7			
Bicycle (Atenolol)	B_A	0.052	2.67	-266.7			
Bicycle (Nifedipine)	B_N	0.056	2.50	-294.0	$B_A=B_N=B_C$	1.5	0.82
Bicycle (Combination)	B_C	0.053	2.72	-260.0			

Table 4.2 Weibull parameter (λ, γ) estimates and values of 2×log likelihood for models applied to all data, by exercise types and by treatment groups, with likelihood ratio statistics and p-values

Similarly, within each exercise type, the Weibull distribution is fitted to the three treatment groups separately, and the likelihoods compared to test whether the survival distributions are different. No evidence of differences can be seen for either treadmill (p=0.97) or bicycle (p=0.82) exercise.

To determine in what way the survival distributions differ between exercise types, the likelihood ratio method can be used to derive joint 95% confidence intervals for the parameter estimates $\hat{\lambda}$ and $\hat{\gamma}$. These are shown in Figure 4.4, and whilst there appears to be little evidence that the shape (γ) of the survival distributions differ between exercise types, the scale (λ) of the distribution is larger under treadmill than bicycle exercise. This corresponds with a shorter mean time to anginal pain, since the mean of a Weibull distributed variable is $\frac{\Gamma\left(1+1/\gamma\right)}{\lambda}$. It is also of note that neither confidence interval is close to the value $\gamma=1$, supporting the earlier observation that the time to anginal pain is not exponetially distributed.

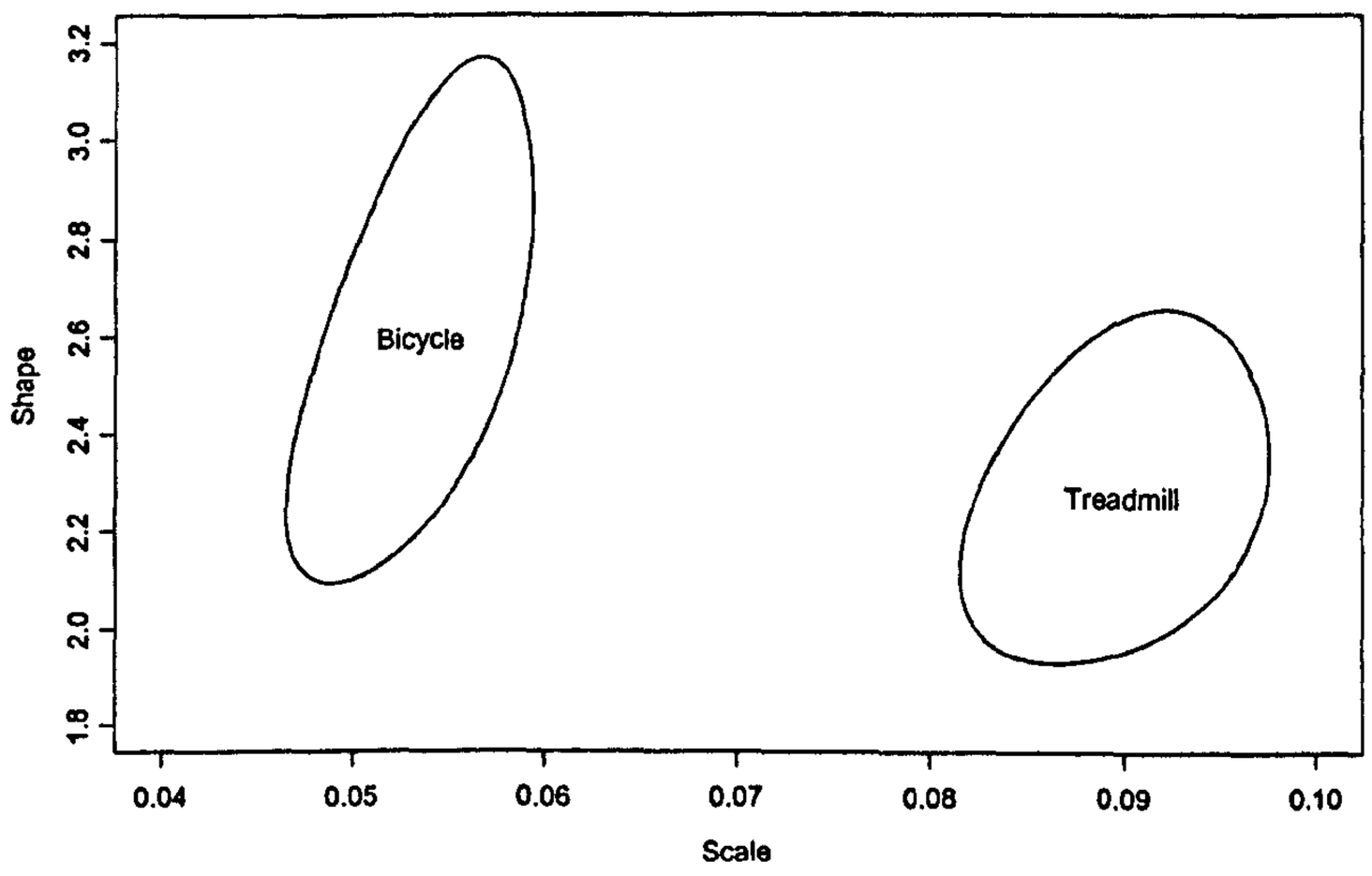


Figure 4.4 Joint 95% confidence intervals derived by the likelihood ratio method for estimates of the scale (λ) and shape (γ) parameters of Weibull survival distributions for the time to anginal pain from trial subjects exercising with treadmill or bicycle.

4.3 Regression Models for Survival Data

In order to estimate differences in treatment effects in clinical trials of antianginal therapies, possibly taking account of individual patient characteristics such as age, sex and weight, it is most efficient to develop regression models. The effects of explanatory variables, or covariates, are quantified by unknown parameters within the framework of a comprehensive model. Differences between groups are modelled by inclusion of 0/1 dummy variables and interaction effects can be examined by the inclusion of additional variables formed by combinations of covariates in the same way as with multiple linear regression models. This Section outlines some of the possible regression models that could be used.

4.3.1 Proportional Hazards Models

As described in Section 3.2, the proportional hazards regression model can be expressed as $\lambda(t|\theta,\beta,\mathbf{z})=\lambda_0(t,\theta)\psi(\mathbf{z},\beta)$. The hazard for an individual is the product of a baseline hazard function, $\lambda_0(t,\theta)$, and a function $\psi(\mathbf{z},\beta)$ of their covariates, constrained so that $\psi(\mathbf{0},\beta)=1$. The parameters θ encapsulate the underlying survival distribution of an individual with covariate vector $\mathbf{0}$, and the parameters β quantify the effects of the covariates $\mathbf{z}$. The most common function used to represent covariate effects is $\psi(\mathbf{z},\beta)=\exp(\mathbf{z}\beta)$. Covariates are commonly organised so that the vector $\mathbf{0}$ has some real

interpretation, for example an individual in the placebo or standard treatment arm of a parallel groups clinical trial, or continuous covariates have the mean or median value subtracted so that 0 corresponds to a central value.

Example 4.3 Weibull Proportional Hazards Regression for TIBET Time to Anginal Pain

Example 4.2 found that a Weibull distribution could adequately describe the baseline distribution of the time to anginal pain during both treadmill and bicycle exercise. A proportional hazards model could be used to estimate the effects of treatment on survival time, with or without adjustment for the effects of other covariates. Table 4.3 shows the maximum likelihood estimates obtained by fitting four models: the null model, with separate Weibull distributions for the times to anginal pain during treadmill and bicycle exercise; a model with two dummy variables to represent treatment differences; a model with covariates to represent gender differences and effects of age and body weight; and a model including both treatment and covariate effects.

For each model parameter, 95% confidence intervals have been constructed by the method of profile likelihood. If the model parameters, θ, are written as a single parameter, θ, and the vector of other parameters, ϕ_θ, then the function $l(\theta, \hat{\phi}_\theta)$ is the profile likelihood for θ, where $l(\,)$ is the log likelihood function and $\hat{\phi}_\theta$ is the maximum likelihood estimate of ϕ_θ. A 95% confidence interval for θ is defined as

$$\left\{\theta : 2\left[l(\hat{\theta}, \hat{\phi}_\theta) - l(\theta, \hat{\phi}_\theta)\right] \leq \chi^2_{1,0.95}\right\},$$

where $l(\hat{\theta}, \hat{\phi}_\theta) = l(\hat{\theta})$ is the global maximum log likelihood, and $\chi^2_{1,0.95}$ is the 95^{th} percentile point of a χ^2 distribution with 1 degree of freedom.

Models with treatment or covariate effects included are compared to models with those effects omitted by means of likelihood ratio tests; to ensure comparability all models have been fitted using only those subjects for whom there is complete data on all covariates. Treatment and covariate effect estimates are interpretable as log hazard ratios.

There is little evidence that treatments affect exercise times (p=0.18 without covariate adjustment, p=0.28 with adjustment), though the confidence interval for the Nifedipine-Atenolol treatment contrast only just includes zero (-0.01 to 0.27 without covariate adjustment, -0.03 to 0.24 with adjustment). Jointly, the three covariates

		Model			
		Null	Treatments	Age, Sex & Weight	Treatments, Age, Sex & Weight
Weibull parameters					
Treadmill	λ	0.085 (0.076, 0.093)	0.079 (0.068, 0.090)	0.086 (0.077, 0.094)	0.081 (0.069, 0.092)
	γ	2.31 (1.94, 2.72)	2.30 (1.93, 2.70)	2.36 (1.98, 2.78)	2.35 (1.97, 2.35)
Bicycle	λ	0.054 (0.048, 0.059)	0.051 (0.044, 0.057)	0.055 (0.049, 0.060)	0.052 (0.045, 0.059)
	γ	2.62 (2.20, 3.07)	2.63 (2.21, 3.09)	2.74 (2.30, 3.21)	2.74 (2.30, 3.22)
Treatment effects					
Nifedipine-Atenolol		-	0.31 (-0.03, 0.66)	-	0.28 (-0.07, 0.62)
Combination-Atenolol		-	0.11 (-0.25, 0.48)	-	0.10 (-0.26, 0.47)
Covariate effects					
Gender (Female-Male)		-	-	0.31 (-0.18, 0.77)	0.29 (-0.20, 0.74)
Age (/10 years, centred at 60 years)		-	-	0.22 (0.02, 0.42)	0.22 (0.02, 0.42)
Weight (/10 kg, centred at 75 kg)		-	-	-0.05 (-0.20, 0.09)	-0.04 (-0.19, 0.11)
Tests					
2×log likelihood		-1483.3	-1479.9	-1475.3	-1472.8
p (treatment effects)		-	0.18	-	0.28
p (covariate effects)		-	-	0.019	0.029

Table 4.3 Maximum likelihood estimates, with 95% confidence intervals calculated by profile likelihood, of Weibull parameters and treatment and covariate effects, with likelihood ratio test results, from proportional hazards models with baseline hazard stratified by exercise type

improve the fit of the model (p=0.019 without adjustment for treatment effects), though age is the only covariate to have an effect estimate significantly different to zero (estimated hazard ratio 1.24, 95% CI 1.02-1.52), so that older patients have increased hazard and therefore shorter average exercise times.

Interaction effects can be included in the model between exercise types and each treatment or covariate effect, in the same way as with standard linear regression models. Building on the model adjusting for treatment and covariate effects, there is evidence of an interaction between exercise type and weight (p=0.0072); the estimated hazard ratios associated with a 10 kg increase in weight are, under treadmill exercise, 1.17 (0.95,

1.44) and under bicycle exercise, 0.79 (0.64, 0.98). The other terms in the model are effectively unchanged compared with the fourth model shown in Table 4.3. Thus, with bicycle exercise, in which the apparatus supports the body, the greater strength associated with increased body weight allows longer exercise times and so is associated with lower hazard with respect to the occurrence of anginal pain. With treadmill exercise, in which body weight must be carried, greater weight will result in higher workload and is therefore associated with increased hazard, though this association does not reach statistical significance in these data.

4.3.2 Accelerated Failure Time Models

The accelerated failure time model can be represented as $S(t|\theta,\beta,\mathbf{z})=S_0(t\psi(\mathbf{z},\beta),\theta)$, so that $\lambda(t|\theta,\beta,\mathbf{z})=\lambda_0(t\psi(\mathbf{z},\beta),\theta)\psi(\mathbf{z},\beta)$. As with the proportional hazards model, $\psi(\mathbf{0},\beta)$ is constrained to be 1. If T is the random variable having survivor function $S(t|\theta,\beta,\mathbf{z})$, then the model can be viewed in terms of the relationship $T=T_0/\psi(\mathbf{z},\beta)$, where T_0 has survivor function $S_0(t,\theta)$. Written in this way, $\psi(\mathbf{z},\beta)$ represents the rate at which an individual "uses up" time, hence the name of the model.

If μ_0 is the expected value of $\log(T_0)$,

$$\log(T)=\mu_0(\theta)-\log\{\psi(\mathbf{z},\beta)\}+\varepsilon$$

where ε has zero mean and is independent of $\mathbf{z}$. A natural formulation of $\psi(\mathbf{z},\beta)$ is again $\psi(\mathbf{z},\beta)=\exp(\mathbf{z}\beta)$, so that

$$\log(T)=\mu_0(\theta)-\mathbf{z}\beta+\varepsilon$$

and $\lambda(t|\theta,\beta,\mathbf{z})=\lambda_0(t.\exp(\mathbf{z}\beta),\theta)\exp(\mathbf{z}\beta)$.

Example 4.4 Weibull Accelerated Failure Time Regression for TIBET Time to Anginal Pain

In the same way as a Weibull baseline hazard was extended in the previous example with a proportional hazards model, the same baseline hazard can be used in an accelerated life model. Table 4.4 shows the maximum likelihood estimates obtained from four models using the same covariates as Example 4.3; a null model with separate Weibull baseline hazard functions for those using treadmill or bicycle exercise (this is identical to the null model used in Example 4.3), a model adjusting for treatment effects, a model adjusting for covariate effects and a model adjusting for both treatments and covariates.

		Model			
		Null	Treatments	Age, Sex & Weight	Treatments, Age, Sex & Weight
Weibull parameters					
Treadmill	λ	0.085 (0.076, 0.093)	0.080 (0.069, 0.090)	0.086 (0.076, 0.094)	0.081 (0.071, 0.092)
	γ	2.31 (1.94, 2.72)	2.31 (1.94, 2.71)	2.29 (1.92, 2.71)	2.30 (1.92, 2.71)
Bicycle	λ	0.054 (0.048, 0.059)	0.051 (0.044, 0.057)	0.055 (0.050, 0.060)	0.053 (0.046, 0.059)
	γ	2.62 (2.20, 3.07)	2.63 (2.21, 3.08)	2.84 (2.37, 3.35)	2.82 (2.36, 3.33)
Treatment effects					
Nifedipine-Atenolol		-	0.12 (-0.01, 0.27)	-	0.10 (-0.03, 0.24)
Combination-Atenolol		-	0.04 (-0.10, 0.19)	-	0.04 (-0.10, 0.18)
Covariate effects					
Gender (Female-Male)		-	-	0.13 (-0.06, 0.31)	0.12 (-0.07, 0.30)
Age (/10 years, centred at 60 years)		-	-	0.09 (0.01, 0.16)	0.09 (0.01, 0.17)
Weight (/10 kg, centred at 75 kg)		-	-	-0.029 (-0.087, 0.029)	-0.023 (-0.082, 0.036)
Tests					
2×log likelihood		-1483.3	-1480.0	-1474.0	-1471.8
p (treatment effects)		-	0.19	-	0.32
p (covariate effects)		-	-	0.010	0.017

Table 4.4 Maximum likelihood estimates, with 95% confidence intervals calculated by profile likelihood, of Weibull parameters and treatment and covariate effects, with likelihood ratio test results, from accelerated failure time models for the time to anginal pain with baseline hazard stratified by exercise type

The conclusions reached by this model are very similar to those reached using a proportional hazards model. Overall, there is no evidence of treatment effects, though the Nifedipine-Atenolol treatment effect contrast approaches significance. Inclusion of other covariates does improve the fit of the model, primarily through the effect of age. As with the proportional hazards model, there is evidence of an interaction between exercise type and body weight (p=0.010). The effect estimates for a 10 kg increase in weight, adding an interaction between exercise type and weight to the last model shown in Table 4.4 are for treadmill exercise, 0.06 (-0.02, 0.16) and for bicycle exercise, -0.08

(-0.15, 0.01). Neither effect is significantly different to null, despite the strong evidence that the two effects differ from each other.

4.3.3 Other Regression Models

There are some less common alternatives that can be used to model survival data. Their lack of popularity amongst the medical literature could be due to a number of factors. There is a vast literature on the use of proportional hazards and accelerated failure time models, which are implemented within a number of statistical software packages, with a small selection of baseline hazard functions, at least. The most widely used regression method, that of the Cox proportional hazards model, is highly flexible with respect to the baseline hazard function, can be easily applied with standard software, and effect estimates are simple to report and are widely understood in terms of hazard ratios. The alternative methods shown here are more difficult to implement and would be less readily accepted unless a considerable benefit over more commonly used methods could be demonstrated.

Under an additive hazards model, hazard functions are parallel between groups, rather than proportional. This can be expressed as

$$\lambda(t|\theta,\beta,\mathbf{z})=\lambda_0(t,\theta)+\psi(\mathbf{z},\beta), \text{ with } \psi(\mathbf{z},\beta)\geq 0.$$

The transferred origin model incorporates covariate effects through a translation in time, so that

$$\lambda(t|\theta,\beta,\mathbf{z})=\lambda_0(t+\psi(\mathbf{z},\beta),\theta)$$

and, if $\lambda_0(x)$ is defined to be zero for negative x, then if $\psi(\mathbf{z},\beta)<0$, $-\psi(\mathbf{z},\beta)$ can be thought of as a hazard free interval, during which time an individual with covariates $\mathbf{z}$ is immune from failure. Otherwise, if $\psi(\mathbf{z},\beta)>0$, the individual can be considered to be at a more advanced stage relative to one with a covariate vector of $\mathbf{0}$.

In principle, different types could be combined within the same model, so that, for example, a model could consist of proportional hazards and a transferred origin parts,

$$\lambda(t|\theta,\beta,\gamma,\mathbf{z})=\lambda_0(t+\psi(\mathbf{z},\beta),\theta)\phi(\mathbf{z},\gamma)$$

where some components of β and γ could be zero, corresponding to whether a particular element of $\mathbf{z}$ acts only through the proportional hazards or transferred origin parts of the model.

4.4 Non-parametric Survival Methods

The log rank test[96] is used to compare survival distributions between subgroups of a population and can be reduced to a comparison of the observed and expected numbers of events using a χ^2 test[94]. If r_j is the size of the risk set at the j^{th} failure time, with r_{jk} being the size of the risk set amongst the k^{th} subgroup (k=1,2,...,K), and d_j the number of events occurring at the j^{th} failure time, then the expected number of events in the kth subgroup is

$$E_k = \sum_j d_j \frac{r_{jk}}{r_j}.$$

The test statistic is defined as

$$X^2 = \sum_k \frac{(D_k - E_k)^2}{E_k}.$$

where D_k is the number of events observed in the k^{th} subgroup. If the survival distributions of the K subgroups are the same, then $X^2 \sim \chi^2_{k-1}$, so that large values of X^2 provide evidence of differences between the K survival distributions.

Example 4.5 Log Rank Test for Treatment Effects on Time to Anginal Pain, TIBET Study

Table 5.2 shows the results of log rank tests applied to the time to anginal pain data from the TIBET Study. The test is applied to all data, to data from treadmill and bicycle tests separately, and to all data, stratified by exercise type. There is no evidence from any of these tests of any differences in survival between treatment groups.

4.5 Standard Regression Methods

The methods considered thus far are designed for use with censored data. These methods are preferable for use with exercise test data, because endpoints such as the occurrence of anginal pain, or significant ST-segment depression need not necessarily occur before the patient has to stop exercising for some other reason, such as fatigue or breathlessness.

However, the subjects recruited into a clinical trial of antianginal therapies may be selected on the basis of having poor exercise tolerance, and even under an improved treatment regime, would most likely experience an ischaemic event during exercise. As

Treatment			All	Treadmill	Bicycle	Stratified by Exercise Type
Atenolol	Subjects		218	114	104	218
	Events	Obs	85	53	32	85
		Exp	90.4	56.3	34.1	90.3
Nifedipine	Subjects		219	111	108	219
	Events	Obs	96	60	36	96
		Exp	88.3	58.2	30.9	89.2
Combination	Subjects		211	107	104	211
	Events	Obs	85	54	31	85
		Exp	87.3	52.5	34	86.5
χ^2 (2 df)			1.1	0.3	1.2	0.9
p			0.59	0.86	0.54	0.64

Table 4.5 Log rank test results for time to anginal pain for all subjects, separately by exercise type and stratified by exercise type; numbers of subjects, numbers of events (Obs), expected numbers of events (Exp), with associated χ^2 statistics and p-values

a result, the proportion of exercise tests producing censored observations might be small, as would any loss of efficiency from applying standard regression techniques instead of survival analysis methods. Also, it is likely that withdrawals due to non-ischaemic events would occur after a long period of exercise, so that times to ischaemic and non-ischaemic events would be positively correlated. Furthermore, the outcome of total exercise time is never censored, unless one considers exercising to the end of the exercise protocol, or withdrawal due to particular causes, such as leg pain or cramp, to determine an unobserved total exercise time. If biases or loss of efficiency due to ignoring censoring are of concern, then analyses can be repeated with censored observations included or excluded, to compare the results and assess the impact of censoring.

Nevertheless, the use of standard methods that do not take account of censoring where it occurs, is conceptually incorrect and is theoretically and empirically biased[52]. For the analysis of clinical studies involving human subjects, the most efficient methods should be used[97], so the analysis of censored data using these methods could not be advocated.

Effect Estimate (s) (95% CI) p-value	All Data	Censored Observations Excluded
Intercept	490.5 (459.3, 521.8) <0.0001	388.9 (343.8, 434.0) <0.0001
Exercise (Bicycle – Treadmill)	198.2 (168.4, 228.0) <0.0001	185.2 (142.5, 227.9) <0.0001
Treatment (Nifedipine – Atenolol)	-0.5 (-34.9, 33.9) 0.98	3.6 (-45.7, 52.9) 0.89
Treatment (Combination – Atenolol)	-7.7 (-42.1, 26.7) 0.66	10.1 (-41.9, 62.0) 0.70
Gender (Female – Male)	-116.9 (-159.7, -74.1) <0.0001	-101.0 (-170.7, -31.4) 0.0050
Age (/10 years)	-55.6 (-74.6, -36.7) <0.0001	-39.6 (-64.9, -14.4) 0.0024
Weight (/10 kg : Treadmill)	-20.9 (-42.5, 0.7) 0.058	-10.4 (-39.4, 18.6) 0.48
Weight (/10 kg : Bicycle)	58.9 (41.6, 76.2) <0.0001	41.5 (14.7, 68.4) 0.0028

Table 4.6 Linear regression model effects estimates with 95% confidence intervals and p-values, for models of time to anginal pain or end of exercise, using all data or restricted to subjects experiencing anginal pain

4.5.2 Exercise Times

When analysing exercise times using classical statistical techniques, the natural methods to use would be based on Normal theory, such as t-tests, ANOVA or, more generally, multiple linear regression. These methods are robust to mild departures from the assumption of Normality, though the data may require transformation to ensure that model residuals are Normally distributed. Otherwise, non-parametric methods, such as the Wilcoxon Mann-Whitney test could be used.

Example 4.6 Multiple Linear Regression Analysis of Time to Pain, TIBET Study

Table 4.6 shows the effect estimates obtained by fitting a linear regression model to the time to anginal pain data from the TIBET Study, with those subjects who did not experience pain during the test having their total exercise time substituted. Also shown are the estimates from the same model applied only to those subjects that experienced pain. The model shown includes effects of exercise type, treatment, gender, age and, for

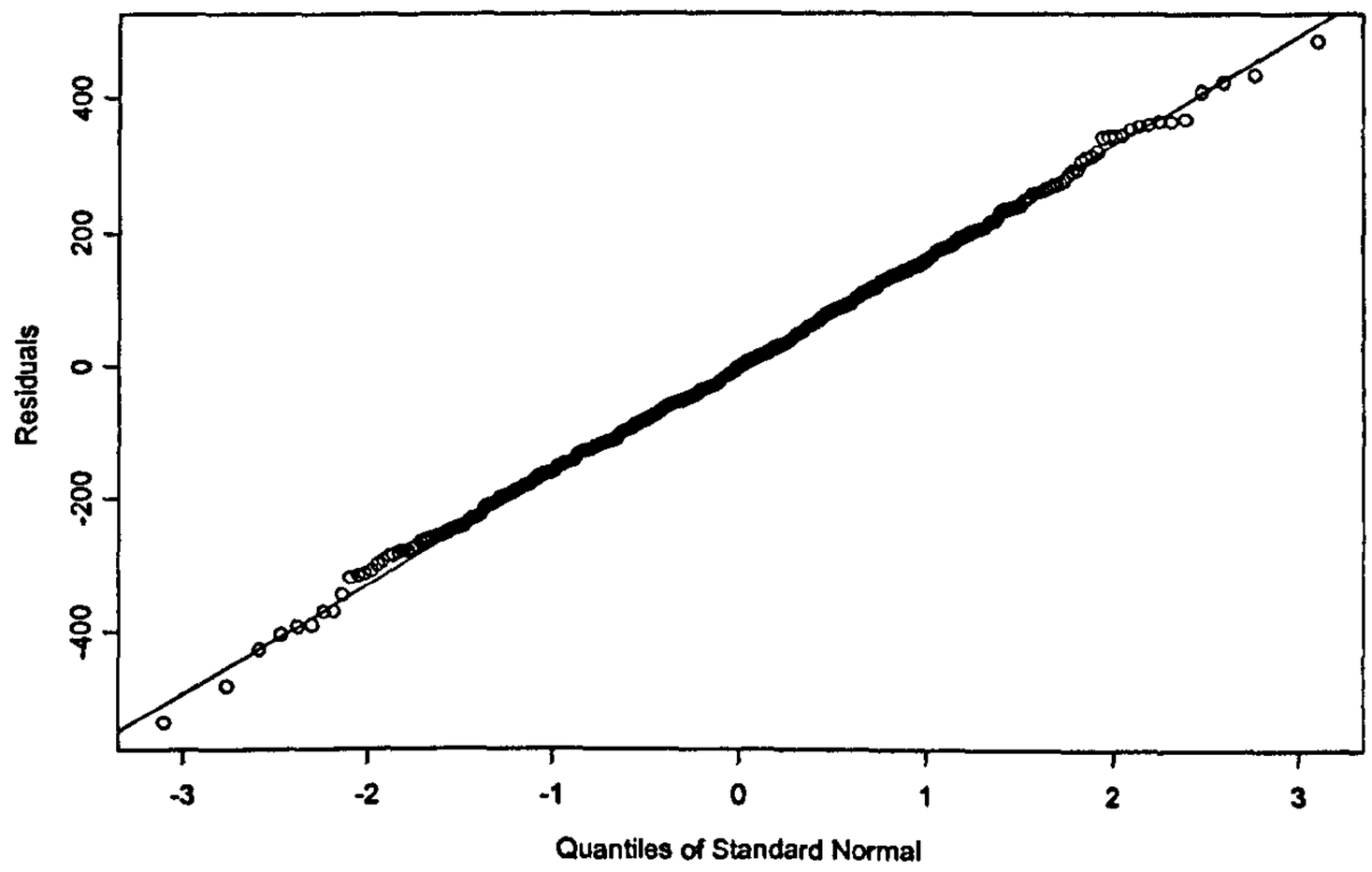


Figure 4.5 Normal probability plot of residuals from linear regression model of time to anginal pain or end of exercise, as shown in Table 4.6

each exercise type separately, weight. This was consistently the best model found amongst those that included treatment effects, regardless of whether or not censored individuals were included in the analysis.

There is no evidence that exercise times are any different under the three treatments. Exercise times are more than 3 minutes longer on a bicycle than on a treadmill, and older patients have shorter exercise times. The weight effects are different under the two exercise types, with heavier patients exercising for longer with a bicycle, and, if there is any real effect, for a shorter period on a treadmill.

Whereas gender was not found to have any effect on exercise times when using survival models, it has a large effect in this analysis. Using all data, women are seen to exercise for nearly two minutes less than men, an effect that persists when those who do not experience anginal pain are excluded.

Perhaps surprisingly, it was not necessary to transform exercise times to obtain an adequate model fit in terms of the assumption of Normally distributed errors. Figure 4.5 shows the Normal probability plot of the residuals from the first model shown in Table 4.6. The graph is similar when the model is restricted to those who experienced anginal pain during exercise.

4.5.3 Endpoints

The analysis could be focused on the occurrence of ischaemic endpoints, such as anginal pain or significant ST-segment depression. The proportions of individuals

		Treatment Group				
		Atenolol	Nifedipine	Combination	χ^2	p-value
Anginal Pain	All	87/221 (39.4%)	96/219 (43.8%)	88/214 (41.1%)	0.9	0.63
	Treadmill	54/116 (46.6%)	60/111 (54.1%)	56/109 (51.4%)	1.3	0.52
	Bicycle	33/105 (31.4%)	36/108 (33.3%)	32/105 (30.5%)	0.2	0.90
1mm ST-Segment Depression	All	157/226 (69.5%)	135/232 (58.2%)	126/224 (56.2%)	9.7	0.0078
	Treadmill	90/119 (75.6%)	74/119 (62.2%)	71/116 (61.2%)	6.9	0.032
	Bicycle	67/107 (62.6%)	61/113 (54.0%)	55/108 (50.9%)	3.2	0.20

Table 4.7 Numbers and percentages of subjects experiencing anginal pain and ST-segment depression during exercise, for all subjects and by exercise type

suffering an event in different treatment groups could then be compared using χ^2-tests or, to take account of covariate information, logistic regression.

Example 4.7 TIBET Ischaemic Endpoints

Table 4.7 shows the numbers of individuals suffering the ischaemic endpoints of anginal pain and 1mm ST-segment depression during exercise, with χ^2 tests for association. There is no evidence of any treatment differences in the proportions suffering anginal pain, though there is evidence that more of those using Atenolol suffer 1mm ST-depression during exercise (p=0.0078). The data as shown also suggest this difference to be apparent during treadmill rather than during bicycle exercise, though a formal test of this difference, using logistic regression, is required.

Table 4.8 shows the results of applying logistic regression models to the occurrence of anginal pain and 1mm ST-segement depression during exercise. All models include exercise type and treatment, though models are shown with and without adjustment for gender, age and body weight. Though the Nifedipine-Atenolol treatment contrast appears to suggest some effect upon the occurrence of anginal pain, as a whole there is no evidence of any treatment effect (joint treatment effect; p=0.12 without covariates, p=0.088 with covariates). However, there is evidence that treatments affect the occurrence of 1mm ST-segment depression during exercise, with (p=0.023) or without (p=0.020) adjustment for covariates. Under the adjusted model, both Nifedipine (odds ratio [OR]=0.64 [95% CI: 0.41, 0.99], p=0.045) and combination therapy

Odds Ratio (95% CI) p-value	Endpoint			
	Anginal Pain		1mm ST-Segment Depression	
	Without Covariates	With Covariates	Without Covariates	With Covariates
Exercise Type (Bicycle / Treadmill)	0.56 (0.39, 0.80) 0.0017	0.57 (0.39, 0.82) 0.0028	0.78 (0.55, 1.10) 0.16	0.73 (0.51, 1.05) 0.090
Treatment (Nifedipine / Atenolol)	1.55 (1.00, 2.39) 0.052	1.60 (1.03, 2.50) 0.038	0.63 (0.41, 0.96) 0.031	0.64 (0.41, 0.99) 0.045
Treatment (Combination / Atenolol)	1.10 (0.70, 1.72) 0.69	1.09 (0.70, 1.71) 0.70	0.56 (0.37, 0.86) 0.0087	0.56 (0.36, 0.86) 0.0089
Gender (Female / Male)	0.67 (0.38, 1.20) 0.18		1.06 (0.62, 1.83) 0.83	
Age (/10 years)	0.96 (0.75, 1.22) 0.73		1.49 (1.18, 1.89) 0.0010	
Weight (/10 kg)	1.04 (0.86, 1.24) 0.71		1.14 (0.96, 1.36) 0.15	

Table 4.8 Logistic regression model effect estimates (as odds ratios), with 95% CIs and p-values, for models of occurrence of anginal pain or 1mm ST-segment depression during exercise

(OR=0.56 [0.36, 0.86], p=0.0089) reduce the likelihood of suffering 1mm ST-segment depression during exercise.

For the endpoint of anginal pain, there is no evidence that any of the covariates considered influences the occurrence of the endpoint. There is evidence that advanced age increases the occurrence of 1mm ST-segment depression during exercise (OR=1.49 /10 years [1.18, 1.89], p=0.0010). With regard to the type of exercise performed, those using a bicycle appear less likely to suffer anginal pain (adjusted model, OR=0.57 [0.39, 0.82], p=0.0028), and though the trend is in the same direction, does not reach statistical significance for the effect estimate for the occurrence of 1mm ST-segment depression (adjusted model, OR=0.73 [0.51, 1.05], p=0.090).

Since no account has been taken of the time spent exercising in this analysis, the results are difficult to interpret. For example, older patients are no more likely to experience anginal pain than younger patients according to this analysis, though previous analyses suggest that older patients experience pain after a shorter period of exercise. Consequently, an analysis based on the occurrence of ischaemic events after a fixed period of exercise could be used to avoid this source of confounding.

CHAPTER 5 Interval Censoring

5.1 Introduction

Myocardial ischaemia during an exercise test can be determined by the occurrence of anginal pain by the subject or through the use of an electrocardiogram (ECG). Figure 5.1 shows a normal ECG trace, and the trace of a patient who is suffering from ischaemia. The ST-segment is the horizontal section of the trace to immediately to the right of the "spike"; when ischaemia occurs, the ST-segment of the ECG trace tends to move downwards, as in Figure 5.1(b). By monitoring the ECG of a patient during exercise, it is possible to detect the incidence of myocardial ischaemia in the absence of anginal pain. The ECG can also inform of adverse events such as severe ischaemia or dysrhythmia that indicate that the test should cease for the safety of the patient.

The occurrence of ≥1mm ST-segment depression is often used to indicate myocardial ischaemia, though the cut-off value is to some degree arbitrary. Larger values will be more specific, but less sensitive for detecting ischaemia. Other values that have been chosen vary from 0.5 to 2mm. In terms of patient safety, reasons to stop an exercise test would include rapidly increasing ST-segment depression or a severe level such as ≥5mm. Electronic monitoring of the ECG allows for automatic detection of dangerous levels of ST-segment depression and direct recording of results for later analysis.

In principle, it is possible to measure ST-segment depression at every heartbeat of the patient during a test. However, this would result in large amounts of data being collected on each patient, and for studies with hundreds of participants, with more than one exercise test, each lasting for several minutes, the quntity of data generated would be extremely large. Consequently, levels of ST-segment depression are usually measured at regular intervals, often every minute, with each reading being an average taken over several heartbeats. The capacity of an individual to endure an increased workload to the heart is measured by the time for which he/she can exercise against a

Figure 5.1 Examples of ECG traces from (a) a normal subject and (b) a subject with ST-segment depression

standard exercise protocol before myocardial ischaemia occurs. This can be measured by the time until the first occurrence of ≥1mm ST-segment depression.

5.1.2 Interval Censoring

The time until ≥1mm ST-segment depression is "interval censored"; that is, given the time until the first *observed* occurrence of ≥1mm ST-segment depression, it is known only that the first *actual* occurrence of the event took place during the interval since the previous recording of ST-segment depression. (In fact, it is possible that the ST-segment could have become depressed by more than 1mm during a previous interval but returned to a level of depression below 1mm before the end of that interval, when the next reading was taken. For the sake of simplicity, however, this possibility will be ignored, since it would be assumed that ischaemia would worsen during exercise.)

This chapter will outline some methods of analysing interval censored survival data, with examples of the methods being applied to data from the TIBET study on the time until ≥1mm ST-segment depression.

5.2 Standard Methods

The time until the first observed occurrence of ≥1mm ST-segment depression is often analysed as though it is not interval censored. That is, the time for which a patient exercises until ≥1mm ST-segment is observed is assumed to be a continuous random variable, even though it is not. Also, it is not uncommon for the fact that exercise times are survival times (which may be censored) to be ignored. Exercise times are often analysed by standard statistical techniques, such as t-tests or linear regression. An allowance for observations that are censored, where ≥1mm ST-segment depression was not seen to occur, might be to re-analyse the data with these subjects omitted, and compare the results to those obtained using all observations (Example 4.6).

Since the publication of simulation studies comparing these simple approaches to survival methods for the analysis of exercise times[52], the latter have been used more often. Generally, groups of exercise times are compared with log rank tests and covariate information is adjusted for by use of Cox proportional hazards or other survival regression methods.

Example 5.1 TIBET Study, Time to ≥1mm ST-Segment Depression

Table 5.1 shows the results of linear regression modelling of the times until ≥1mm ST-segment depression from the TIBET Study. Models were fitted to data from treadmill and bicycle exercise tests combined, with the type of exercise treated as a covariate. Interaction terms between exercise type and all other terms in the model were fitted, and included in the final model if this resulted in an improved model fit. Models were fitted using all data, in which those subjects that did not experience ≥1mm ST-segment depression during the exercise test had their total exercise time substituted, and to a reduced dataset, in which these subjects, whose times to ≥1mm ST-segment depression were strictly censored, were excluded.

The only significant interaction with exercise type was in the effect of body weight; every 10 kg increase in weight was associated with a 48.5 sec (95% CI, 32.1-65.0 sec) greater exercise time if exercising using a bicycle ergometer (according to the

Effect Estimate (secs) (95% CI) p-value	All Data	Censored Observations Excluded
Intercept	438.0 (408.5, 467.5) <0.0001	407.5 (373.7, 441.3) <0.0001
Exercise (Bicycle – Treadmill)	199.4 (171.3, 227.5) <0.0001	191.3 (157.0, 225.5) <0.0001
Treatment (Nifedipine – Atenolol)	5.8 (-26.5, 38.2) 0.73	0.3 (-38.8, 39.4) 0.99
Treatment (Combination – Atenolol)	0.7 (-31.7, 33.2) 0.96	-5.8 (-45.0, 33.5) 0.77
Gender (Female – Male)	-106.1 (-146.4, -65.7) <0.0001	-102.6 (-152.9, -52.3) 0.0001
Age (/10 years)	-68.4 (-86.2, -50.6) <0.0001	-49.9 (-71.8, -28.1) <0.0001
Weight (/10 kg : Treadmill)	-10.2 (-30.5, 10.1) 0.33	-5.1 (-29.1, 19.0) 0.68
Weight (/10 kg : Bicycle)	48.6 (32.1, 65.1) <0.0001	46.1 (26.0, 66.1) <0.0001

Table 5.1 Linear regression model effects estimates with 95% confidence intervals and p-values, for models of time to ≥1mm ST-segment depression or end of exercise, using all data or restricted to subjects experiencing ≥1mm ST-segment depression

model using all data), whilst for those exercising using a treadmill, body weight showed no association with the time to ≥1mm ST-segment depression.

Increased age and female gender were both associated with reduced exercise time, but there was no evidence that any treatment affected the time to ≥1mm ST-segment depression (or end of exercise).

Excluding those individuals that did not reach ≥1mm ST-segment depression introduces more uncertainty about the effect estimates, as the sample size was reduced by 33% amongst those using treadmill exercise and 41% in those using a bicycle. However, the estimates were not substantively affected, and so the conclusions of the model using all subjects could be adopted with the knowledge that the results are not overly influenced by those that do not reach the specified endpoint of ≥1mm ST-segment depression.

Figure 5.2 shows the Normal probability plot of the residuals from the model using all data, suggesting an adequate model fit in respect of the residual distribution. This figure is similar for the model using uncensored observations only.

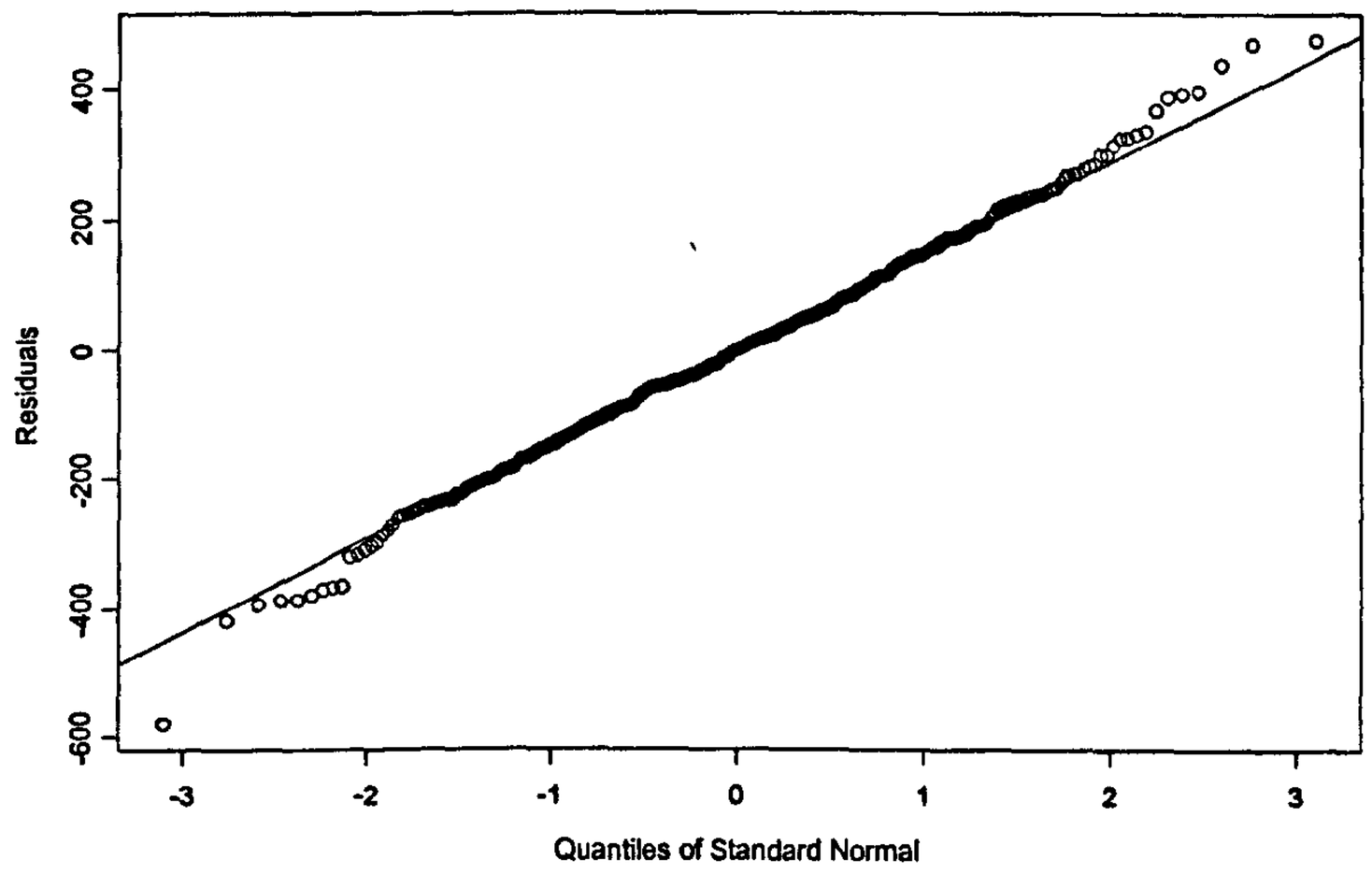


Figure 5.2 Normal probability plot of residuals from linear regression model of time to ≥1mm ST-segment depression or end exercise

Since the time to ≥1mm ST segment depression is often censored, it is preferable to use survival techniques[52]. Table 5.2 gives the results of applying log rank tests to these data. There was no evidence of differences between treatment groups (p=0.14), when all data were combined. However, previous analyses would suggest fundamental differences between bicycle and treadmill exercise, and when the test was stratified by exercise type there was some evidence of differences between treatments (p=0.095). Splitting the data by mode of exercise, it appeared that under treadmill exercise, those on Atenolol were more likely to experience ≥1mm ST-segment depression than would be expected. Those using bicycle exercise, however, showed no differences in survival from ≥1mm ST-segment depression.

To estimate treatment and covariate effects simultaneously requires the use of regression models for survival data; the most commonly applied method is the Cox proportional hazards model[83]. Table 5.3 shows the effect estimates obtained from a Cox model for the time to ≥1mm ST-segment depression, fitted with separate baseline hazard functions for each exercise type. The model shown includes terms for treatment, gender and age, with gender and age having different effects under each exercise type, since these interactions were found to improve the fit of the model. Body weight was not found to have a significant effect in this model.

Those on either Nifedipine or Combination therapy had lower hazard for suffering ≥1mm ST-segment depression than those on Atenolol only, suggesting a protective

Treatment			All	Treadmill	Bicycle	Stratified by Exercise Type
Atenolol	Subjects		218	115	103	218
	Events	Obs	157	90	67	157
		Exp	140.1	74.4	64.2	138.6
Nifedipine	Subjects		217	111	106	217
	Events	Obs	135	74	61	135
		Exp	137.6	81.9	56.9	138.9
Combination	Subjects		211	110	101	211
	Events	Obs	126	71	55	126
		Exp	140.2	78.7	61.8	140.5
χ^2 (2 df)			4.0	5.6	1.3	4.7
p			0.14	0.061	0.52	0.095

Table 5.2 Log rank test results for time to ≥1mm ST depression for all subjects, separately by exercise type and stratified by exercise type; numbers of subjects, numbers of event occurrences (Obs), expected numbers of event occurrences (Exp), with associated χ^2 statistics and p-values

	Exercise Type	Hazard Ratio			PH Test
		Estimate	95% CI	p	χ^2 p
Treatment Effects					
Nifedipine – Atenolol		0.81	(0.64, 1.02)	0.076	0.1 0.79
Combination – Atenolol		0.76	(0.60, 0.97)	0.027	0.5 0.48
Covariate Effects					
Gender (Female – Male)	Treadmill	1.26	(0.84, 1.89)	0.26	0.5 0.49
	Bicycle	2.92	(1.89, 4.51)	<0.0001	0.5 0.47
Age (/10 years)	Treadmill	1.36	(1.15, 1.62)	0.0005	2.6 0.11
	Bicycle	1.91	(1.53, 2.38)	<0.0001	0.3 0.56

Table 5.3 Effect estimates, 95% CIs and p-values from Cox proportional hazards model for time to ≥1mm ST-segment depression, with baseline hazard function stratified by exercise type, with χ^2 statistics and p-values for goodness-of-fit with respect to proportional hazards assumption, as determined by the time varying coefficients method (section 3.5.2.5)

effect of Nifedipine compared to Atenolol. Under treadmill exercise, gender did not affect the hazard for suffering an ischaemic event, whereas on a bicycle, the hazard for women was nearly three times that for men. Similarly, the effect of age was greater with bicycle compared to treadmill exercise, though under both forms, older patients were more likely to suffer the event.

Goodness-of fit with respect to the proportional hazards assumption was tested by the time varying coefficients method[93] (section 3.5.2.2). None of the variables in the model showed any sign of non-proportionality.

5.3 Logistic Model

Thompson[98] suggested a logistic regression model for grouped survival data, modelling the conditional probability of surviving to the end of an interval given that a subject survived to the end of the previous interval without experiencing an event. In other words, if π_{ij} is the probability that the i^{th} subject survives to the end of the j^{th} interval, conditional upon their survival through the first j-1 intervals, then the model assumes that

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta' \mathbf{z}_{ij} + \gamma_j \qquad \textbf{(Eq. 5.1)}$$

where $\mathbf{z}_{ij}$ is a vector of (possibly interval dependent) covariates, β a vector of associated parameters and γ_j is a parameter to model the j^{th} interval effect.

This model can be fitted directly using logistic regression methods as supplied by standard statistical packages, though this would require that censoring times are interval censored in the same way as failure times. That is, when an individual is withdrawn from observation without experiencing an event, this happens at the end of an interval. With exercise test data the censoring times are not so well behaved, since a subject may withdraw from a test at any time due to fatigue or pain, whether anginal or not. The exact time that a subject withdraws from a test will be recorded, and at withdrawal, the ECG and with it the level of ST-segment depression will be measured.

Consequently, data on each individual will fall into one of three categories. For those who first experience ≥1mm ST-segment depression at the end of an interval (the j^{th} interval, say), their contribution to the data will be that no event occurred during the first j-1 intervals, but it did occur at some time during the j^{th}. Those who do not experience an event at all, and withdraw from exercise during the j^{th} interval, will contribute that no event occurred during the first j-1 intervals or the start if the j^{th} interval. That is not to say that the event would not have occurred had the patient been able to continue exercising to the end of that interval. Most subjects will fall into one of these two types. Some, however, will withdraw from exercise during an interval (the j^{th}, say) and when their ECG is recorded will be suffering ≥1mm ST-segment depression.

Such an individual would contribute the information that no event occurred during the first j-1 intervals, but there was an event at some time during the start of the j^{th} interval.

A decision must therefore be made about how to treat partially observed intervals. Either some of the information about these intervals must be ignored, or an attempt must be made to model these data.

5.3.2 Ignoring Partially Observed Intervals

To ignore some of the information will involve either treating these partial intervals as if they were complete (ignoring the fact that only part of the interval was observed), or ignoring them completely, and only using the information gathered on wholly observed intervals.

Example 5.2 Logistic regression model for interval censored times to ≥1mm ST-segment depression, ignoring partial observation of final intervals

To fit the logistic model for interval censored survival data, interval effects are included as categorical variables with as many levels as there are intervals. In the case of those exercising with a treadmill, the longest exercise time without ≥1mm ST-segment depression was into the 13^{th} minute, and with a bicycle ergometer, the longest test lasted into the 21^{st} minute. When fitting any logistic regression model with a large number of categorical variables, care must be taken that there are no levels of the predictor variables for which there are either no successes or no failures. In other words, the γ_j in (Eq. 5.1) are estimable for levels of predictor variables at which there occur both ischaemic events and survivals.

As with previous models, the data strongly suggested that the relationships between time and the probability of surviving an interval were different for the two types of exercise, so that the model included an interaction between exercise type and interval of exercise. As a result, no data could be used from the 13^{th} minute of treadmill exercise, and the 2^{nd}, 17^{th}, 19^{th}, 20^{th} and 21^{st} minutes of bicycle exercise.

Figure 5.3 shows the interval effect estimates and Table 5.4 shows the treatment and covariate effect estimates from the final models fitted to these data. Table 5.4 shows the results from the same model applied by either treating partially observed intervals as complete, or by ignoring these intervals completely. Out of 641 subjects without missing data that were included in these analyses, there were 110 partially observed intervals, each being the last interval of observation for a single patient. Of these 110

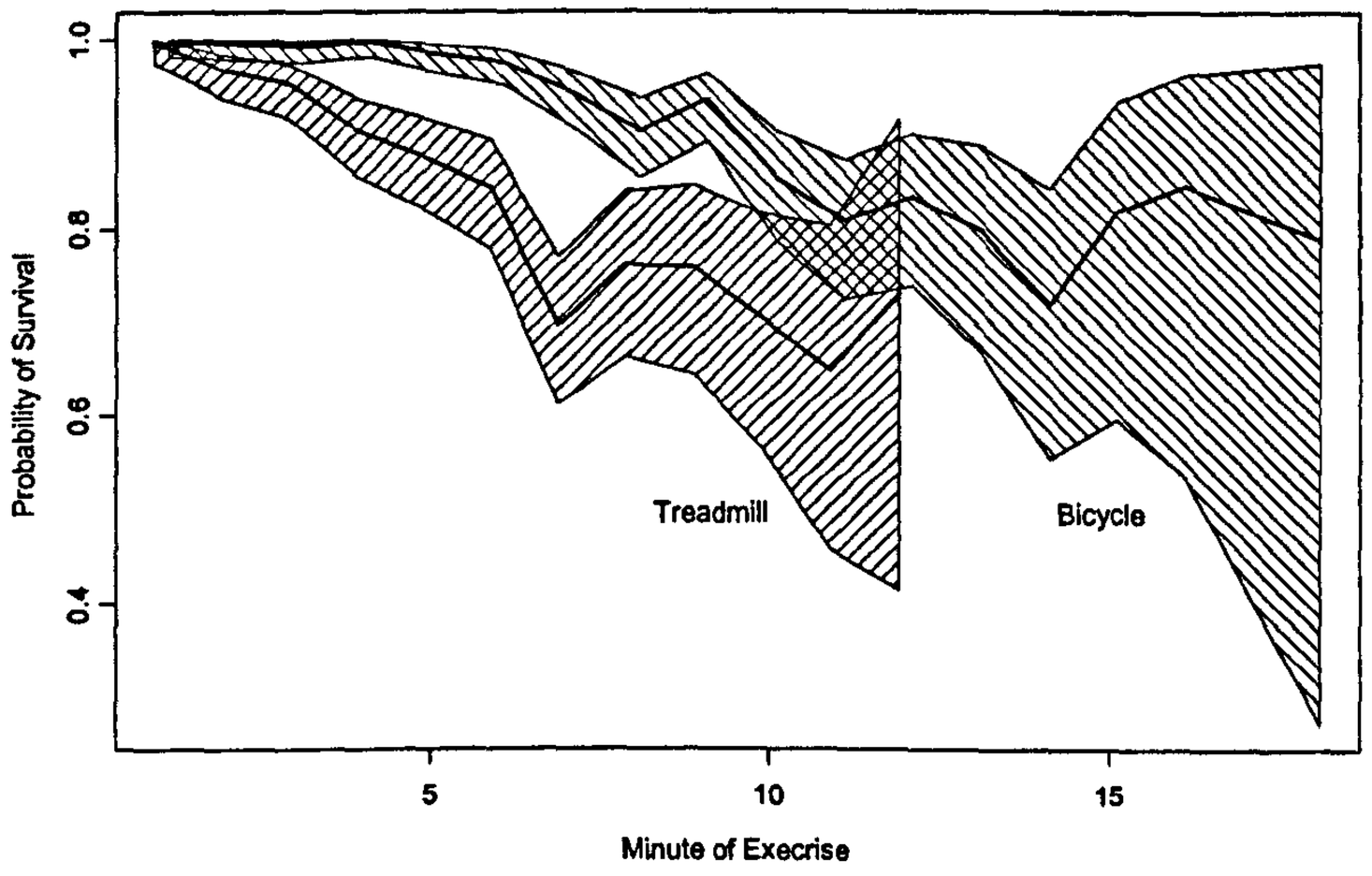


Figure 5.3 Interval effect estimates (for a male patient aged 60 years, treated with Atenolol alone), with pointwise 95% CIs from logistic model for interval censored data applied to the TIBET time to ≥1mm ST-segment depression data, treating partially observed intervals as complete

patients, 22 (20.0%) suffered the event of ≥1mm ST-segment depression during their final interval, accounting for 5.3% of the 413 events that occurred.

The probability of survival through each interval for patients exercising on a treadmill clearly decreases more rapidly than for those on a bicycle, for whom there is very little probability of experiencing ≥1mm ST-segment depression for the first 6 minutes.

Overall, there are statistically significant differences between treatment groups (p=0.048), with Nifedipine appearing to reduce the occurrence of ischaemic events relative to Atenolol. Gender did not appear to influence the probability of experiencing ≥1mm ST-segment depression whilst exercising with a treadmill, though with a bicycle, women were more than 3 times as likely to have an event (p-value for heterogeneity, 0.0029). Similarly, there were significant differences in the effect of age between exercise types (p=0.021), with older patients less likely to survive an interval without an event, particularly when using a bicycle.

There was little difference in effect estimates between the two models, though CIs are wider when partially observed intervals are ignored; this greater uncertainty is inevitable given the loss of data. The loss of precision is small, however, since only 5% of events are lost.

Odds Ratios		Treating partial intervals as complete			Excluding partial intervals		
		Estimate	95% CI	p	Estimate	95% CI	p
Treatment Effects							
Nifedipine – Atenolol		1.26	(0.96,1.64)	0.093	1.24	(0.94,1.63)	0.13
Combination – Atenolol		1.39	(1.06,1.82)	0.016	1.44	(1.09,1.90)	0.011
Covariate Effects							
Gender (Female – Male)	Treadmill	0.89	(0.57,1.41)	0.63	0.76	(0.47,1.23)	0.26
	Bicycle	0.32	(0.19,0.52)	<0.0001	0.31	(0.19,0.52)	<0.0001
Age (/10 years)	Treadmill	0.72	(0.59,0.88)	0.0011	0.71	(0.58,0.87)	0.0011
	Bicycle	0.50	(0.39,0.64)	<0.0001	0.49	(0.38,0.63)	<0.0001

Table 5.4 Treatment and covariate effect estimates (as odds ratios) with 95% CIs and p-values from logistic model for interval censored data applied to the TIBET time to ≥1mm ST-segment depression data, either treating as complete or excluding partially observed intervals

5.3.3 Modelling Partially Observed Intervals

In Example 5.2 it was noted that about 1 in 6 patients were not observed to suffer ≥1mm ST-segment depression up to and including their last scheduled ECG recording prior to stopping exercise, so that their final measurement came at withdrawal, allowing a shorter time for the event to occur. These patients accounted for only 5% of all occurrences of the event, suggesting that the occurrence of ≥1mm ST-segment depression is less likely in these partial intervals.

This could be tested within the logistic model, by including a term to indicate that an interval is only partly observed and/or by the addition of a variable, q_{ij}, denoting the proportion of the j^{th} interval for which the i^{th} subject was unable to exercise (so that q_{ij} is zero in most instances and between 0 and 1 for partially observed intervals). If the effect of q_{ij} is positive, this would imply that the chance of suffering an event is greatest in fully observed intervals.

Alternatively, if an assumption is made about the distribution of failures within intervals, the whole of the data can be modelled. For subject i during interval j, denote the covariate vector by $\mathbf{z}_{ij}$,, the event indicator by d_{ij}, and the proportion of the interval for which the subject was able to exercise by $p_{ij} = 1 - q_{ij}$. p_{ij} will take the value 1 for each interval except possibly the last for that subject, when it will take a value in the interval (0, 1].

When $p_{ij} = 1$ and a subject is able to exercise throughout interval j, the expectation of $1 - d_{ij} = s_{ij}$ is π_{ij}. However, when $p_{ij} < 1$, and the subject can exercise for only part of the interval, then assuming that the hazard for the event is constant during that interval, we can write the expected value of the survival indicator (s_{ij}) as

$$\mu_{ij} = E(s_{ij}) = \pi_{ij}^{p_{ij}}.$$

For a standard logistic regression model, $\mu_{ij} = \pi_{ij}$, and the link function is given by

$$g(\mu_{ij}) = \text{logit}(\mu_{ij}) = \frac{\mu_{ij}}{1-\mu_{ij}} = \frac{\pi_{ij}}{1-\pi_{ij}};$$

in this case, however, to preserve the original form of the logistic model, the link function can be written as

$$g(\mu_{ij}) = \frac{\mu_{ij}^{1/p_{ij}}}{1-\mu_{ij}^{1/p_{ij}}} = \frac{\pi_{ij}}{1-\pi_{ij}}. \qquad \textbf{(Eq. 5.2)}$$

Given the link function, it is now a relatively simple matter to fit this as a generalised linear model, using iteratively reweighted least squares[99]. Treatment effect differences and the effects of other covariates can be estimated in the same way as with other linear regression models.

Example 5.3 Logistic regression model for time to ≥1mm ST-segment depression with adjustment for partially observed intervals

Table 5.5 shows the treatment and covariate effects from the logistic model applied to the time to ≥1mm ST-segment depression, adjusting for the proportion of each interval for which each patient is under observation. Also shown are the estimates from the same model applied without including information about partially observed intervals, in which all intervals of observation are assumed to be complete. The estimates from the two models are very similar, suggesting that explicitly modelling the partially observed intervals is unnecessary.

This is confirmed by extending the models shown in Table 5.4 to include terms for the proportion of each interval for which each subject was under observation and/or the fact of whether each interval was fully or partially observed. None of these modifications offered any improvement to the fit of the model (data not shown), indicating that suffering ≥1mm ST-segment depression was not associated with withdrawal from the test during an interval.

		Ignoring Partial Intervals	Modelling Partial Intervals		
		Odds Ratio	Odds Ratio	(95% CI)	p
Treatment Effects					
Nifedipine – Atenolol		1.26	1.26	(0.96, 1.65)	0.099
Combination – Atenolol		1.39	1.36	(1.03, 1.79)	0.030
Covariate Effects					
Gender (Female – Male)	Treadmill	0.89	0.82	(0.51, 1.31)	0.40
	Bicycle	0.32	0.31	(0.19, 0.52)	<0.0001
Age (/10 years)	Treadmill	0.72	0.71	(0.58, 0.87)	0.0009
	Bicycle	0.50	0.50	(0.39, 0.64)	<0.0001

Table 5.5 Treatment and covariate effect estimates from logistic model incorporating information regarding partially observed intervals, with estimates from model ignoring partially observed interval shown for comparison

5.3.4 Goodness of Fit

An effective method for testing goodness-of-fit for logistic regression models is to rank observations in terms of their associated risk as predicted by the model, and to compare the observed and predicted numbers of events that occur in groups defined by this measure, for example in quintiles or deciles of predicted risk[100].

Example 5.4 Goodness-of-fit of logistic regression model for time to ≥1mm ST-segment depression

The different applications of the logistic model for the time to ≥1mm ST-segment depression shown in Example 5.2 and Example 5.3 appear to fit the model equally well, so the standard model treating partial intervals as complete, might be considered the best choice, on the grounds that it is simpler than attempting to model the partially observed intervals, and it does not waste data by excluding these intervals.

This example explores the goodness-of-fit of these models formally, by comparing the observed and expected numbers of occurrences of ≥1mm ST-segment depression in subgroups defined by quintiles of predicted risk. Predicted risk was defined as the average predicted risk under the three models considered.

Quintile of Predicted Risk		Numbers of Events (≥1mm ST-Segment Depression) Observed: Total	Expected: Model 1	Expected: Model 3	Observed: Completed Intervals	Expected: Model 2
1st		4	3.3	3.3	4	3.3
2nd		15	20.3	20.2	14	19.9
3rd		59	54.3	54.0	59	52.9
4th		117	112.0	108.9	111	105.2
5th		218	223.1	225.8	203	209.7
Goodness-of-fit:						
All Data	χ^2	-	2.3	3.7	-	4.9
	p		0.68	0.44		0.30
Treadmill	χ^2	-	0.9	2.7	-	2.7
	p		0.93	0.60		0.61
Bicycle	χ^2	-	2.2	2.6	-	3.1
	p		0.70	0.63		0.54

Table 5.6 Observed and expected numbers of occurrences of ≥1mm ST-segment depression according to Model 1 (logistic model, treating partially observed intervals as complete), Model 2 (logistic model, excluding partial intervals) and Model 3 (logistic model adjusting for partial intervals), with corresponding χ^2 goodness-of-fit statistics and p-values as a global test, and applied to treadmill and bicycle data separately

Writing O_i and E_i for the observed and expected numbers of events in quintile i, the statistic used was $\sum_i \frac{(O_i - E_i)^2}{E_i}$ which, if the models fits, will have a χ^2 distribution on 4 df. Furthermore, since there is a clear distinction between those using treadmill and bicycle exercise, the fit is assessed in quintiles of predicted risk stratified by exercise type. For the treadmill data, the statistic was compared to a χ^2 distribution on 3 df, since none of those using a treadmill were in the lowest quintile of predicted risk.

Table 5.6 shows the results of this assessment. None of the models demonstrate any lack of fit, and the earlier view that the simple logistic model, treating partial intervals as complete, would be the most useful in practice.

5.4 Proportional Hazards Model

The proportional hazards model has, since the development of the semi-parametric Cox model, been the most widely used method of analysing survival data[84]. It has the property of robustness to non-proportional hazards[101], making it a reliable tool for many situations. A proportional hazards model applied to interval censored data

might show similar properties to the continuous data model. The Cox model can be adapted to interval censored data, under the umbrella of generalized linear models[102].

Under the proportional hazards model, the hazard for subject i at time t is

$$\lambda(t \mid \mathbf{z}_i) = \lambda_0(t)\exp(\mathbf{z}_i\beta)$$

where $\mathbf{z}_i$ is a vector of covariates for the i^{th} individual. Interest lies in the conditional probability of failure during any particular interval,

$$\theta_{ij} = P(t_{j-1} < T_i < t_j \mid T_i > t_{j-1})$$
$$= 1 - \frac{S(t_j)}{S(t_{j-1})}$$

where $S_i(.)$ is the survivor function for the i^{th} subject. Now,

$$S(t) = \exp\left(- \int_0^t \lambda(u)\,du \right)$$
$$= \exp\left(- \exp(\mathbf{z}_i\beta) \int_0^t \lambda_0(u)\,du \right)$$

so that

$$\theta_{ij} = \exp\left(- \exp(\mathbf{z}_i\beta) \int_{t_{j-1}}^{t_j} \lambda_0(u)\,du \right)$$

and

$$\log(-\log(1-\theta_{ij})) = \log(-\log(1-\theta_{0j})) + \mathbf{z}_i\beta$$

where θ_{0j} is the conditional probability of failure during interval j given survival to the end of the previous interval for a subject with covariates **0**.

The model can be fitted using the methodology of Generalized Linear Models, with the link function

$$g(\mu) = \log(-\log(1-\mu)).$$

Again, as for the logistic model (Eq. 5.2), information about the proportion, p_{ij}, for which subject i was observed to exercise during interval j, can be incorporated into the link function. Using an assumption of piecewise constant hazards, the link function becomes

		Ignoring Partial Intervals		Modelling Partial Intervals	
		Hazard Ratio (95% CI)	p	Hazard Ratio (95% CI)	p
Treatment Effects					
Nifedipine – Atenolol		1.08 (0.96, 1.21)	0.20	1.08 (0.96, 1.21)	0.22
Combination – Atenolol		1.13 (1.01, 1.27)	0.041	1.12 (1.00, 1.26)	0.059
Covariate Effects					
Gender (Female – Male)	Treadmill	0.94 (0.77, 1.15)	0.55	0.91 (0.74, 1.12)	0.37
	Bicycle	0.69 (0.56, 0.85)	0.0004	0.69 (0.56, 0.85)	0.0005
Age (/10 years)	Treadmill	0.87 (0.79, 0.95)	0.0015	0.86 (0.79, 0.95)	0.0016
	Bicycle	0.79 (0.72, 0.87)	<0.0001	0.79 (0.72, 0.87)	<0.0001
Goodness-of-Fit	χ^2	3.1		3.2	
	p	0.54		0.53	

Table 5.7 Model effect estimates, with 95% confidence intervals and p-values, from proportional hazards regression models for interval censored data, either ignoring partial intervals (treating as if they were complete) or assuming a constant hazard rate within intervals and adjusting for partial intervals, with χ^2 goodness-of-fit statistics

$$g(\mu) = \log\left(-\log\left(1-\mu^{1/p}\right)\right).$$

The model can now be fitted by iteratively reweighted least squares, and treatment and covariate effects estimated.

Example 5.5 Proportional hazards model for interval censored times to ≥1mm ST-segment depression

Table 5.7 shows the effect estimates from proportional hazards models for interval censored data applied to the time to ≥1mm ST-segment depression data from the TIBET Study. Two models are shown, the first treating intervals that were only partially observed as if they were complete intervals, and the second modelling data from the intervals using an assumption of constant hazards within an interval. The two models give almost identical effect estimates. Also shown are goodness-of-fit statistics calculated in the same way as in Example 5.4, comparing the observed and expected numbers of events in quintiles of predicted risk (calculated in this instance as the

average predicted risk over the two models). Neither model shows any lack of fit, and so the first model would appear to be the best choice in practice, since any benefit to be gained by modelling partially observed intervals in this way would seem to be negligible.

5.5 Imputation

An alternative method for analysing interval censored data is to consider the time that the event occurred as missing data. The actual survival time can be imputed, or estimated, given that failure occurred at some point during a particular interval, and standard survival methodology then applied to the imputed survival times. There are a number of methods by which survival times could be imputed; some of these are outlined below.

5.5.1 Right Imputation

One possibility for imputing the time that an event occurred, given that it occurred at some point during an interval $(t_{j-1}, t_j]$, is to use the right hand end of the interval, t_j. This is identical to ignoring the interval censoring, and using standard techniques on the observed failure times, as if they were fully observed. The results of applying this approach have been given in Example 5.1.

5.5.2 Single Imputation

If it is assumed that more accurate results are obtained if a more accurate guess at the true failure time is made, then it would seem probable that using any value within the interval during which failure occurred would be better than simply using the maximum possible value. For example, the mid-point of the interval in which the event occurred is a natural candidate.

With exercise test data, however, it is possible to make a more considered estimate. As well as the fact that ≥1mm ST-segment depression has occurred during an interval, the level of ST-segment depression will be recorded at the start and the end of the interval in question. If, for example, the ST-segment was depressed by 0.9mm immediately prior to the interval, but by 1.5mm at the end of the interval, then it might be deduced that the event of ≥1mm ST-segment depression occurred near to the start of that interval. Another natural assumption would be that changes in ST-segment depression occur linearly during the interval, so that the time at which it crosses the

threshold of 1mm can be interpolated; in the case outlined above, if the interval is one minute long, ≥1mm ST-segment depression would be imputed to have occurred 10 seconds into the interval.

5.5.3 Multiple imputation

The fact that the time at which ≥1mm ST-segment depression occurred is unobserved would imply that any imputed time is liable to a degree of uncertainty. By imputing a single value for each time to ≥1mm ST-segment depression and analysing as if this were the true data, model effect estimates may be obtained with standard errors that are too small. Because they are based on imputed failure times but estimated as if the failure times were observed, they will not reflect the degree of uncertainty with which they are estimated.

To incorporate this uncertainty into the model, the method of multiple imputation[103] can be used. Rather than create a single dataset, with imputed values for every individual for whom the failure time was interval censored, several datasets are created, with imputed values in each that are to some degree random. For example, when creating a dataset with randomly imputed failure times, individual times could be based on sampled values from a uniform distribution. Alternatively, linear interpolation could be used to obtain an expected value for the time that ≥1mm ST-segment depression occurred, and multiple imputations could then be made based on a Beta distribution with a mean value equal to the proportion of the interval after which linear interpolation would suggest the event to have occurred.

Once multiple imputed datasets have been created, each is analysed using the preferred statistical method. For each of the J analyses, a set of model effects estimates, $\hat{\beta}_j$, and their variance-covariance matrices, V_j, are produced. The mean of the effect estimates,

$$\hat{\beta} = \frac{1}{J}\sum_j \hat{\beta}_j ,$$

is used as the final estimate of the model parameters. The variance of $\hat{\beta}$ is then considered to be

$$V = V_W + \left(1 + \frac{1}{J}\right)V_B ,$$

		Method of Single Imputation			
		Midpoint		Linear Interpolation	
		Haz. Ratio (95% CI) p	PH Test χ^2 p	Haz. Ratio (95% CI) p	PH Test χ^2 p
Treatment Effects					
Nifedipine - Atenolol		0.81 (0.64, 1.02) 0.072	0.1 0.78	0.73 (0.57, 0.94) 0.016	0.2 0.70
Combination - Atenolol		0.77 (0.60, 0.97) 0.029	0.5 0.49	0.67 (0.52, 0.86) 0.0022	0.8 0.36
Covariate Effects					
Gender (Female – Male)	Treadmill	1.19 (0.79, 1.78) 0.40	0.5 0.50	1.13 (0.73, 1.75) 0.58	0.6 0.45
	Bicycle	2.91 (1.88, 4.49) <0.0001	0.6 0.44	2.88 (1.82, 4.58) <0.0001	0.3 0.57
Age (/10 years)	Treadmill	1.35 (1.14, 1.61) 0.0007	2.5 0.12	1.31 (1.09, 1.58) 0.0043	2.5 0.11
	Bicycle	1.92 (1.54, 2.39) <0.0001	0.3 0.61	2.11 (1.65, 2.71) <0.0001	0.0 0.94

Table 5.8 Effect estimates, with 95% CIs and p-values, from Cox proportional hazards models for time to 1mm ST-segment depression, with time of event imputed as either the midpoint of final interval of exercise, or by linear interpolation from ST data before and after final interval of exercise. Also shown are tests of proportional hazards assumption for each effect, by the time varying coefficients method (section 0)

where the within- and between imputation variances are estimated by

$$V_W = \frac{1}{J}\sum_j V_j \text{, and}$$

$$V_B = \frac{1}{J-1}\sum_j \left(\beta_j - \hat{\beta}\right)^T \left(\beta_j - \hat{\beta}\right).$$

Example 5.6 Imputation methods for time to 1mm ST-segment depression

Table 5.8 shows the results of fitting Cox proportional hazards models for the time to 1mm ST-segment depression, where the time to event is singly imputed for those who experience the event during exercise. The two methods of imputation used were midpoint and linear interpolation. The estimates obtained after midpoint imputation are almost identical to those obtained from the model applied to the observed times of occurrence of ≥1mm ST-segment depression; using the observed

		Hazard Ratio		
		Estimate	95% CI	p
Treatment Effects				
Nifedipine - Atenolol		0.80	(0.64, 1.02)	0.072
Combination - Atenolol		0.76	(0.60, 0.96)	0.024
Covariate Effects				
Gender (Female – Male)	Treadmill	1.19	(0.79, 1.79)	0.40
	Bicycle	2.92	(1.89, 4.51)	<0.0001
Age (/10 years)	Treadmill	1.35	(1.14, 1.61)	0.0006
	Bicycle	1.92	(1.54, 2.40)	<0.0001

Table 5.9 Treatment and covariate effect estimates, with 95% CIs and p-values from Cox proportional hazards models applied to time to ≥1mm ST-segment depression using multiple imputation

times can be thought of as right imputation, since the time at the right-hand end of the interval in which the event is first known to have occurred is used. The estimates after linear interpolation are also similar, though treatment effects and the effect of weight for those exercising using a bicycle are larger, with the log hazard ratios moving away from zero by approximately one standard error. Furthermore, standard errors of effect estimates using midpoint imputation are very similar to the original model, but those from the linear interpolation model are all larger; in general by a factor of 6 to 9% (as log hazard ratios), though the standard error of the weight effect estimate on a bicycle is 13% larger when using linear interpolation.

Table 5.9 shows the treatment and effect estimates obtained by application of multiple imputation. Ten datasets were imputed, assuming a uniform distribution of failure times within intervals, and Cox proportional hazards models applied to each. The mean effect estimates are shown, with 95% CIs and p-values derived using the estimate of the total variance of the parameter estimates. The estimates obtained by this method, and their confidence intervals are very similar to those obtained by right or midpoint imputation, and the additional variation due to multiple imputation is slight. This would suggest that single imputation methods are adequate in this instance.

CHAPTER 6 Simulation Study I: Analysis of Interval Censored Survival Data

This chapter presents the results of a simulation study to investigate the relative performance of different models for analysing interval censored survival data. The main aim is to investigate the extent to which the analysis of times until significant ST-segment depression might be influenced by the choice of model used. As well as looking at different models for the analysis, the size of the treatment effect and the level of interval censoring (analogous to the time between successive recordings of ST-segment depression) will be varied, in order to find those models that perform best under a range of situations.

6.1 Generation of simulated data

Data were simulated in a parallel groups design, with two groups (to be referred to as groups A and B) and 100 subjects in each group. Simulations were generated in five batches, each with a different "treatment" effect, represented by a constant hazard ratio between groups B and A. The hazard ratios were 1 (corresponding to no treatment effect), 0.80, 0.67, 0.57 and 0.50. In each batch there were 1000 simulated studies. Each study was subjected to different levels of interval censoring, to simulate the effect of the frequency with which the ECG is recorded during an exercise test.

Failure times were generated from a Weibull distribution with a shape parameter of 2. Group A was scaled to have a mean failure time of 400, by a scale parameter of $\gamma_0 = \Gamma\left(1+{}^{1}\!/\!_{\alpha}\right)\Big/_{400}$, where $\alpha = 2$ was the shape parameter. Group B had a scaling factor of $\gamma_0 h^{1/\alpha}$, with h the hazard ratio between groups B and A. Censoring times were generated from an exponential distribution with a mean of 800 in both groups, independently of the failure times.

If a simulated subject had a failure time less than or equal to their censoring time, it was considered to have experienced the event of interest. If the censoring time was after the end of the interval in which the event occurred, then the "observed" failure time was at the end of that interval. If, however, the censoring time came after the failure time, but before the end of that interval, the "observed" failure time was taken to be the censoring time, to reflect the possibility that a subject who withdraws from an exercise test could be observed to have experienced significant ST-segment depression since the previous recording of the ECG. Otherwise, when censoring occurred before failure, the subject was censored without an event at the generated censoring time. Given that simulated failure and censoring times could be arbitrarily large, any large survival times were censored at the end of the first interval after 1000. This is to mimic the situation where an exercise protocol will be of finite length, so that any subject who exercises for more than a predetermined time will discontinue the test.

Several interval widths were considered; 25, 50, 75, 100, 125, 150, 175, 200, 300, 400 and 500. These correspond to 2, 3, 4, 5, 6, 7, 8, 10, 14, 20 and 40 intervals (or recordings of the ECG). This range encompasses the usual range of maximum exercise times (about 10 to 20 minutes, with ECG recordings every minute) as well as exploring the effects of having narrower or broader intervals.

6.2 Models

Six different models were investigated. Two methods based on the t-test were used, either using all survival times regardless of whether or not the time was a failure or a censoring, or using only those survival times that were not censored. In either case, the response variable was the simulated time of the first observation of ≥1mm ST-segment depression. Two Cox proportional hazards models were implemented, one using the exact partial likelihood, and one using the Efron approximation to the partial likelihood. Since the data are interval censored, there will be many tied survival times, and the Efron approximation performs better than the Breslow approximation in the presence of a large number of ties. Two models designed for interval censored survival data were employed: the logistic model of the probability of survival to the end of an interval given survival to the start of that interval, and the proportional hazards model for interval censored survival data. With both of these methods, effect estimates were multiplied by –1 (since they estimate probabilities or hazards of survival to the end of an interval, rather than of failure) to make them comparable to the other methods. No

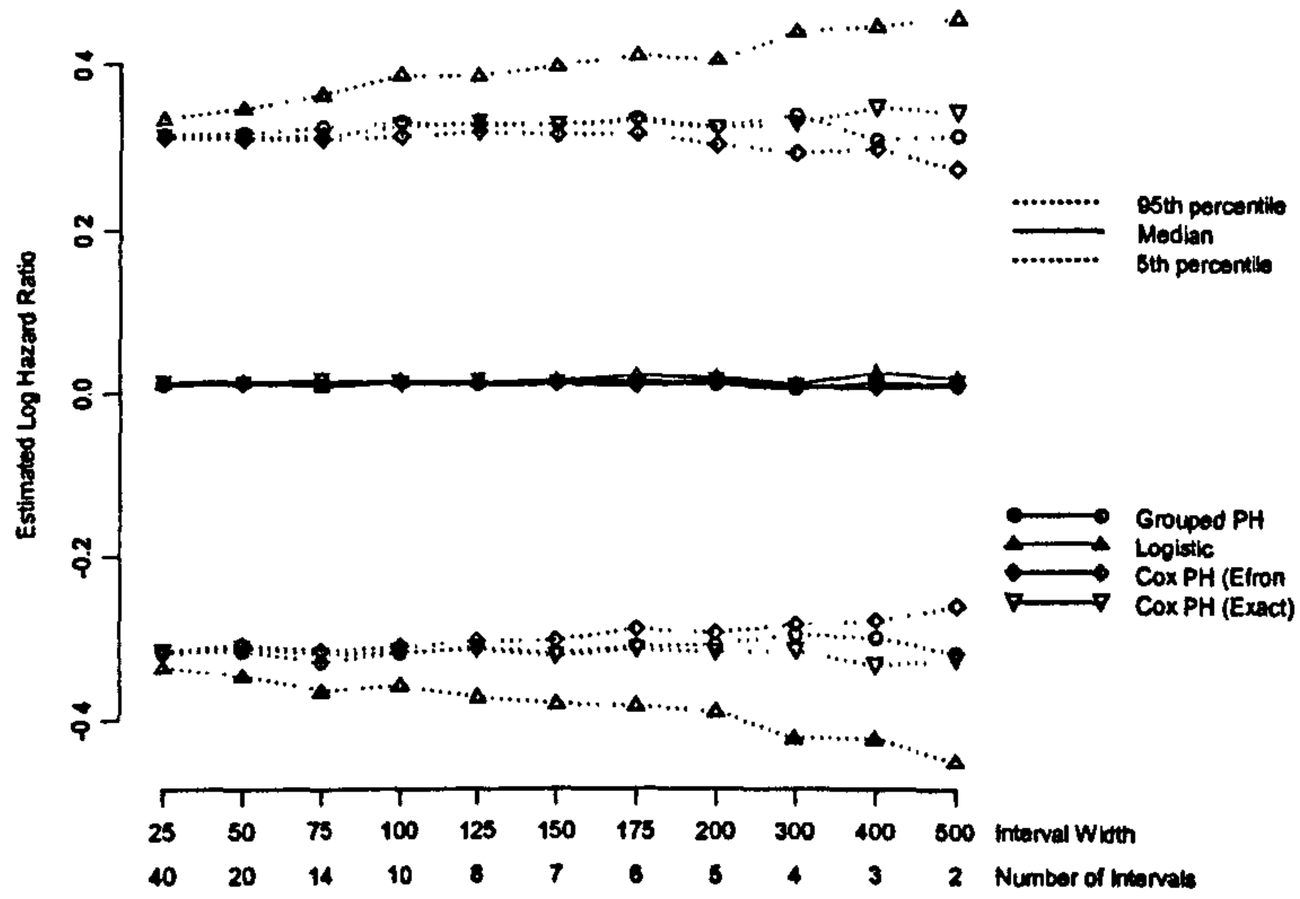


Figure 6.1 Median effect estimate against interval width, with 5th and 95th percentiles, from simulated trials with no treatment effect, comparing the grouped proportional hazards (PH) model, logistic model, Cox PH model using the Efron approximation to the partial likelihood and Cox PH model with exact partial likelihood

allowance was made for whether the failure times fell within or at the end of an interval, so that the only information that is used about survival times is the intervals into which they fall.

6.3 Results

6.3.1 Bias

The grouped data proportional hazards (PH) model and the two Cox PH models are estimating the same quantity in the log hazard ratio between the two groups, and the logistic model, in estimating the log odds ratio between the groups should produce estimates that approach the log hazard ratio as the interval size tends to zero[98]. Figure 6.1 shows the median estimated log hazard ratio plotted against the interval width from the 1000 simulated trials where the hazard ratio was 1, as well as points showing the cut-offs for the lower and upper 5%. The data being analysed by each model is equivalent in the two groups, so each model should on average find no differences between groups, regardless of how well each model fits the data.

Of more interest are the ranges of estimates produced by the different models. The logistic model is estimating a different quantity to that of the PH models, which approximates the log hazard ratio more closely as the interval widths become smaller[98]; the converse of this observation is that as interval widths become larger, the log odds

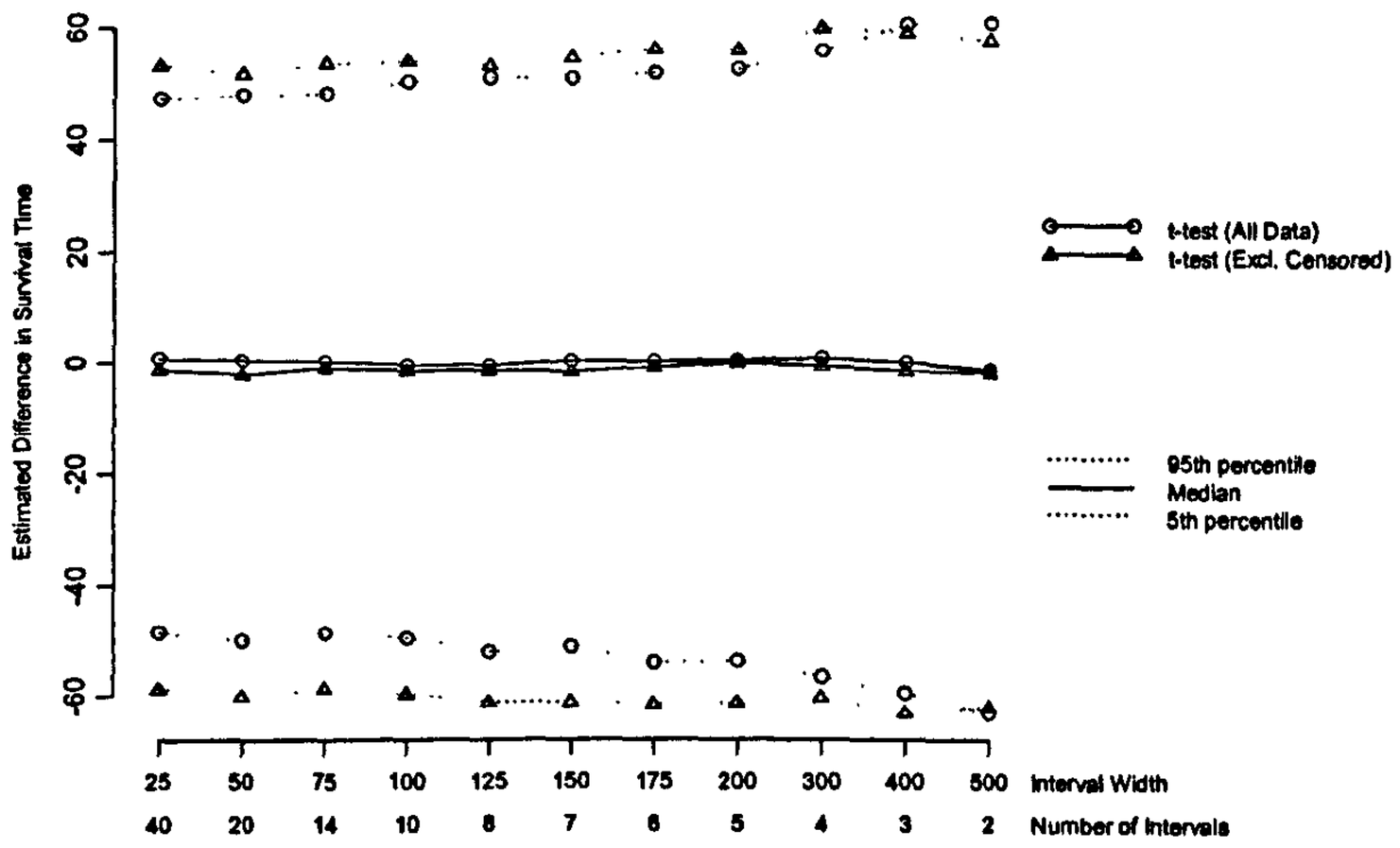


Figure 6.2 Median effect estimates against interval width, with 5th and 95th percentiles, from simulated trials with no treatment effect, comparing t-test of time to failure or censoring and t-test of time to failure excluding censored observations

ratio approximates the log hazard ratio less well, and as a result the range of estimates from the logistic model increases as the intervals become wider. The range of estimates produced by the grouped PH and exact Cox PH models remain approximately constant across all interval widths, though the effect estimates produced by the Cox PH model using the Efron approximation to the partial likelihood show slightly less variability as the intervals become wider.

As interval widths approach zero, the four models will become equivalent. The median limiting value appears to be slightly less than zero, indicating a minor bias over the 1000 simulations.

Figure 6.2 shows the same plot for the t-test methods. These models are estimating the mean difference in survival time between the groups, which will be zero in this case. The model using all data shows less variability than the analyses excluding censored observations, except possibly for the most extreme levels of interval censoring.

Figure 6.3 shows the median bias, with 5th and 95th percentiles, found under each model with a simulated treatment effect of 0.67. As the interval width approaches zero, the survival methods, including the logistic model, give estimates that are unbiased for the log hazard ratio. The logistic model becomes progressively more biased away from zero as the data are interval censored into fewer, wider intervals, except when there as few as two or three intervals, when the estimated log odds ratio between groups

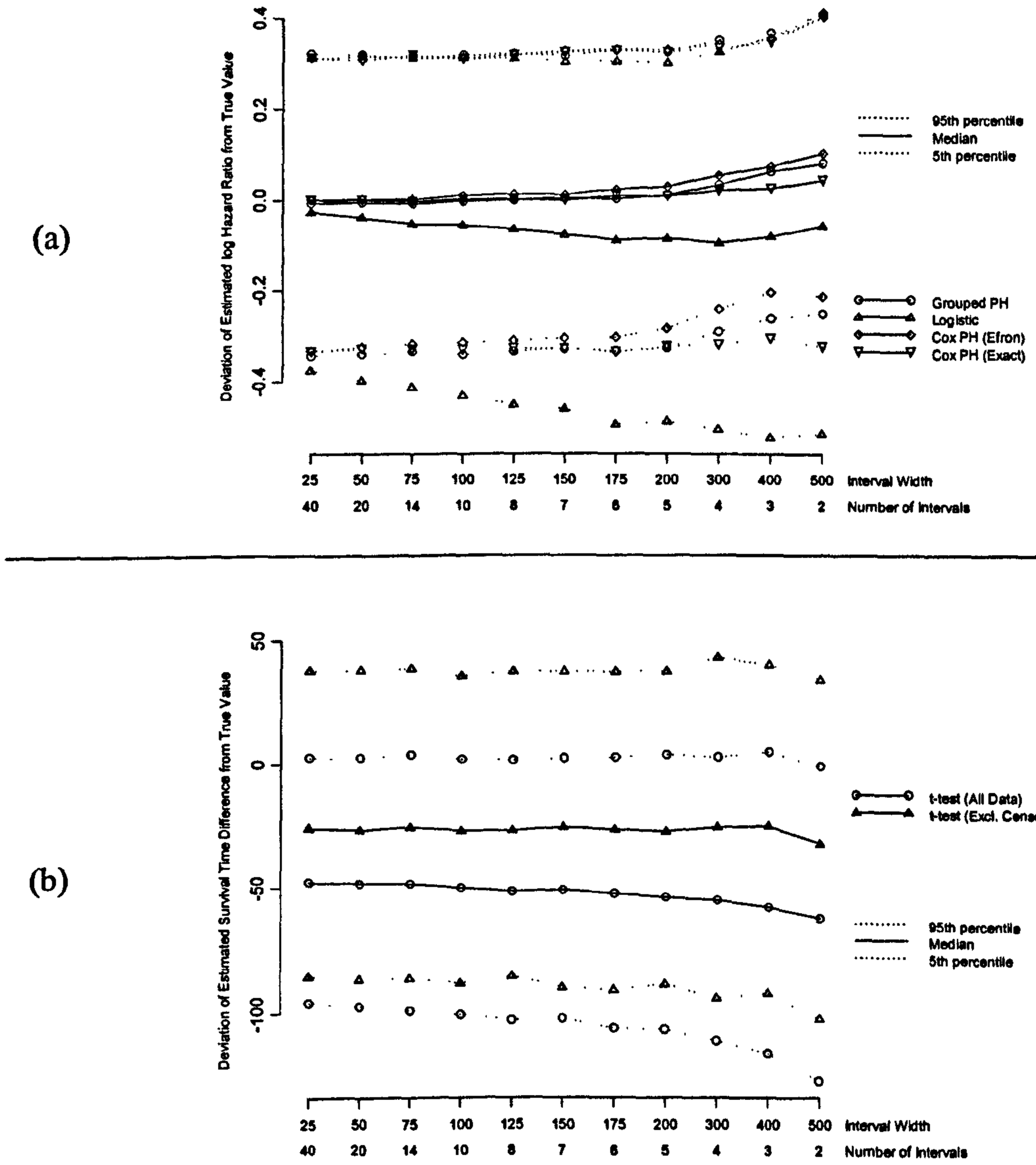
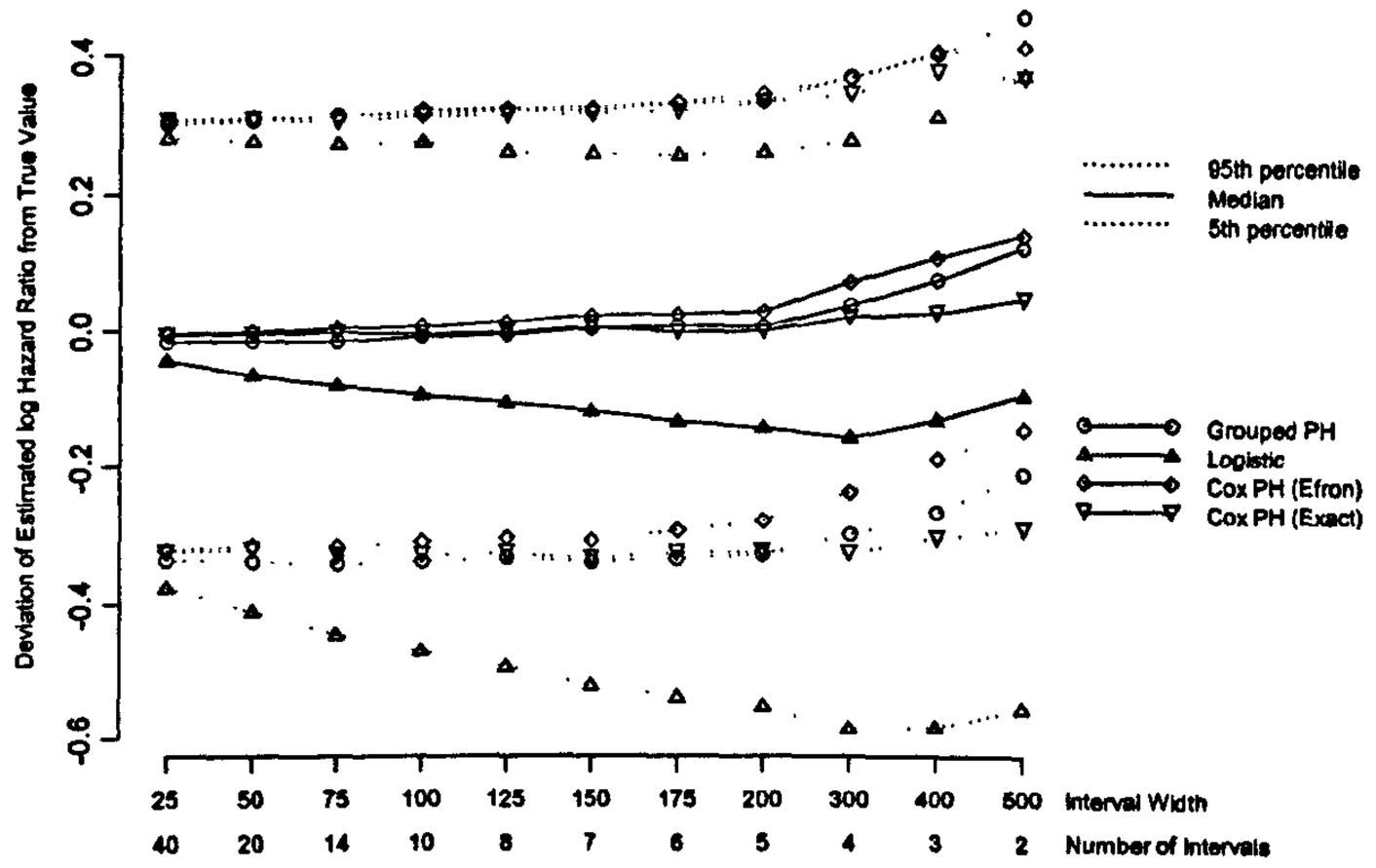


Figure 6.3 Median deviations of effect estimates from target values, with 5th and 95th percentiles found in simulated studies with a hazard ratio between groups of 0.67, using (a) survival analysis methods and (b) t-test methods

approaches the log hazard ratio once more. The three PH models remain relatively unbiased down to as little as five or six intervals, after which they become biased towards zero, with the Cox PH model using the exact partial likelihood being least biased, and the Cox model with the Efron approximation being most biased.

Both t-test methods are negatively biased, underestimating the true difference in mean survival times between groups. The model using all data shows greater underestimation, since it includes observations in both groups that were censored; censoring times were assumed to have the same distribution in each group, so their inclusion will shrink estimates of the true difference towards zero. The extent of the bias

(a)



(b)

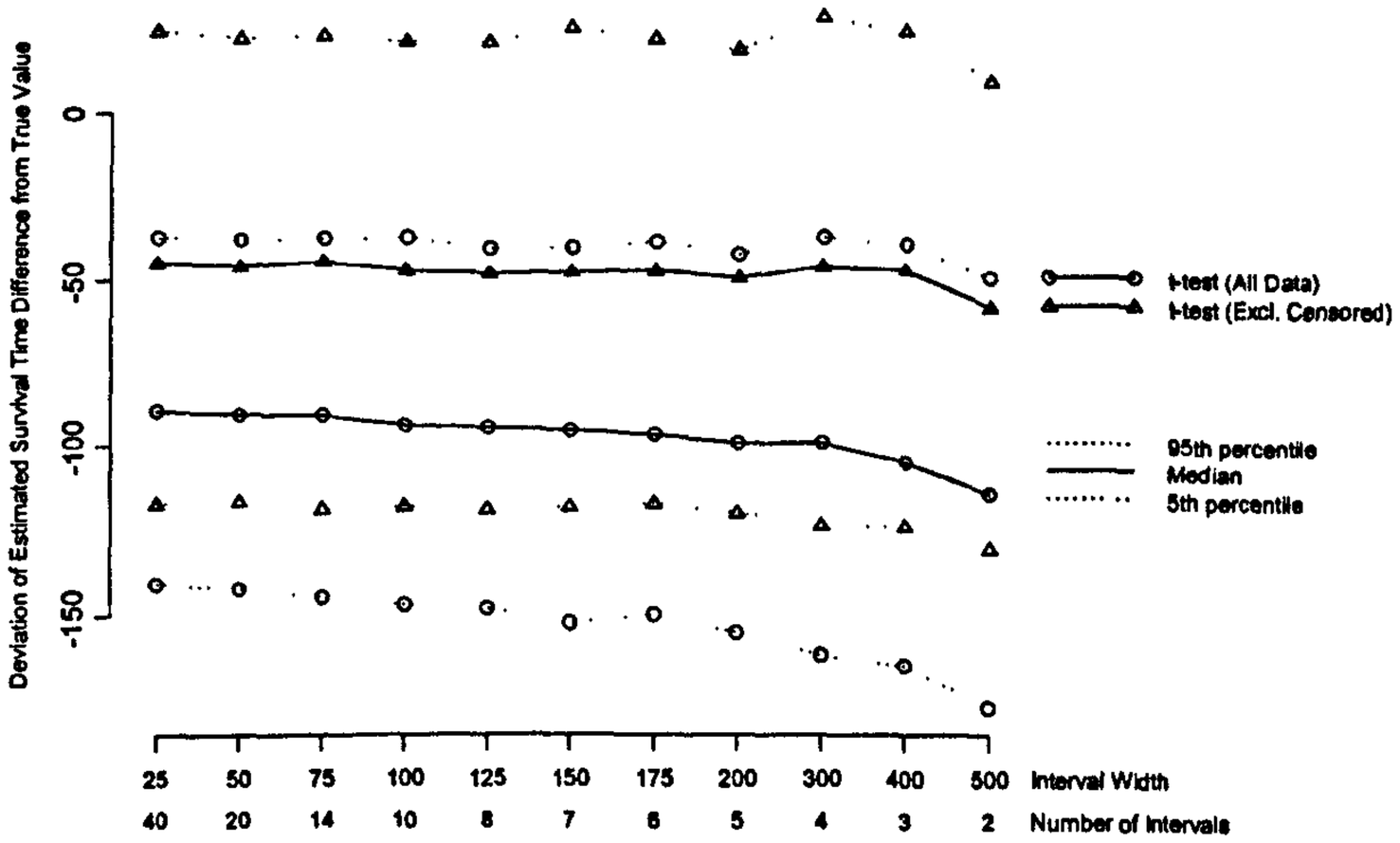


Figure 6.4 Median deviation of effect estimates from target values with 5th and 95th percentiles found in simulated studies with hazard ratio between groups of 0.5, using (a) survival analysis methods and (b) t-test methods

under both approaches does not vary much as interval widths increase, becoming slightly larger in the most extreme cases.

Figure 6.4 shows the same result from simulations with a hazard ratio of 0.50 between groups. Similar patterns are seen, with the Cox PH model using the exact partial likelihood again showing the least bias of the PH models, though none are severely biased except when there are few intervals. The t-test methods are very biased, particularly when censored observations are included, though when these are excluded, there is a greater variability in treatment effect differences over the 1000 simulations.

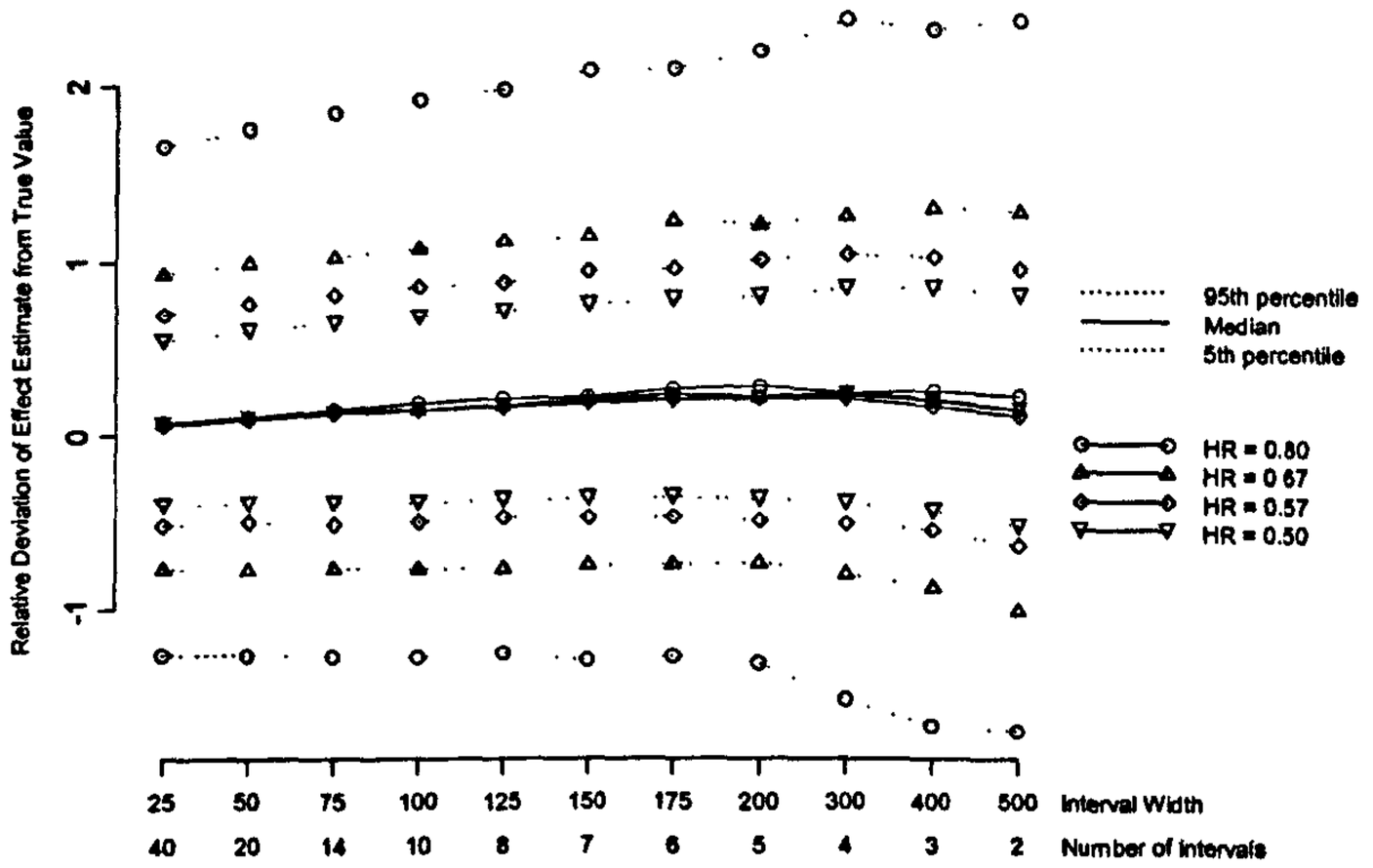


Figure 6.5 Median deviation of effect estimates from true log hazard ratio, with 5th and 95th percentiles found in simulated studies using the logistic model, under a range of between-group hazard ratios

To allow more direct comparisons between the levels of bias found with different simulated treatment effects, deviations of effect estimates from their target values can be expressed relative to the true log hazard ratio or mean survival time differences. Figure 6.5 shows the relative deviation of the logistic model estimates, for which the bias of the log odds ratios estimated for the true log hazard ratio appeared to become more severe with increasing simulated treatment effect in Figure 6.4(a). However, the shapes of the graphs of median relative deviations against interval width are very similar under the differing treatment effects. Between-simulation variability in terms of relative deviation is smaller for larger treatment effects, since absolute levels of variation do not depend greatly on treatment effects.

6.3.2 Error Rates and Power

The Type I error rate of a statistical test is defined as the probability that the null hypothesis will be "rejected" when it is in fact true. This error rate will depend upon the level of statistical significance used to define evidence against the null hypothesis; the traditional value of 5% might be used. If a test performs well, then the Type I error rate will be the same as the significance level, though a conservative test would have an error rate below the level of significance. It would be expected that the Type I error rate would be no more than the level of significance used to perform the test; otherwise an investigator would falsely consider there to be differences between groups more often than would be liked.

No. of Intervals	Interval Width	Model					
		Grouped PH	Logistic	Cox PH (Efron)	Cox PH (Exact)	t-test (All Data)	t-test (Excl. Censored)
40	25	6.5%	6.8%	6.2%	6.4%	6.3%	6.5%
20	50	6.5%	7.0%	6.6%	6.8%	6.7%	6.4%
14	75	6.2%	6.7%	5.9%	6.0%	6.2%	6.2%
10	100	6.8%	6.9%	6.4%	6.8%	6.5%	5.9%
8	125	6.4%	6.8%	5.7%	5.9%	6.6%	5.4%
7	150	5.6%	6.3%	5.7%	6.4%	6.5%	5.9%
6	175	7.0%	7.0%	6.0%	6.5%	7.4%	5.8%
5	200	6.3%	5.8%	5.5%	6.1%	6.5%	5.6%
4	300	5.9%	6.0%	3.8%	4.9%	6.3%	4.6%
3	400	4.9%	5.7%	3.8%	5.5%	5.3%	5.2%
2	500	5.6%	5.4%	3.6%	5.4%	6.1%	5.2%

Table 6.1 Estimated Type I error rates for each model, estimated from 1000 simulated studies, under different levels of interval censoring

Table 6.1 shows the estimated Type I error rates of each model in this simulation study, at every level of interval censoring considered. For a single binomial significance test using a sample of 1000, values of 3.4% or below, or 6.4% or above would indicate evidence that the true proportion is not equal to 5%. Since the same simulated datasets are being analysed by different models under differing levels of interval censoring, these estimates will be inter-related. However, the cut-off of 6.4% is a useful guide to whether the tests are being too liberal. Each of the methods demonstrates some degree of poor performance, though the t-test after excluding censored observations appears to be the least likely to suffer a Type I error. As previously observed, this method provides estimates that are severely biased and the Cox PH models would appear to offer the best alternative in terms of Type I error rates, though the Efron approximation to the partial likelihood may be slightly conservative when the data are extremely interval censored.

Whereas the Type I error rate is the probability of falsely rejecting a null hypothesis, the Type II error rate is the probability of failing to reject the null hypothesis when it is in fact false. This will depend on the significance level being used in the statistical test as well as the true effect size and the sample size. In this case, the Type II error rate will be dependent upon the true treatment effect, though the influence of sample size cannot be examined as this was not varied in this study.

Hazard Ratio	Interval Width	Model					
		Grouped PH	Logistic	Cox PH (Efron)	Cox PH (Exact)	t-test (All Data)	t-test (Excl. Censored)
0.80	50	23.9%	23.9%	23.5%	23.8%	14.5%	16.8%
	100	24.6%	23.8%	23.0%	23.9%	13.1%	16.6%
	200	22.6%	23.0%	21.4%	22.3%	11.2%	15.6%
0.67	50	57.7%	58.0%	57.2%	57.3%	27.7%	41.0%
	100	57.3%	57.0%	56.1%	56.6%	25.3%	40.3%
	200	55.0%	54.4%	52.4%	54.3%	21.9%	39.6%
0.57	50	84.5%	84.0%	83.5%	83.7%	48.1%	63.3%
	100	84.2%	83.5%	82.9%	83.2%	43.0%	61.8%
	200	82.8%	82.2%	81.4%	82.1%	34.5%	59.7%
0.50	50	96.0%	95.8%	95.2%	95.2%	65.5%	82.7%
	100	94.9%	94.7%	95.0%	95.2%	59.7%	81.7%
	200	93.7%	93.1%	93.1%	93.8%	49.2%	78.8%

Table 6.2 Power (%) of statistical models to detect treatment effects, simulated as constant hazard ratios of 0.80, 0.67, 0.57 and 0.50, for selected levels of interval censoring (interval widths of 50, 100 and 200 units)

Type II error rates are often described in terms of the power of a statistical test, defined as 1-β, where β is the Type II error rate. Power is lowest, or Type II error rates are highest, when effects are small or the sample size is small. Power increases with larger effects and larger samples; this is intuitively obvious, since the chances of detecting a difference will clearly be greatest when the effect is large or there are large amounts of data.

Table 6.2 shows the estimated power of the six models studied, for each treatment effect other than the no effect scenario (hazard ratio = 1), for three levels of interval censoring, chosen as interval widths of 50, 100 and 200, which correspond to 20, 10 or 5 intervals respectively. Power is greater when the treatment effect is larger, and though usually greater when the degree of interval censoring is lower, the dependence on censoring is small.

The t-test methods clearly have much lower power than the survival analysis methods, and for every combination of treatment effect and interval width, excluding the censored observations leads to a more powerful test. Amongst the survival methods, power levels are very similar, with none having consistently better performance.

No. of Intervals	Interval Width	Model					
		Grouped PH	Logistic	Cox PH (Efron)	Cox PH (Exact)	t-test (All Data)	t-test (Excl. Censored)
40	25	2	1	4	3	6	5
20	50	1	2	4	3	6	5
14	75	1	2	4	3	6	5
10	100	1	3	4	2	6	5
8	125	1½	3	4	1½	6	5
7	150	1	3	4	2	6	5
6	175	1½	1½	4	3	6	5
5	200	1	2	4	3	6	5
4	300	3	2	4	1	6	5
3	400	3	2	4	1	6	5
2	500	3½	2	3½	1	6	5

Table 6.3 Ranking of the six methods in terms of average power over the range of effect sizes simulated within each level of interval censoring (1=most powerful, 6=least powerful)

To investigate more closely the method with the best power under a range of situations, the six models were ranked within each treatment effect × interval width combination with the most powerful method being given the rank 1 and the least powerful the rank 6. The ranks for each method × interval width combination were then averaged over the four treatment effect sizes, and finally, for clarity the six methods were ranked within each interval width.

These ranks are shown in Table 6.3. The least powerful methods, in order of increasing power are the t-test applied to all data, the t-test with censored observations excluded and the Cox PH model using the Efron approximation to the partial likelihood. Of the remaining three methods, the exact Cox PH model appears to be least powerful when there are a large number of intervals, but most powerful when there are few intervals.

The better performance of the exact Cox model when there are a few large intervals is likely to be due to the way failure times were simulated to have been recorded. In general, simulated failure times were rounded up to the time of the end of the interval within which it fell, so that the Cox PH models and the two grouped survival data methods were analysing essentially the same data. For observations such

as these, the approximate Cox PH model would not perform as well as the exact Cox PH model or the grouped survival methods, since it approximates the partial likelihood. When the failure time and censoring time were both simulated to have occurred in the same interval, if the failure time was less than the censoring time, the grouped survival models were using only the fact that a failure occurred within that interval, whereas the Cox PH models used the observed censoring time as the first time when the observation was known to have failed.

When there are many short intervals, there would be few of these observations, since it would be quite rare for both the failure and censoring times to fall within the same interval. When there are few large intervals, such occurrences would be more common, and the exact Cox PH model would gain by using observed failure times closer to the true failure times than the grouped survival models, which would be forced to use only the information that a failure fell within some wide interval. This could also explain the better performance of the exact Cox PH model in terms of bias for large intervals.

6.4 Summary

The two grouped survival analysis methods appear to perform better than the other methods over the range of situations representative of most exercise tests, that is a maximum of 10 to 20 intervals of observation. The two Cox PH models are nearly as good in terms of power, and equally effective in terms of bias and coverage, and would be more straightforward to apply in practice, since there is no need to separate the observation period of each individual into distinct intervals. The ANOVA methods are clearly less adequate in every respect other than coverage, and cannot be recommended based on these results.

The most important decisions about how to analyse such data would most likely be based on goodness-of-fit assessments, particularly in terms of the proportional hazards assumption (for the PH models). It has not been assessed in these simulations how the different methods perform when the data do not satisfy a PH assumption, and the robustness of the various methods when this assumption is not met needs to be investigated further.

CHAPTER 7 Repeated Survival Times

One of the features of exercise test data is that a test can be repeated on the same subject. In a clinical trial, a positive exercise test may be one of the inclusion criteria for candidate participants[53], and an off-treatment test may be one of the baseline measurements. Once a subject has been included in a trial, any number of tests can be performed during the follow up of each subject, whether for monitoring of the patient and adjustment of treatment, or for use in the final analysis[52]. However, it will often be the case that an off-treatment, baseline exercise test will be compared to an end-of-study test to evaluate treatment effects.

Another setting in which repeated exercise tests occur is in crossover trials of angina treatments. Since angina is a chronic condition, without the possibility for cure by drug treatment, it is a suitable condition for a crossover study[50]. Such studies include repeated exercise tests by design, and may include tests conducted at the start of each treatment period, after a suitable washout period, as a baseline for each end-of-treatment test.

When trials are designed to include several exercise tests the resulting data should be analysed accordingly. If two or more exercise times have come from the same subject, it would be expected that the exercise times produced will be correlated, and this should be accounted for in the analysis, if at all possible. By taking account of the information from more than one test, it would be hoped that some of the within-subject variation in exercise performance can be removed and hence a better understanding of between-subject differences can be achieved. This is particularly important in a crossover trial, since the purpose of the design is to reduce within-subject variation.

Repeated survival data can be analysed in a number of ways. Standard (but inappropriate) statistical methods such as paired t-tests and repeated measures ANOVA can been used, but these ignore the censored nature of the data, and will not be considered in detail here. This chapter will consider methods for repeated observations where the data are (possibly censored) survival times.

Non-parametric methods, similar to the log-rank test[96], have been developed for the comparison of paired survival times, and some of these will be described. However, these methods do not allow for covariate adjustment, which is often desired of an analysis. Analogous to the paired t-test, a parametric form may be applied to the difference between two survival times, and one such approach will be presented. Finally, methods will be considered that seek to extend survival regression techniques to multivariate failure time data. These methods can be divided into two strands. Marginal models, in which models for the distribution of each survival time are fitted, whilst adjusting variance estimates of the resultant parameters to account for the effect of correlation between survival times, treat this correlation as a nuisance. Frailty models allow the analysis of correlated survival data through a random effects model, in which survival times become correlated through the sharing of unobserved covariates, which can be thought of as an unknown susceptibility to failure. Some of these models will be applied to data from the TIBET Study.

7.1 Treatment Preference

One area in which repeated exercise tests arise is in crossover trials of anti-anginal therapies[104]. For a two-treatment, two-period crossover trial, the standard proportional hazards model, $\lambda_i(t)=\lambda_0(t)\exp(\mathbf{z}_i\beta)$, can be applied, where $\mathbf{z}_i$ includes indicator variables to represent treatments and periods. The likelihood, however, reduces to a function of the numbers of individuals in each order of treatment allocation that have a preference for one treatment or the other.

These methods are not immediately applicable to a parallel groups study in which baseline and on-treatment exercise tests are to be analysed, since each patient receives exactly one treatment, and their preference for one treatment over any other cannot be determined. To extend the methods would result in comparing the numbers of patients that prefer treatment to no treatment between different treatment groups, which would result in a between-, rather than a within-subject comparison. Since it is likely that the majority would perform better under an exercise test whilst on treatment in all groups, this approach would lack power to detect treatment effects in study designs other than a crossover.

7.2 Paired Rank Tests

Numerous approaches[105,106,107,108,109] have been developed based on rank testing procedures that test for differences between correlated survival times. In a simulation study[110], the paired Prentice-Wilcoxon (PPW) test[111] and the test devised by Akritas[112] were found to perform consistently better than other tests considered.

The PPW test derives a score for each observation based on the rank of the survival time amongst those from the whole dataset, using both observations from each pair. The Akritas test is similar, but uses the Kaplan-Meier survival probability of each observation amongst the whole dataset as the basis for the score. In both settings, the scores for censored observations are penalised to adjust for the fact that the actual survival time is larger than that observed.

7.2.1 Paired Prentice-Wilcoxon Test

Calculation of the paired Prentice-Wilcoxon (PPW) test statistic[111] requires that all observed failure and censoring times, that is both members of each pair, be considered as a single set of data. Let D be the number of distinct failure times, and define n_j to be the number of observation times greater than or equal to the j^{th} ordered failure time, for $j = 1, 2, \ldots, D$. Then, for $i = 1, 2, \ldots, D$, define

$$s_i = \prod_{j=1}^{i} \frac{n_j}{n_j + 1}.$$

If failure times are tied, those subjects seen to fail at a particular time are assigned distinct failure times slightly smaller than the observed time. For example, if there are k observed failure times equal to $t_{(i)}$, the i^{th} ordered failure time, then these k times are changed to $t_{(i)} - \varepsilon < t_{(i1)} < t_{(i2)} < \ldots < t_{(ik)} < t_{(i)}$, for suitably small ε. The allocation of these new times is arbitrary. The quantities, s_i, are then calculated as above, and those values corresponding to $t_{(i1)}$, $t_{(i2)}$, ... $t_{(ik)}$ are averaged to produce s_i for the i^{th} ordered failure time.

Each subject seen to fail at the i^{th} ordered failure time is assigned a Prentice-Wilcoxon score of $1-2s_i$, whereas each subject who is censored at a time as least as large as the i^{th}, but less than the $(i+1)^{th}$ ordered failure time is given a score of $1-s_i$. Then, for each pair of observations, calculate Δ_i as the difference between the Prentice-Wilcoxon scores for the i^{th} pair. The PPW test statistic is then defined to be

$$Z_{PPW} = \sum_{i=1}^{n} \Delta_i \Big/ \left(\sum_{i=1}^{n} \Delta_i^{\,2} \right)^{1/2} \qquad \textbf{(Eq. 7.1)}$$

which, under the null hypothesis that there is no difference between the survival distributions of each pair member, converges in distribution to a standard normal, to which the statistic should be compared to assess the evidence against the null hypothesis.

Alternatively, the statistic Z_{PPW} (Eq. 7.1) could be recalculated for all 2^n permutations of survival times within pairs (or a random sample thereof, if n is large). Under the null hypothesis, each value of the statistic obtained in this way is equally likely, and if the value corresponding to the observed data should be extreme within the permutation sample, this would constitute evidence against the null hypothesis.

7.2.2 Akritas Test

An alternative to the PPW test, which gives similar results but is simpler to evaluate, is the Akritas test[112]. Its derivation hinges on the definition of the ranks of observed and censored failure times. If S(t) is the Kaplan-Meier estimator of the survival function on the set $\{ T_i, \delta_i; i = 1, 2, ..., n \}$, then the rank of an individual seen to fail at time t is defined to be nS(t). The rank of an individual censored at time t is defined to be n[½+½S(t)]; the true failure time is known to lie between t and +∞, so the average of the Kaplan-Meier estimator at these two time points, S(t) and 1, is used to define the rank.

The paired test is carried out by computing Kaplan-Meier estimates of the survival function for the first and second pair members, $S_1(t)$ and $S_2(t)$. The average survival function is then defined as

$$\bar{S}(t) = \frac{1}{2}\{S_1(t) + S_2(t)\}.$$

The ranks of all observations are then calculated using the function $\bar{S}(t)$, and the test statistic is that of a paired t-test applied to these ranks.

Example 7.1 Akritas Test for Time to Anginal Pain

Table 7.1 shows the mean difference in ranks, as defined for the Akritas test, between the third exercise test, taken after six weeks of treatment, and the first, taken after a two-week washout period before study treatments were given. All six mean

Number of Patients Mean (SD) Rank Difference		Treatment		
		Atenolol	Nifedipine	Combination
Exercise Type	Treadmill	113 68.4 (226.6)	110 12.5 (214.9)	107 47.8 (214.1)
	Bicycle	104 95.7 (200.2)	106 66.2 (212.4)	104 104.6 (193.5)

Table 7.1 Numbers of patients, with mean and standard deviation (SD) of the difference in rank between the third and first exercise tests, calculated from the time to anginal pain by exercise type and study treatment

values are positive, indicating that patients tended to exercise for longer on the later test. The values for those using a bicycle are larger than for those using a treadmill, indicating a greater increase in exercise time. This may be artefactual, since ranks were calculated on a pooled dataset, including both treadmill and bicycle exercise times at both time points; the greater increase in ranks for the bicycle data may simply reflect the shorter timescale of treadmill exercise tests, rather than a greater improvement in exercise times per se.

Table 7.2 shows the results of fitting simple linear regression models to these data. Two models are shown, both controlling for the effects of the type of exercise and estimating the effects of treatment. Age and weight were found to have no significant effects in this model, though gender is seen to have some effect, for those using a treadmill, at least, with women showing a greater increase in exercise times. There is some evidence that treatment with Atenolol, either alone or in combination with Nifedipine, is associated with a greater increase in the ranks of exercise times than treatment with Nifedipine alone.

7.3 Model for the Difference in Survival Times

The t-test is a powerful tool for analysing paired continuous data; one of its advantages is that it performs well so long as the differences between pair members are roughly Normally distributed, regardless of the distributions of the actual observations in each pair. Pair differences are, in practice, often approximately Normally distributed. This view could be extended to look at paired survival data, and assume that the difference between two survival times will follow a Normal distribution.

Let the data be denoted by $\{(T_{i1}, \delta_{i1}), (T_{i2}, \delta_{i2}); i = 1, 2, ..., n\}$; $T_{ij}=\min(D_{ij}, C_{ij})$, where D_{ij} is the failure time and C_{ij} the censoring time (assumed to be non-informative

	Estimate (95% CI)	p	Estimate (95% CI)	p
Intercepts				
Treadmill	61.2 (29.1,93.2)	0.0002	50.2 (16.7,83.6)	0.0034
Bicycle	106.4 (73.9,138.8)	<0.0001	105.4 (72.0,138.9)	<0.0001
Treatment Effects				
Nifedipine – Atenolol	-43.4 (-82.8,-4.0)	0.031	-43.1 (-82.6,-3.6)	0.033
Combination – Atenolol	-8.0 (-47.6,31.6)	0.69	-6.8 (-46.3,32.8)	0.74
Gender Effects (Female – Male)				
Treadmill			74.4 (9.7,139.1)	0.025
Bicycle			2.6 (-62.0,67.3)	0.94

Table 7.2 Effect estimates from linear regression models for changes in ranks of exercise times to anginal pain between first and third exercise tests

for D_{ij}) for the i^{th} individual on the j^{th} occasion, and δ_{ij} is the indicator variable, $I(D_{ij} \leq C_{ij})$, equal to one if a failure is observed, or zero if the time is censored. Let $\Delta_i=D_{i2}-D_{i1}$, and assume that

$$\Delta_i \sim N(\mu_i, \sigma^2). \qquad \textbf{(Eq. 7.2)}$$

From this basic model, a natural extension to allow the estimation of treatment and covariate effects is to parameterise μ_i as $z_i\beta$. In this case, (Eq. 7.2) can be written as

$$\frac{\Delta_i - z_i\beta}{\sigma} \sim N(0,1).$$

Regarding the observed data, there are four situations to consider, corresponding to whether δ_{i1} and δ_{i2} are 0 or 1; that is, whether the individual survival times are censored or not. If $\delta_{i1} = \delta_{i2} = 1$, then $T_{i2} - T_{i1} = \Delta_i$. If $\delta_{i1} = \delta_{i2} = 0$, then there is no information regarding the value of Δ_i, and the i^{th} pair are effectively useless. If $\delta_{i1} = 1$ and $\delta_{i2} = 0$, so that T_{i1} is observed exactly whilst T_{i2} is censored, then $T_{i1} - T_{i2}$ is right censored, i.e. $\Delta_i > T_{i2} - T_{i1}$. If $\delta_{i1} = 1$ and $\delta_{i2} = 0$, then $T_{i1} = T_{i2}$ is censored on the left; $\Delta_i < T_{i2} - T_{i1}$.

	Mean Difference (secs)	95% CI	p
Treatment (Nifedipine – Atenolol)	-44.6	-103.1, 13.9	0.13
Treatment (Combination – Atenolol)	24.5	-34.4, 83.4	0.42
Age (/10 years)	7.9	-23.0, 38.9	0.62
Gender (Female – Male)	-15.1	-92.1, 61.9	0.70
Weight (/10 kg)	14.7	-9.7, 39.0	0.24

Table 7.3 Treatment and covariate effect estimates, with 95% CIs and p-values from a model assuming that differences in exercise times to anginal pain are Normally distributed

If $\phi(z)$ and $\Phi(z)$ are the probability density and cumulative distribution functions of the standard normal distribution, and if $u_i = \frac{[T_{i2} - T_{i1}] - \mathbf{z}_i\beta}{\sigma}$, then we can write the likelihood of the data as

$$L(\beta,\sigma|\Delta_i,\delta_{i1},\delta_{i2}) = \prod_i f(\Delta_i)^{\delta_{i1}\delta_{i2}} \{1-F(\Delta_i)\}^{\delta_{i1}(1-\delta_{i2})} F(\Delta_i)^{(1-\delta_{i1})\delta_{i2}} \quad \textbf{(Eq. 7.3)}$$

$$= \prod_i \left\{\frac{1}{\sigma}\phi(u_i)\right\}^{\delta_{i1}\delta_{i2}} \{1-\Phi(u_i)\}^{\delta_{i1}(1-\delta_{i2})} \Phi(u_i)^{(1-\delta_{i1})\delta_{i2}}$$

so that doubly-censored pairs, where $\delta_{i1} = \delta_{i2} = 0$, do not contribute to the likelihood. Maximum likelihood estimates of the parameters β and σ can be found, along with their standard errors from the second derivative of minus the log likelihood evaluated at $\hat{\beta}$ and $\hat{\sigma}$. The log likelihood, together with first and second derivatives, are given in Appendix A.

Example 7.2 Normal Assumption for Difference in Survival Times, Time to Anginal Pain

Table 7.3 shows the effect estimates, with 95% CIs and p-values from this model applied to data from the TIBET Study. There is no evidence that either treatments or other covariates influence the improvement of exercise times between visits 3 and 5.

7.4 Models for Correlated Survival Times

Whilst modelling the difference between correlated survival times is conceptually straightforward, giving a summary of effects in terms of increases in survival time, it

relies on assumptions about the distribution of these differences, and any pairs of survival times that are doubly censored provide no information to the model and are lost from the analysis. Furthermore, this approach is applicable to analysing pairs (in particular ordered pairs) of survival times only; larger clusters of survival times could only be analysed by multiple pairwise comparisons. However, more general methods for the analysis of clusters of arbitrary size, or for clusters of times with no inherent order might be preferred.

To analyse arbitrary clusters of correlated data with a general linear model, there are essentially two methods that can be adopted. The generalized estimating equations (GEE) approach[113] models the marginal (population-averaged) effects of independent variables on outcomes, whilst simultaneously modelling the correlation between members of the same cluster. In general, this results in inflation of the variance of the parameter estimates according to the correlation between outcomes within clusters. The correlations themselves are treated as nuisance parameters and are not of primary interest. An alternative to GEE is mixed effects modelling[114], in which the effects of covariates can be modelled as fixed effects and are interpreted in the same way as a general linear model, though from the point of view of effects within clusters; correlation within clusters is accounted for by assuming that study units within clusters share the same value of some unobserved random variable. With some restrictions and assumptions about the distribution of these random effects (e.g. zero mean Normal distribution), the analysis focuses on estimation of distributional parameters such as the variance. This allows the separation of variability in the population as a whole into variation between clusters and variation within clusters, i.e. between individuals.

These two schools of thought have also surfaced in the analysis of clustered survival data, motivated by problems in matched case-control studies, litter studies and family studies in humans.

7.4.1 Marginal Models

In the analysis of clustered survival data, marginal models can be applied by assuming a population-averaged model, such as the Cox proportional hazards model[83]. For a sample of independent observations, $\{T_i, \delta_i: i=1, 2, ..., n\}$ the Cox model would normally be expressed in terms of the hazard function,

$$\lambda_i(t) = \lambda_0(t)\exp(\mathbf{z}_i\beta) \qquad \textbf{(Eq. 7.4)}$$

though from this specification, the model survivor function $S_i(t)$ is fully determined; considered as a random variable, $S_i(T_i)$ has a uniform distribution on the interval (0, 1).

For a clustered sample of observations, $\{T_{ij}, \delta_{ij}: i=1, 2, ..., n; j=1, 2, ..., n_i\}$, the same model will apply in the population, such as (Eq. 7.4). Each $S_{ij}(T_{ij})$ has a uniform distribution on (0, 1), though there is a degree of dependence between observations within clusters. This can be parameterised by a class of distributions known as copulas[115], so that $\{S_{ij}(T_{ij}): i=1, 2, ..., n_i\}$ is an n_i-variate distribution with uniform marginal distributions on (0, 1).

In the bivariate case, such as the analysis of the time to anginal pain during the first and third exercise tests in the TIBET Study, a pseudo-likelihood approach[116] can be employed. If the marginal distribution is correctly specified, then an independence working model, whereby the marginal model is fitted to the data assuming that pairs of survival times are independent, will give parameter estimates, $\hat{\beta}$ that are consistent[117]. The variance of $\hat{\beta}$ will not be consistent; however a robust variance estimate for $\hat{\beta}$ can be calculated as

$$V^*(\hat{\beta}) = V(\hat{\beta})\left\{\sum_j U_j(\hat{\beta})U_j(\hat{\beta})'\right\}V(\hat{\beta})$$

where $V(\hat{\beta})$ is the usual variance estimator from the independence model, derived as the inverse of the observed information matrix, and $U_j(\hat{\beta})$ is the contribution of the j^{th} pair of observations to the score vector.

Example 7.3 Marginal Regression Model for Time to Anginal Pain

Table 7.4 shows the estimated hazard ratios from Cox proportional hazards models for the time to anginal pain in the TIBET Study. Each model includes separate baseline hazard functions for each exercise type, and includes terms for the effects of treatment, gender, age and weight. The first model uses data from the third exercise test only; this was the principal exercise test within the study. The two further models also incorporate data from the first exercise test, taken prior to the start of study treatment, after a washout period without active treatment. The baseline hazard functions with these models are stratified by occasion as well as exercise type. An alternative approach would be to include test occasion as a binary predictor variable; in this instance the models reach similar conclusions in each case.

	3rd Test Data Only		1st and 3rd Test Data Independence model		1st and 3rd Test Data Marginal model	
	Haz. Ratio (95% CI)	p	Haz. Ratio (95% CI)	p	Haz. Ratio (95% CI)	p
Nifedipine-Atenolol	1.33 (0.94,1.87)	0.11	1.33 (0.94,1.87)	0.11	1.33 (0.94,1.87)	0.11
Combination-Atenolol	1.06 (0.74,1.52)	0.76	1.06 (0.74,1.52)	0.76	1.06 (0.74,1.53)	0.76
Gender (Female – Male)	1.21 (0.76,1.94)	0.43	1.35 (1.03,1.77)	0.032	1.35 (0.95,1.91)	0.095
Age (/10 years)	1.19 (0.97,1.45)	0.089	1.26 (1.12,1.42)	0.0001	1.26 (1.09,1.46)	0.0022
Weight (/10 kg)	0.96 (0.83,1.12)	0.60	0.99 (0.90,1.08)	0.74	0.99 (0.88,1.11)	0.80

Table 7.4 Effect estimates (as hazard ratios, with 95% CIs and p-values) from Cox proportional hazards models for the time to anginal pain, stratified by exercise type and dependent upon treatment, gender, age and weight

All effect estimates from the latter two models are identical, since they are estimated in exactly the same way. Treatment effect estimates are also the same under the model based on data from the third exercise test only, because the first test was carried out off treatment, and including these data adds no information to inferences about treatment effect differences; this is also apparent from the fact that treatment effect confidence intervals are the same when the additional data are included.

Uncertainty in the covariate effect estimates is reduced by the addition of data from the first test, though confidence intervals around the estimates under the marginal model are wider than under the independence model, as the correlation between pairs of observations is accounted for. The increase in variation around the treatment effect estimates is negligible, and since the first exercise test contains no information regarding treatment effects, the use of the additional data in this instance would seem unnecessary. However, covariate effects are estimated with greater precision under the marginal model using data from two occasions, than under the model based on data from a single test, demonstrating the potential advantages to using these methods in studies where multiple on-treatment tests are administered, such as crossover trials.

7.4.2 Mixed Effects Models

Mixed effects analysis of survival data assumes that different individuals have a different susceptibility (or frailty) to the event of interest. This heterogeneity is

modelled, not as effects of covariates that can be measured, but as the combined effects of unmeasured and unmeasurable factors. These methods have been applied in the analysis of survival data from animals within litters[118], family members[119] and matched case control studies, as well as repeated exercise test data[120]. This section will describe how to fit these models.

7.4.2.1 Frailty Distributions

Frailties can be thought of as unmeasurable information, such as the genetic susceptibility to disease shared by members of the same family, or the motivation of an individual to exercise. They can also be thought of as unmeasured information, such as shared environmental exposures experienced by siblings during childhood, or variables that might influence exercise capacity but are not recorded, such as percentage body fat or normal daily exercise levels.

In general, after allowing for the effects of those factors that are measured, the remaining heterogeneity between individuals could be due to factors of both kinds. Over the population as a whole, there will be a distribution of frailties. The methods considered here will assume that the unobserved frailties come from a continuous distribution. There are situations where a discrete distribution would suffice; for example we might assume that the population consists of those with or without coronary artery disease. Alternatively, a mixed distribution might be appropriate, for example in a population made up of some immune and some susceptible subjects, with those susceptible having a continuous distribution of frailties.

7.4.2.2 Incorporating Frailty Effects

Fitting regression models to survival data involves making a number of assumptions. For example, we may assume a proportional hazards model, or choose a particular distribution for the true failure times. Fitting frailty models involves making further assumptions about how the frailty factors influence survival times, and what the distribution of the factors is in the population under study. In practice, these decisions are often based on convenience, to make the process of fitting the model simpler.

The most common regression model for survival data assumes proportional hazards. A parsimonious extension would be to include frailty factors as acting multiplicatively on the hazard function. That is, an individual with frailty factor ω and covariates $\mathbf{z}$ has a hazard function of

$$\lambda(t|\omega, z) = \omega\lambda_0(t)\exp(z\beta), \qquad \textbf{(Eq. 7.5)}$$

where $\omega > 0$ has some distribution over the population under investigation. To make the parameters identifiable, the distribution of ω has some fixed mean, usually 1. Thus the "average" hazard is set by the baseline hazard function, $\lambda_0(t)$, and the spread in the distribution of ω describes the degree of heterogeneity in the population. If the variance of ω is large, then multiple survival times will show a large degree of correlation within individuals or clusters. If the variance of ω is small, there will be little correlation within groups, meaning that multiple survival times are essentially independent, after allowing for the effects of measured covariates.

The effects of measured covariates are modelled through β, viewed as log hazard ratios between members of the same cluster. At the population level, hazard ratios associated with covariate differences are not necessarily constant over time[121]. For example, consider a population made up of families, within which individuals share the same susceptibility for some disease, even though these susceptibilities have some distribution across different families. Now assume that men have a greater hazard for the disease than women, and that within any single family, the hazard for men is a constant multiple of that for women. Over time, more men will contract the disease, with those that do being more likely to come from families with high susceptibility. As a result, the average susceptibility amongst the total population of men that remains disease free will progressively decrease, more so than amongst women, and the hazard ratio between the sexes over the surviving population will fall.

7.4.2.3 EM Algorithm

One method of fitting a frailty model is to consider the problem to be one of incomplete data, and to use the EM algorithm[122]. The frailties common to each individual (in the context of repeated exercise tests) are viewed as unobserved variables. These variables are assumed to come from some specified distribution with a known mean of 1 and an unknown variance, θ.

This method can be applied to a proportional hazards model where the baseline hazard function is defined parametrically or semi-parametrically. The notation $\lambda_0(t)$ and $\Lambda_0(t)$ shall be used to denote the baseline hazard and cumulative hazard functions at time t, which will depend upon a vector of parameters. In the case of a semi-parametric

model, these parameters will be the set of baseline cumulative hazard increments at the distinct failure times.

Assuming that the density of the frailty factors can be written as $g(\omega|\theta)$, then the likelihood of the data $\{t_{ij}, \delta_{ij}: i = 1, 2, ..., n; j = 1, 2, ..., n_i\}$, where i indexes individuals or clusters, and j indexes the multiple survival times or individuals within clusters, could be written as

$$L\left(\theta,\beta,\lambda_0\middle|t_{ij},\delta_{ij},\omega_i\right)=\prod_{i=1}^{n}\prod_{j=1}^{n_i}S\left(t_{ij}\middle|\omega_i,\beta\right)\lambda\left(t_{ij}\middle|\omega_i,\beta\right)^{\delta_{ij}}g\left(\omega_i\middle|\theta\right)$$

$$=\prod_{i=1}^{n}\prod_{j=1}^{n_i}\exp\left[-\Lambda\left(t_{ij}\middle|\omega_i,\beta\right)\right]\lambda\left(t_{ij}\middle|\omega_i,\beta\right)^{\delta_{ij}}g\left(\omega_i\middle|\theta\right)$$

$$=\prod_{i=1}^{n}\prod_{j=1}^{n_i}\exp\left[-\omega_i\Lambda_0\left(t_{ij}\right)\exp\left(\mathbf{z}_{ij}\beta\right)\right]\left\{\omega_i\lambda_0\left(t_{ij}\right)\exp\left(\mathbf{z}_{ij}\beta\right)\right\}^{\delta_{ij}}g\left(\omega_i\middle|\theta\right)$$

if the frailty factors ω_i were observed. Thus the log likelihood can be written as

$$l\left(\theta,\beta,\lambda_0\middle|t_{ij},\delta_{ij},\omega_i\right)$$

$$=\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left\{\delta_{ij}\left(\log\left(\lambda_0\left(t_{ij}\right)\right)+\log\left(\omega_i\right)+\mathbf{z}_{ij}\beta\right)-\Lambda_0\left(t_{ij}\right)\exp\left(\log\left(\omega_i\right)+\mathbf{z}_{ij}\beta\right)\right\}$$

$$+\sum_{i=1}^{n}\left\{n_i\log\left(g\left(\omega_i\middle|\theta\right)\right)\right\}$$

$$=l_1\left(\beta,\lambda_0\middle|t_{ij},\delta_{ij},\omega_i\right)+l_2\left(\theta\middle|t_{ij},\delta_{ij},\omega_i\right)$$

so that the full log likelihood, given the true values of the frailty factors, consists of two parts, the first dependent upon β and λ_0, the second upon θ.

Since the distributions of the ω_i are assumed to have a mean of 1, the initial values of the ω_i can be taken as 1 (i.e. the initial value of θ is 0). Initial values for β and λ_0 can be taken from the fit of the standard model, assuming that there are no frailty effects.

The E-step of the EM algorithm will entail calculating the expected values of ω_i, given t_{ij}, δ_{ij}, and the current values for θ, β and λ_0. The M-step will require the maximisation of the log likelihood, that is, maximisation of $l_1\left(\beta,\lambda_0\middle|t_{ij},\delta_{ij},\omega_i\right)$ and $l_2\left(\theta\middle|t_{ij},\delta_{ij},\omega_i\right)$.

At this point, note that $l_1\left(\beta,\lambda_0\middle|t_{ij},\delta_{ij},\omega_i\right)$ is exactly the log likelihood of a standard proportional hazards model which includes an offset term of $\log(\omega_i)$ in the linear

predictor, and the maximum likelihood estimates for β and λ_0 can be found using standard statistical software. Maximisation of $l_2\left(\theta\middle|t_{ij},\delta_{ij},\omega_i\right)$ may be achieved algebraically, or by numerical methods, depending on the form of $\log\left(g\left(\omega_i\middle|\theta\right)\right)$. It will often be possible to automate this step by using the previous value of θ as the starting point for the subsequent numerical maximisation.

The only remaining problem is to determine the expectation of ω_i given the observed data and the current values of the parameters, θ, β and λ_0. That is, to find the expectation of

$$f\left(\omega_i\middle|t_{ij},\delta_{ij},\mathbf{z}_{ij},\theta,\beta,\lambda_0\right)$$

$$=\frac{f\left(t_{ij},\delta_{ij}\middle|\omega_i,\mathbf{z}_{ij},\theta,\beta,\lambda_0\right)\times f\left(\omega_i\middle|\mathbf{z}_{ij},\theta,\beta,\lambda_0\right)}{f\left(t_{ij},\delta_{ij}\middle|\mathbf{z}_{ij},\theta,\beta,\lambda_0\right)}$$

$$\propto g\left(\omega_i\middle|\theta\right)\times\prod_{j=1}^{n_i}S\left(t_{ij}\middle|\omega_i,\mathbf{z}_{ij},\theta,\beta,\lambda_0\right)\lambda\left(t_{ij}\middle|\omega_i,\mathbf{z}_{ij},\theta,\beta,\lambda_0\right)^{\delta_{ij}}$$

$$=g\left(\omega_i\middle|\theta\right)\times\prod_{j=1}^{n_i}\exp\left(-\omega_i\Lambda_0\left(t_{ij}\right)\exp\left(\mathbf{z}_{ij}\beta\right)\right)\left(\omega_i\lambda_0\left(t_{ij}\right)\exp\left(\mathbf{z}_{ij}\beta\right)\right)^{\delta_{ij}}$$

$$\propto g\left(\omega_i\middle|\theta\right)\omega_i^{\,D_i}\exp\left(-\omega_i\sum_{j=1}^{n_i}\Lambda_0\left(t_{ij}\right)\exp\left(\mathbf{z}_{ij}\beta\right)\right)$$

where $D_i=\sum_{j=1}^{n_i}\delta_{ij}$ is the total number of observed failures for subject or cluster i.

A convenient form for $g(\omega|\theta)$ is therefore the gamma distribution, with variance θ, so that

$$g\left(\omega\middle|\theta\right)=\frac{\omega^{1/\theta-1}\exp\left(-\omega/\theta\right)}{\Gamma\left(1/\theta\right)\theta^{1/\theta}}. \qquad \textbf{(Eq. 7.6)}$$

Thus

$$f\left(\omega_i\middle|t_{ij},\delta_{ij},\mathbf{z}_{ij},\theta,\beta,\lambda_0\right)\propto\omega_i^{\,D_i+1/\theta-1}\exp\left(-\omega_i\left\{\frac{1}{\theta}+\sum_{j=1}^{n_i}\Lambda_0\left(t_{ij}\right)\exp\left(\mathbf{z}_{ij}\beta\right)\right\}\right),$$

so that, conditional upon the observed data and the current values of the parameters, the individual frailty factors have gamma distributions with shape parameters equal to

$A_i = D_i + 1/\theta$ and scale parameters $C_i = \left\{\frac{1}{\theta} + \sum_{j=1}^{n_i} \Lambda_0(t_{ij}) \exp(z_{ij}\beta)\right\}^{-1}$. Consequently, the expected values of ω_i can be written as $A_i C_i$.

Convergence of the EM algorithm provides maximum likelihood estimates of the parameters, but not the standard errors of these parameter estimates. Profile likelihood methods could be used to determine the statistical significance of parameter values and their confidence intervals. Alternatively, standard errors could be estimated from the observed information matrix, i.e. the second derivative of the negative log likelihood.

Klein[123] applied the EM algorithm to a model with a semi-parametric baseline hazard and calculated the information matrix by substituting the maximum likelihood estimates of the cumulative hazard function, $\hat{\Lambda}_0(t)$, into the observable log likelihood, and differentiating with respect to θ and β. It was subsequently noted[124] that this would produce standard errors for the parameters that were too small, since the estimator $\hat{\Lambda}_0(t_{(k)})$ is a random variable and is not independent of $\hat{\theta}$ and $\hat{\beta}$. To derive unbiased estimates of the variance of $\hat{\theta}$ and $\hat{\beta}$, it is necessary to calculate the derivatives of the log likelihood with respect to θ, β and λ_{0k}, the increments of the cumulative hazard at the unique death times.

This parameterisation introduces an increasing number of unknowns into the likelihood, as the sample size and hence the number of unique event times increases, so that normal likelihood theory may not apply and variance estimates may not be consistent. As noted by Andersen et al[124], it is possible to use the non-parametric information matrix to gain consistent and asymptotically normal estimates in a frailty model. Once the correct information matrix is used, standard errors of the parameter estimates can be obtained, allowing the construction of confidence intervals and performance of hypothesis tests for these quantities.

More complex models than the Gamma frailty model can also be fitted using the EM algorithm. The Cox model has been extended to incorporate random genetic and environmental effects on the age of onset of disease[125], written as

$$\lambda(t_{ij} | \omega_i, z_{ij}, g_{ij}) = \omega_i \lambda_0(t_{ij}) \exp(z_{ij}\beta + \mu_{g_{ij}}),$$

where ω_i are Gamma distributed frailty terms representing random environmental effects and $\mu_{g_{ij}} = 0$ if the $(i,j)^{th}$ individual is not susceptible to the disease (i.e. g_{ij} = aa,

according to the single Mendelian diallelic locus model) or $\mu_{g_{ij}} = \mu$ if the $(i,j)^{th}$ individual is susceptible to disease (g_{ij} = Aa or AA). The E-step of the EM algorithm cannot be carried out directly, but a Monte Carlo method based on the Gibbs sampler[126] can be used.

7.4.2.4 Maximum Likelihood

The EM algorithm approach can be used for either parametric or semi-parametric baseline hazard models. However, the algorithm may be slow when nearing the maximum likelihood estimates of the parameters. Also, if the likelihood must be differentiated twice to obtain the observed information matrix, it may be preferable to use numerical techniques to maximise the likelihood directly. If the frailty factors of each individual or cluster are themselves of interest, these can be calculated once the maximum likelihood estimators have been reached.

As before, a proportional hazards model is assumed, with frailty factors acting multiplicatively on the hazard function. Frailty factors are assumed to have a gamma distribution, with mean 1 and variance θ, so that the density of ω is given by (Eq. 7.6). Conditional upon the frailty term for subject i, ω_i, that subject's survivor function can be written as

$$\begin{aligned} S(t_i|\omega_i) &= P(T_i \geq t_i|\omega = \omega_i) \\ &= \exp\{-\omega_i \Lambda_0(t_i)\exp(\mathbf{z}_i\beta)\} \end{aligned}$$

where $\Lambda_0(t_i) = \int_0^{t_i} \lambda_0(t)dt$ is the baseline cumulative hazard function. Since ω_i is unobserved, the population (rather than the individual) survivor function,

$$\begin{aligned} S(t_i) &= P(T_i \geq t_i) \\ &= E_\omega[P(T_i \geq t_i|\omega = \omega_i)] \\ &= E_\omega[\exp\{-\omega_i \Lambda_0(t_i)\exp(\mathbf{z}_i\beta)\}] \end{aligned} \qquad \textbf{(Eq. 7.7)}$$

which is the Laplace transform of ω_i evaluated at $\Lambda_0(t_i)\exp(\mathbf{z}_i\beta)$, is used. Since ω_i is gamma distributed, (Eq. 7.7) can be evaluated explicitly as

$$S(t_i) = \{1 + \theta\Lambda_0(t_i)\exp(\mathbf{z}_i\beta)\}^{-1/\theta}$$

and

$$\lambda(t_i) = -\frac{S'(t_i)}{S(t_i)}$$
$$= \frac{\lambda_0(t_i)\exp(\mathbf{z}_i\beta)}{1+\theta\Lambda_0(t_i)\exp(\mathbf{z}_i\beta)}.$$

Thus, given the data of survival times t_{ij} and censoring indicators δ_{ij} observed on n subjects, with n_i (≥ 1) observations made of each subject, the likelihood function can be written as

$$L(\theta,\beta,\lambda_0(t)) = \prod_{i=1}^{n}\prod_{j=1}^{n_i} S(t_{ij})\lambda(t_{ij})^{\delta_{ij}}$$
$$= \prod_{i=1}^{n}\prod_{j=1}^{n_i} \left\{1+\theta\Lambda_0(t_{ij})\exp(\mathbf{z}_{ij}\beta)\right\}^{\frac{1}{\theta}-\delta_{ij}} \left\{\lambda_0(t_{ij})\exp(\mathbf{z}_{ij}\beta)\right\}^{\delta_{ij}}.$$

The gamma distribution is a special case of a broader class of distributions, the positive stable class[127], for which similar algebra can be used to derive a closed form for the likelihood of the data.

This approach is particularly advantageous when using a parametric baseline hazard function, since there are few parameters associated with the hazard. With a semi-parametric hazard, however, a large number of parameters may be associated with the baseline hazard, since one parameter must be included for every unique event time. Even for moderately large sample sizes, this could mean a few hundred parameters, and the second derivative of the likelihood can become extremely unwieldy, slowing down the maximisation routine.

Example 7.4 Gamma Frailty Model with Weibull Baseline Hazard Function for Time to Anginal Pain

A convenient and flexible parametric baseline hazard function is the Weibull distribution (so that $\lambda_0(t) = \alpha t^{\alpha-1}$ and $\Lambda_0(t) = t^{\alpha}$), so that the log likelihood can be written as

$$l(\theta,\beta,\alpha) = -\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left\{\left(\frac{1}{\theta}+\delta_{ij}\right)\log\left(1+\theta t_{ij}^{\alpha}\exp(\mathbf{z}_{ij}\beta)\right)+\delta_{ij}\left(\log(\alpha)+(\alpha-1)\log(t_{ij})+\mathbf{z}_{ij}\beta\right)\right\}.$$

This is equivalent to the usual two-parameter form of the Weibull distribution (Table 4.1) if the covariate matrix **z** includes a column of ones, as with the intercept term of a standard linear regression model.

	Hazard Ratio	95% CI	p
Test Occasion (Test 3 : Test 1)	0.13	0.08, 0.20	<0.0001
Treatment (Nifedipine : Atenolol)	1.43	0.83, 2.48	0.19
Treatment (Combination : Atenolol)	0.98	0.56, 1.70	0.93
Gender (Female : Male)	2.15	1.30, 3.54	0.0028
Age (/10 years)	1.55	1.27, 1.90	<0.0001
Weight (Treadmill) (/10 kg)	1.33	1.05, 1.68	0.019
Weight (Bicycle) (/10 kg)	0.77	0.64, 0.92	0.0054

	Estimate	95% CI
Gamma Frailty Variance	2.35	2.04, 2.66
Baseline Weibull Shape	4.02	3.67, 4.37
Baseline Weibull Scale (Treadmill)	0.0033	0.0031, 0.0036
Baseline Weibull Scale (Bicycle)	0.0019	0.0017, 0.0020

Table 7.5 Period, treatment and covariate effect estimates, with 95% CIs and p-values from Gamma frailty model with Weibull baseline hazard function for repeated exercise times to anginal pain, plus estimates and 95% CIs for frailty variance, common baseline shape parameter and exercise-type-specific baseline scale parameters

This can be maximised with respect to θ, β and α, subject to $\theta, \alpha > 0$, to obtain the maximum likelihood estimators of these parameters, $\hat{\theta}$, $\hat{\beta}$ and $\hat{\alpha}$, and the variance-covariance matrix of these estimates can be estimated from the observed information matrix. The first and second derivatives of this likelihood function are given in Appendix B.

Table 7.5 shows the estimates from fitting this model to data from the TIBET Study. The exercise times used were the times to anginal pain during both the first (off-treatment) and the third (on-treatment) exercise tests. The hazard for suffering anginal pain during exercise was much reduced at the third exercise test, with a hazard ratio of 0.13 (95% CI 0.08-0.20). Women and older patients were seen to be more likely to suffer an event, and the effect of weight was dependent upon the type of exercise, with heavier patients being more likely to experience anginal pain when using a treadmill, but on a bicycle, increased weight was associated with lower hazard. There was no evidence of any treatment effects based on this model.

There was a considerable frailty effect, with a frailty variance estimate of 2.35. This can be interpreted by considering the hazard ratio between two individuals with identical covariates, one of whom lies at the 90th percentile of the population frailty distribution and the other who lies at the 10th percentile. Since the frailty terms are Gamma distributed with mean 1 and variance θ, the shape and scale of this Gamma distribution will both be θ^{-1}. The hazard ratio between these extreme individuals would be 16.2, demonstrating the degree of variability within the population being studied.

7.4.2.5 Penalized Likelihood

Penalized likelihood can be incorporated into survival regression models to allow the application of shrinkage methods and fitting of smoothing splines for the estimation of non-linear covariate effects[128]. To illustrate, consider a statistical model that is parameterised by β and γ, so that the log likelihood can be written as

$$\log \text{lik} = \log \text{lik}\left(\beta, \gamma | \text{data}\right).$$

The method of penalized likelihood defines the parameters γ to be "constrained", and reflects this in the log likelihood function through a penalty function that assigns large values to unwanted values of γ,

$$\text{pen log lik} = \log \text{lik}\left(\beta, \gamma | \text{data}\right) - f(\gamma, \theta), \qquad \textbf{(Eq. 7.8)}$$

where the penalty function f( ) depends upon θ, a vector of "tuning" parameters.

If the Cox model partial likelihood is used, along with the penalty function

$$f(\gamma, \theta) = -\sum_{i=1}^{n} \frac{1}{\theta}\left[\gamma_i - \exp(\gamma_i)\right] - \log\theta - \log\Gamma(1/\theta),$$

then the penalized log likelihood can be shown[93] to be equivalent to the gamma frailty model, with the gamma frailties $\omega_i = \exp(\gamma_i)$.

An alternative penalty function,

$$f(\gamma, \theta) = \frac{1}{2\theta}\sum_{i=1}^{n} \gamma_i^{\,2},$$

results in a Cox model with frailties from a Normal distribution[87] with variance θ; these techniques can be extended further to allow frailty terms to have any viable covariance matrix, such as an autoregressive correlation[129]. However, for the analysis of exercise

	Standard Cox Model		Cox Model with Gamma Frailty		Cox Model with Gaussian Frailty	
	Haz. Ratio (95% CI)	p	Haz. Ratio (95% CI)	p	Haz. Ratio (95% CI)	p
Nifedipine-Atenolol	1.33 (0.94,1.87)	0.11	1.55 (1.00,2.40)	0.050	1.47 (0.95,2.26)	0.083
Combination-Atenolol	1.06 (0.74,1.52)	0.76	0.85 (0.54,1.34)	0.49	0.89 (0.57,1.39)	0.61
Gender (Female - Male)	1.35 (1.03,1.77)	0.032	1.87 (1.04,3.36)	0.035	1.86 (1.10,3.14)	0.020
Age (/10 years)	1.26 (1.12,1.42)	0.0001	1.58 (1.24,2.00)	0.0002	1.53 (1.23,1.92)	0.0002
Weight (/10 kg)	0.99 (0.90,1.08)	0.74	0.91 (0.76,1.09)	0.32	0.94 (0.79,1.11)	0.47
Frailty Variance	-		2.48		2.52	

Table 7.6 Effect estimates (as hazard ratios, with 95% CIs and p-values) from Cox proportional hazards models for the time to anginal pain, stratified by exercise type and dependent upon treatment, gender, age and weight, fitted with and without frailty

test data, such generality would seem unnecessary, and a simple shared frailty model should suffice.

Example 7.5 Cox Regression Model with Gamma Frailties Fitted by Penalized Likelihood to Time to Anginal Pain

Table 7.6 shows treatment and covariate effect estimates from Cox proportional hazards models for the time to anginal pain during the first or the third exercise test of the TIBET Study. All models are fitted with separate baseline hazard functions according to the type of exercise performed and the occasion (off-treatment or on-treatment) of the test. The first model is a standard Cox model, as shown in Table 7.4, also shown are the results obtained by fitting the Cox model with frailty terms included, in these examples using frailty terms from Gamma and Gaussian (Normal) distributions.

The Gamma and Gaussian frailty models give very similar estimates of model effects as well as the frailty variance. Both give greater variance to these estimates than the standard Cox model, though the effect estimates tend to be larger, particularly for gender and age.

In the penalized likelihood formulation of the model (Eq. 7.8), the frailty terms are incorporated into the hazard through the linear predictor

$$\lambda(t) = \lambda_0(t)\exp(z\beta + x\gamma), \qquad \textbf{(Eq. 7.9)}$$

where the **x** matrix consists of indicator variables for membership of each cluster of observations and the frailty terms, γ, relate to the multiplicative definition (Eq. 7.5) as $\omega = \exp(\gamma)$, so that whilst ω follows a Gamma distribution with mean 1, γ follows a log Gamma distribution, subject to the constraint that $E[\exp(\gamma)] = 1$, though the variance reported in the model fit is that of γ. The Gaussian frailty model, however, is fitted according to (Eq. 7.9) with γ following a Normal distribution with zero mean, and the variance of this distribution is reported with the model fit.

The interpretation of these variances in terms of statistical significance is unclear. On the one hand, from the point of view of the classical frailty model, there is one additional parameter being fitted to each frailty model, and so the change in log likelihood from the non-frailty model should be compared to a χ^2 distribution with one degree of freedom. However, in the penalized likelihood framework[128,87], the degrees of freedom attached to the estimation of the frailty variances are much larger; 396.6 in the Gamma model and 351.4 in the Gaussian. There is as yet no consensus as to the best method of testing the goodness-of-fit of frailty models fitted in this way, in terms of likelihood comparisons with standard survival regression models.

CHAPTER 8 Simulation Study II: Analysis of Paired Survival Data

In this chapter the results of a second simulation study will be presented. The aim of this study was to compare different methods of analysing paired survival data, designed to mimic pairs of exercise tests carried out in a clinical trial, in order to gain some insight into the situations that contribute to the performance of these methods. With this in mind, pairs of correlated survival times were simulated, the first corresponding to an "off treatment" exercise test, and the second "on treatment". Two groups of pairs were generated, corresponding to two treatment groups, so that the distribution of exercise times in the two groups was the same for the "off treatment" test, but different in the "on treatment" test.

8.1 Generation of simulated data

Data were simulated for two parallel groups of individuals, as if each performed two exercise tests, with the first test being carried out off treatment and the second test being carried out on treatment, with different treatments being used in the two groups of individuals. Three sample sizes were considered; 100, 200 and 400 individuals per treatment group.

Varying levels of dependence within pairs of exercise times on the same simulated individual was imposed by assuming a Normal frailty model. Each individual in a simulated trial was assigned a frailty, taken to be an observation from a standard Normal distribution. Taking α to be the shape parameter and γ to be the baseline scale parameter in the simulated off treatment exercise times, then if ω_i is the simulated frailty of the i^{th} individual, their off treatment survival time was an observation from a Weibull distribution with shape α and scale

$$\gamma_{i1} = \gamma_0 \exp\left(-\frac{\phi\omega_i}{\alpha}\right)$$

where ϕ is the desired standard deviation of the frailty distribution. In these simulations, α was taken as 2, and γ_0 as

$$\frac{\Gamma\left(1+\frac{1}{\alpha}\right)}{400}$$

so that an individual with mean (zero) frailty would have a mean failure time of 400. Five values of ϕ were considered; 0, 0.1, 0.3, 0.9 and 2.

On treatment survival times were also generated from a Weibull distribution with a shape parameter of 2, with scale parameters related to the baseline scale parameter through constant hazard ratios. The on treatment scale parameter of the i^{th} individual was taken to be

$$\gamma_{i2} = \gamma_0 \exp\left(-\frac{1}{\alpha}\left[\phi\omega_i + \beta_1 + \beta_2 x_i\right]\right)$$

where β_1 was given the value 0.1, to represent a period effect. x_i took values 0 or 1 depending on the treatment group so that β_2 represents the simulated treatment effect. Four values of β_2 were considered; 0, -0.1, -0.3 and -1.

For each simulated failure time, a corresponding censoring time was generated independently as an observation from an exponential distribution with a mean of 800. A fixed maximum observed time of 1000 was also imposed, to mimic this aspect of an exercise test. The lesser of the failure and censoring times for each individual was taken as their observed exercise time, and failure indicators were calculated accordingly.

Under each of the 60 combinations of sample size, frailty variance and treatment effect, 1000 simulated studies were generated and analysed by each of the methods being compared.

8.2 Models

Four different methods were applied to analyse each set of simulated data. Of the rank tests available for correlated survival data, the Akritas test was used, since a previous simulation study[110] has suggested this to be as good as, if not better than other similar methods. Three Cox proportional hazards models were considered. The first was a standard model, ignoring baseline exercise times and estimating the hazard ratio between the two groups of on treatment times. The other two models were a marginal model, and a frailty model with Gamma-distributed frailties. For these, all observations

were analysed, with a single categorical variable included in the model with three levels corresponding to whether the test was off treatment, on treatment (group A) and on treatment (group B). The treatment effect was determined by the contrast between the two on-treatment levels of this variable.

Three methods based on the t-test were applied. The first simply performed a test on the difference between on-treatment and off-treatment survival times, comparing the two treatment groups. The second was to perform the same test using only those pairs of observations where both survival times corresponded to failures. Finally, the method outlined in Section 7.3 was applied, in which the differences between pair members are assumed to come from a Normal distribution, and maximum likelihood is used to estimate the difference between treatment groups taking account of the left- and right-censoring.

8.3 Results

8.3.1 Bias

Table 8.1 shows the mean estimated treatment effects from the proportional hazards models under each combination of simulated treatment effect, frailty standard deviation (SD) and sample size. As expected, all models show no bias when there is no treatment effect. All models are unbiased when there is no frailty variability, since all models are correctly specified under these circumstances. Otherwise, the frailty model performs best, with no discernible bias from these results except when the treatment effect is at its largest in this set of simulations and the frailty SD is 0.9 or 2.0. In these extreme conditions, the use of a Gamma frailty model to estimate parameters when the data were generated from a Normal frailty model leads to underestimation of treatment effects by approximately 5%. However, with the same simulation parameters, the other models used underestimate treatment effects by more than 50%. The degree of bias is related to the extent of frailty variability, with estimates from the marginal and standard Cox models being approximately half of the true value when the frailty SD is 2, and about 70% of the true value when the frailty SD is 0.9. The degree of bias seems to reflect the degree of model mis-specification, since with smaller levels of frailty SD, there is no detectable bias with any of the models.

		Model								
		Frailty			Marginal			Cox (2nd period)		
Sample Size		100	200	400	100	200	400	100	200	400
β_2	ϕ									
0	0	-0.01	0.01	0.01	-0.01	0.01	0.01	-0.01	0.01	0.01
0	0.1	0.01	-0.01	0.01	0.01	-0.01	0.01	0.01	-0.01	0.01
0	0.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0	0.9	0.01	0.01	0.00	0.00	0.01	0.00	0.01	0.01	0.00
0	2	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.01
-0.1	0	-0.11	-0.09	-0.10	-0.10	-0.09	-0.10	-0.10	-0.09	-0.10
-0.1	0.1	-0.10	-0.11	-0.10	-0.09	-0.10	-0.10	-0.09	-0.10	-0.10
-0.1	0.3	-0.11	-0.09	-0.10	-0.10	-0.08	-0.09	-0.10	-0.08	-0.09
-0.1	0.9	-0.09	-0.10	-0.09	-0.07	-0.08	-0.07	-0.07	-0.08	-0.07
-0.1	2	-0.10	-0.09	-0.09	-0.05	-0.04	-0.04	-0.05	-0.04	-0.04
-0.3	0	-0.32	-0.31	-0.30	-0.31	-0.31	-0.30	-0.31	-0.31	-0.30
-0.3	0.1	-0.31	-0.32	-0.30	-0.30	-0.31	-0.30	-0.31	-0.31	-0.30
-0.3	0.3	-0.31	-0.29	-0.29	-0.30	-0.28	-0.28	-0.30	-0.28	-0.28
-0.3	0.9	-0.28	-0.30	-0.30	-0.21	-0.22	-0.22	-0.21	-0.22	-0.22
-0.3	2	-0.29	-0.28	-0.28	-0.14	-0.13	-0.14	-0.15	-0.13	-0.14
-1	0	-1.03	-1.01	-1.01	-1.01	-1.00	-1.00	-1.00	-1.00	-1.00
-1	0.1	-1.04	-1.03	-1.01	-1.02	-1.02	-1.00	-1.02	-1.02	-1.00
-1	0.3	-1.02	-0.99	-0.98	-0.96	-0.95	-0.95	-0.97	-0.95	-0.95
-1	0.9	-0.97	-0.97	-0.98	-0.74	-0.73	-0.73	-0.75	-0.74	-0.74
-1	2	-0.94	-0.93	-0.94	-0.48	-0.47	-0.47	-0.48	-0.47	-0.47

Table 8.1 Mean treatment effect estimates from three Cox proportional hazards models (Gamma frailty model, marginal model for clustered data and standard Cox model using 2nd period data only) for each combination of simulated treatment effect (β_2), frailty standard deviation (ϕ) and sample size

The methods that assume a Normal distribution for the difference between survival times are not estimating a quantity that can be directly specified from the parameters of the model. The off-treatment baseline mean failure time was designed to be 400, though the expected failure time for the i^{th} individual, given their frailty, would be

$$\mu_{i1} = 400\exp\left(-\frac{\phi\omega_i}{\alpha}\right)$$

and their expected on-treatment failure time would be

$$\mu_{i2} = 400\exp\left(-\frac{1}{\alpha}\left[\phi\omega_i + \beta_1 + \beta_2 x_i\right]\right).$$

The population expected difference between treatment groups in their change in survival time (which is the same as the population expected difference between treatment groups in on-treatment survival times) can be written as

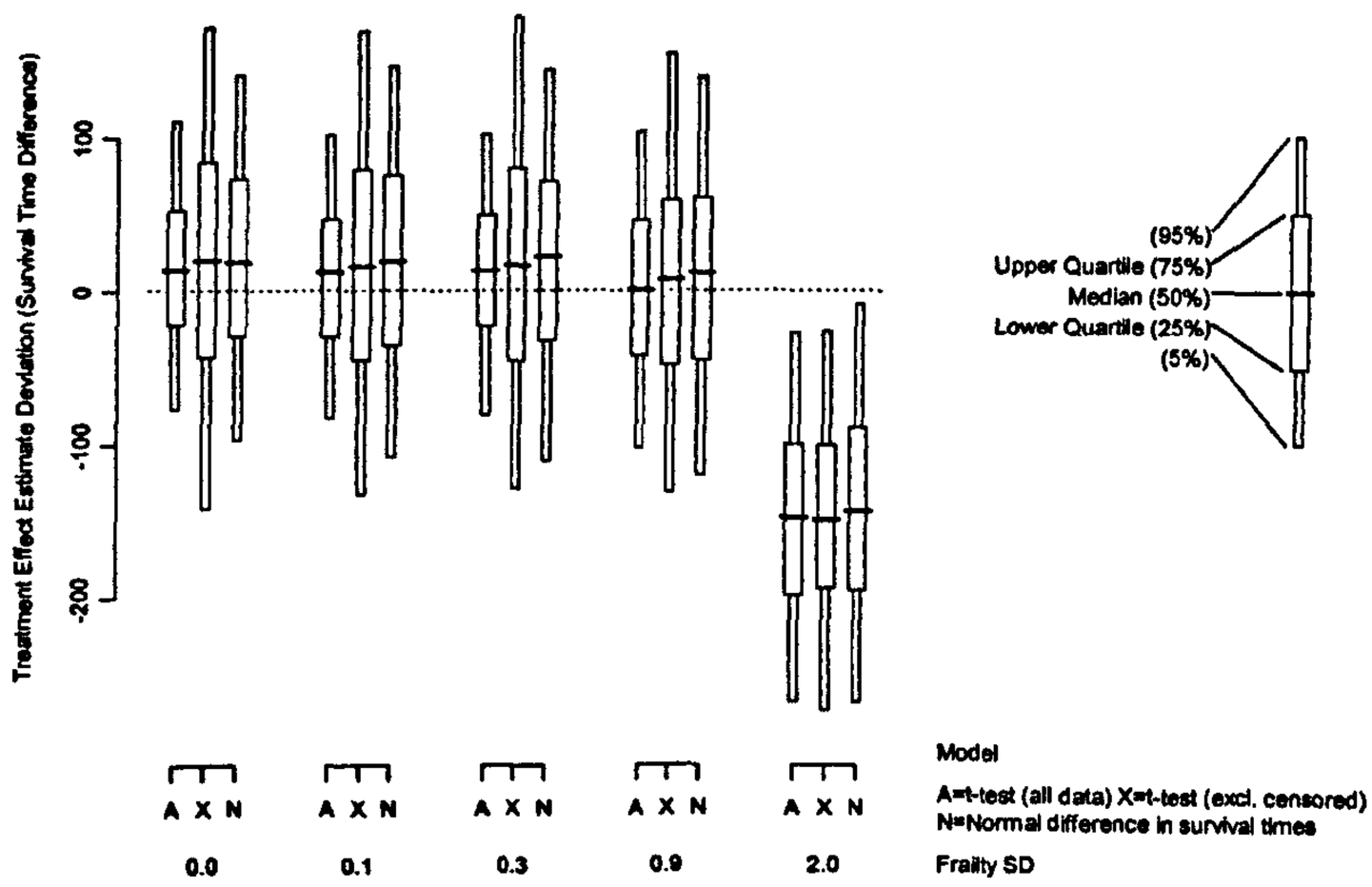


Figure 8.1 Distribution of deviations of treatment effect estimates from target values under t-test using all data, t-test using fully observed pairs and maximum likelihood method assuming differences in survival times to be Normally distributed, where sample size is 100 and β_2= -0.3, for increasing levels of frailty SD

$$E[\Delta_i] = E_\omega\left[400\exp\left(-\frac{1}{\alpha}\phi\omega_i\right)\exp\left(-\frac{1}{\alpha}\beta_1\right)\left\{\exp\left(-\frac{1}{\alpha}\beta_2\right)-1\right\}\right]$$

$$= 400\,E_\omega\left[\exp\left(-\frac{1}{\alpha}\phi\omega_i\right)\right]\exp\left(-\frac{1}{\alpha}\beta_1\right)\left\{\exp\left(-\frac{1}{\alpha}\beta_2\right)-1\right\}$$

for which the expectation can be evaluated numerically, since the distribution of the ω_i is known to be a standard Normal.

By these means, the expected difference between treatment groups in the change in survival time can be evaluated for each simulated treatment effect and frailty SD, so that the deviation of each simulation estimate from its expected value can be derived, and over the set of simulations, give an estimate of mean bias of each method. There was no evidence of bias under the simulations with no treatment effect. When treatment effects were non-zero, the estimates differed from their target values as shown in Figure 8.1, for simulations where β_2= -0.3 and the sample size was 100.

When the frailty SD was 0.9 or less, there was a slight positive bias for each method over a set of 1000 simulations, though when the frailty SD was large, all three methods drastically underestimated the true difference in survival time between treatment groups. This picture became more extreme as simulated treatment effects (and sample sizes) were increased, as shown in Figure 8.2. For moderate simulated frailty

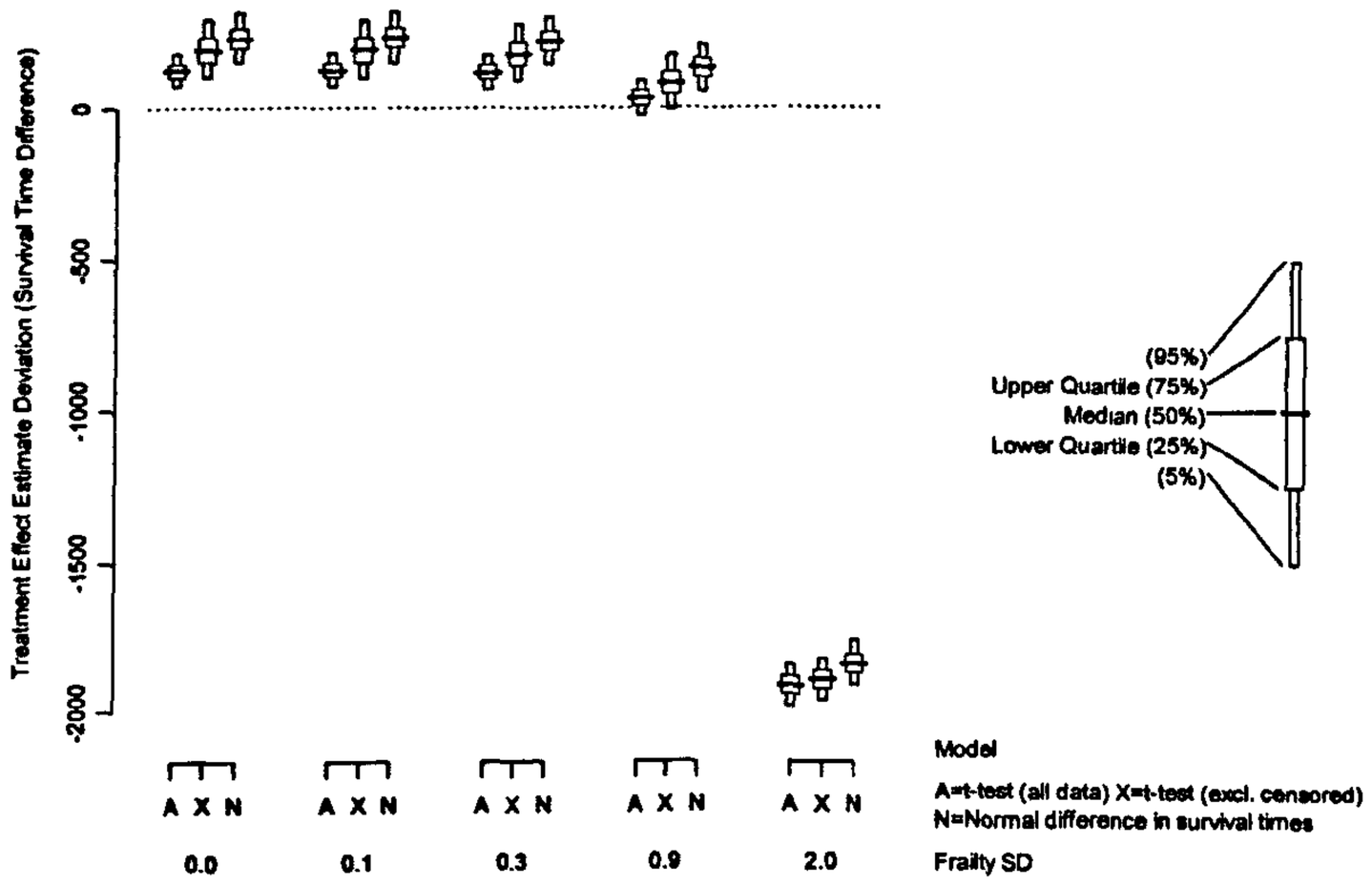


Figure 8.2 Distribution of deviations of treatment effect estimates from target values under models assuming the difference in survival times to be Normally distributed, where sample size is 400 and $\beta_2 = -1$, for increasing levels of frailty SD

SDs, the three methods are biased, producing estimates of the difference between treatment groups that are too large. When the frailty SD is 2, there will be larger numbers of simulated pairs for whom the hazard will be small, and the expected failure times will be relatively large, so that the difference in mean failure times will also be large. However, these pairs will be censored at the maximum censoring time of 1000, and so would provide a zero estimate of the difference in survival time under the t-test using all data, and would be excluded from the t-test that used only fully observed pairs, and would not contribute to the likelihood of the third method, since both observations were censored.

In fact, the likelihood based method, assuming differences in survival times within pairs were Normally distributed, showed very poor convergence when the frailty SD was at its largest value, with between 15% and 45% of simulations in any set of 1000 failing to converge. A number of different algorithms were tried to find starting values for which the minimisation routine would converge, but none proved successful. However, it is most likely a consequence of the way the data were generated, with a common fixed maximum survival time that causes this poor performance, rather than a failing of the method itself.

Sample Size	Frailty SD	Method						
		Proportional Hazards				Normally Distributed Differences		
		Gamma Frailty Model	Marginal Model	Standard Cox Model	Akritas Test	Gamma Frailty Model	Marginal Model	Standard Cox Model
100	0	5.4%	5.7%	5.5%	6.4%	5.6%	5.7%	6.8%
	0.1	5.3%	5.0%	5.1%	5.2%	5.3%	5.5%	5.1%
	0.3	5.6%	5.2%	4.6%	4.8%	4.0%	4.7%	4.4%
	0.9	6.6%	6.6%	6.1%	6.4%	4.8%	3.7%	5.8%
	2	4.7%	4.3%	4.0%	4.2%	5.6%	5.0%	6.5%
200	0	5.8%	6.5%	5.9%	5.0%	6.3%	5.1%	4.9%
	0.1	4.4%	4.7%	4.4%	5.0%	4.7%	4.7%	5.0%
	0.3	5.3%	4.8%	4.5%	4.6%	6.0%	5.5%	5.4%
	0.9	5.1%	4.2%	4.0%	4.1%	4.5%	5.5%	6.4%
	2	5.1%	4.6%	4.7%	5.4%	5.3%	4.3%	6.2%
400	0	5.2%	4.8%	4.7%	5.1%	5.1%	5.3%	4.3%
	0.1	5.1%	4.8%	4.8%	4.7%	5.2%	4.2%	4.4%
	0.3	4.6%	4.9%	4.2%	5.5%	5.3%	5.1%	4.5%
	0.9	6.7%	5.9%	5.6%	5.3%	5.3%	4.5%	4.9%
	2	5.0%	6.4%	6.3%	6.0%	5.6%	5.7%	4.4%

Table 8.2 Type I error rates (%) of each method under different simulated sample sizes and frailty standard deviations (SDs), based on a 5% significance test

8.3.2 Error Rates and Power

Table 8.2 shows the Type I error rates of each method based on statistical tests at the 5% significance level under each simulated scenario (Type I error rates are the proportion of trials in which a significant result is found when there is no treatment effect, i.e. β_2=0). Based on these simulations, all the methods considered appear to perform adequately in this respect.

The power of the different methods are shown in Table 8.3, for those simulations in which the treatment effect was simulated to be a log hazard ratio of -0.1 between groups for the on-treatment exercise test. None of the sample sizes is adequate to detect this small an effect using any method. The marginal model in achieving only 12.5% power with a sample size of 400 per group when the frailty SD is a modest 0.1 gives the best performance of any of these methods.

When the simulated treatment effect is increased to -0.3, the power of all of these methods increases, as shown in Table 8.4. Again none of the methods is powerful

Sample Size	Frailty SD	Method						
		Proportional Hazards				Normally Distributed Differences		
		Gamma Frailty Model	Marginal Model	Standard Cox Model	Akritas Test	Gamma Frailty Model	Marginal Model	Standard Cox Model
100	0	6.6%	6.8%	6.2%	5.6%	5.8%	5.7%	4.9%
	0.1	7.8%	8.0%	7.3%	6.6%	5.8%	5.6%	5.7%
	0.3	7.1%	7.1%	6.3%	6.2%	4.4%	6.2%	6.6%
	0.9	7.7%	7.0%	6.2%	8.1%	4.4%	5.7%	5.7%
	2	4.4%	6.0%	6.0%	5.4%	5.0%	5.8%	7.8%
200	0	7.9%	8.1%	8.3%	7.2%	5.6%	5.8%	6.3%
	0.1	9.3%	9.9%	8.8%	7.9%	5.5%	4.8%	6.8%
	0.3	7.7%	7.7%	6.6%	6.9%	6.7%	5.6%	5.8%
	0.9	7.7%	7.9%	7.5%	8.9%	4.2%	4.8%	6.7%
	2	5.8%	4.6%	4.6%	6.6%	5.3%	3.9%	5.0%
400	0	12.0%	12.2%	12.0%	8.8%	7.5%	6.8%	8.7%
	0.1	12.4%	12.5%	11.8%	10.8%	7.1%	7.8%	9.1%
	0.3	10.7%	10.2%	10.0%	9.4%	8.2%	7.3%	8.6%
	0.9	8.4%	8.0%	7.5%	8.4%	5.6%	5.9%	7.2%
	2	6.2%	6.7%	6.5%	6.8%	6.1%	5.2%	6.3%

Table 8.3 Power (%) of each method based on a 5% significance test, under different simulated sample sizes and frailty standard deviations (SDs) for treatment effect simulated as a log hazard ratio of -0.1

enough for practical use; the best performance in this set of simulations was 64%, observed using the marginal model for a sample size of 400 with no frailty effect. However, a pattern is seen whereby the three proportional hazards models have approximately the same power when there is no or moderate frailty, but when the frailty SD is large (0.9 or 2) the Gamma frailty model out performs the other two models. It is of note that the marginal model fails to demonstrate clearly better power than the standard Cox model using only the on-treatment data. The Akritas test has levels of power close to those of the marginal and standard Cox models.

When the simulated frailty SD is 2, the method that has greatest power, after the (almost) correctly specified Gamma frailty model, is the maximum likelihood method that assumes the differences in survival times to be Normally distributed. This method has consistently greater power than either t-test, though itself has low power compared to the proportional hazards models and the Akritas test when there is little or no simulated within-pair correlation.

Sample Size	Frailty SD	Method						
		Proportional Hazards				Normally Distributed Differences		
		Gamma Frailty Model	Marginal Model	Standard Cox Model	Akritas Test	Gamma Frailty Model	Marginal Model	Standard Cox Model
100	0	21.5%	22.3%	20.5%	19.1%	7.8%	7.1%	12.6%
	0.1	20.4%	21.6%	20.3%	18.9%	8.2%	8.4%	12.6%
	0.3	19.4%	19.6%	18.5%	16.7%	8.9%	6.5%	9.9%
	0.9	13.8%	13.1%	12.0%	13.0%	6.6%	8.3%	12.9%
	2	10.5%	8.7%	8.6%	7.9%	7.0%	7.0%	9.1%
200	0	36.7%	38.1%	36.6%	31.1%	13.3%	12.4%	19.2%
	0.1	38.5%	38.9%	37.6%	33.2%	13.3%	12.2%	19.3%
	0.3	32.5%	32.7%	31.8%	28.9%	10.7%	8.3%	17.5%
	0.9	25.0%	21.9%	21.4%	22.9%	11.9%	12.2%	16.7%
	2	16.5%	12.3%	12.3%	11.5%	6.4%	7.2%	13.2%
400	0	63.1%	64.0%	62.8%	52.7%	22.6%	16.7%	32.2%
	0.1	62.0%	61.1%	60.4%	53.2%	22.2%	16.4%	29.8%
	0.3	57.7%	57.8%	56.9%	52.0%	20.4%	18.0%	32.5%
	0.9	43.5%	38.5%	38.2%	36.7%	16.7%	19.4%	27.2%
	2	27.7%	17.8%	17.3%	17.8%	11.8%	14.6%	23.5%

Table 8.4 Power (%) of each method based on a 5% significance test, under different simulated sample sizes and frailty standard deviations (SDs) for treatment effect simulated as a log hazard ratio of -0.3

When the simulated treatment effect is at is largest, as shown in Table 8.5, the Gamma frailty model achieves in excess of 80% power in every combination of sample size and frailty effect, except in the extreme frailty case with a sample of 100 per group. Otherwise, the patterns are similar to those observed previously.

8.4 Summary

In these simulations, when the data were generated according to a Weibull baseline hazard with proportional hazards effects and Normally distributed individual frailty terms, the Gamma frailty model performs best, with the least bias and greatest power in most circumstances, particularly when the extent of correlation due to the frailty distribution is large. This is as would be expected since this model matches most closely the way in which the data were generated.

The marginal model does not out-perform the standard Cox proportional hazards model using data from the 2nd period only, and both are similar to the frailty model

Sample Size	Frailty SD	Method: Proportional Hazards: Gamma Frailty Model	Proportional Hazards: Marginal Model	Proportional Hazards: Standard Cox Model	Akritas Test	Normally Distributed Differences: Gamma Frailty Model	Normally Distributed Differences: Marginal Model	Normally Distributed Differences: Standard Cox Model
100	0	93.6%	93.6%	92.4%	87.7%	43.7%	34.6%	70.4%
	0.1	94.1%	94.6%	93.9%	89.9%	44.1%	35.0%	69.5%
	0.3	91.8%	92.1%	91.2%	87.0%	44.2%	35.7%	67.5%
	0.9	80.3%	74.2%	73.2%	73.2%	37.2%	33.4%	64.3%
	2	59.3%	38.1%	35.9%	41.6%	19.9%	27.0%	52.9%
200	0	100.0%	100.0%	100.0%	99.7%	71.8%	59.8%	93.4%
	0.1	99.9%	99.8%	99.8%	99.4%	76.6%	64.7%	94.5%
	0.3	99.7%	99.7%	99.7%	99.7%	71.9%	65.1%	94.2%
	0.9	97.5%	96.4%	96.2%	95.5%	63.2%	56.9%	89.0%
	2	86.7%	65.4%	64.6%	68.5%	38.4%	48.4%	78.1%
400	0	100.0%	100.0%	100.0%	100.0%	96.9%	92.0%	99.8%
	0.1	100.0%	100.0%	100.0%	100.0%	95.6%	91.7%	100.0%
	0.3	100.0%	100.0%	100.0%	100.0%	94.7%	91.1%	99.9%
	0.9	100.0%	100.0%	100.0%	99.9%	90.2%	88.1%	99.8%
	2	99.5%	92.5%	92.7%	93.5%	63.0%	77.0%	97.4%

Table 8.5 Power (%) of each method based on a 5% significance test, under different simulated sample sizes and frailty standard deviations (SDs) for treatment effect simulated as a log hazard ratio of -1

when within-pair correlation in survival times are small. Their bias and reduced power when the frailty SD becomes larger might reflect the breakdown of the proportional hazards assumption at the population level, since the frailty model assumes proportional hazards within pairs of observations, but this assumption will be incorrect over the population as a whole.

The Akritas test offers a good alternative to proportional hazards models when the assumptions underlying these models might be in doubt, since it has good power over the range of frailty SDs considered here. The maximum likelihood technique, assuming differences in survival times to be Normally distributed, appears to perform well when heterogeneity due to frailty is large, but the convergence problems experienced in this situation and the extreme levels of bias found make it difficult to conclude that the method is of practical use. However, these features were observed only in the most extreme frailty case, and may be an artefact of the way that the data were generated, whereby the imposition of an upper limit to observed survival times will result in

double censoring of the least frail individuals, in whom the largest differences in exercise times would be expected. In practice, where initial exercise tests might be used as a final exclusion test, so that all (or most) first period tests are uncensored and therefore no (or few) pairs will be doubly censored, this may not be an issue.

CHAPTER 9 Further Work

In this penultimate Chapter some avenues for further work will be explored. Firstly, the application of competing risks methods to exercise test data will be considered. These methods can be used to model survival times when there is more than one type of failure. There are several reasons for which an exercise test can be stopped besides anginal pain, including fatigue, dyspnoea and muscular pain. Secondly, methods for analysing multivariate repeated measures data will be reviewed.

The driving principle behind the exploration of these methods is the desire to analyse the total response of test subjects to exercise. This mirrors the way that exercise tests are assessed by the experienced physician, who considers the haemodynamic and electrocardiographic response of each individual, before, during and after exercise, as well as the occurrence of various endpoints, both ischaemic and non-ischaemic.

Wherever possible, methods will be illustrated through examples using data from the TIBET Study.

9.1 Competing Risks

The theory of competing risks has its origins in the work of Daniel Bernoulli in 1760, when he presented his solution to the problem of describing the effect on population mortality of preventing deaths due to smallpox. The solution he provided rested on the assumption that the individuals who would otherwise have died from smallpox would have survival distributions for other causes of death that were the same as those of the rest of the population. Then, and in similar analyses since, the assumption of independence among types of failure in competing risks was made for reasons of computational feasibility.

Much work in the field of competing risks has been conducted in industrial statistics, in the analysis of failure times of systems made up of separate components, where the whole system is observed to fail at the first time that any single component fails. Applications in the medical arena have included demographic studies of

populations subject to several competing causes of death, or in cancer studies where the analysis models the times to either remission or death, where times to both types of failure could be censored due to withdrawal or loss to follow up.

In the analysis of exercise test data, competing risks could be applied to total exercise times, with the different possible reasons for stopping the test representing the different types of failure. The possible reasons for stopping an exercise test might include anginal pain, muscle pain (e.g. in the legs), fatigue, breathlessness, severe ST-segment depression or a sudden fall in systolic blood pressure. Within any particular study, the numbers of occurrences of some of these endpoints might be small, and the analysis may need to be restricted to the most common forms of stopping, with other failure types being regarded as censoring, or being combined into a catchall "other" category.

9.1.1 Model Specification

9.1.1.1 Independent Failure Types

As with other methods of survival analysis, models can be specified in terms of the hazard function, $\lambda(t)$. In the competing risks setting, each individual is at risk of a number of events, so there are several hazard functions, and

$$\{\lambda_{ij}(t): i = 1,2,...,n; j = 1,2,...,J\}$$

are the J hazard functions for the n individuals under study. One of these J types of failure could be censoring in the sense of withdrawal of consent or loss to follow up in a cohort study, so that the hazard function for being censored in the study is modelled specifically. In the case of exercise test data, it would be unlikely that a participant would withdraw consent during an exercise test, and so the failure types observed should fall into one of a small number of recognised categories.

Each hazard function defines a corresponding survivor function

$$P(T_i > t_i | \varphi_i = j) = S_{ij}(t_i) = \exp\left(- \int_0^{t_i} \lambda_{ij}(t)\, dt \right)$$

where φ_i indexes the failure type of the i^{th} individual. Since the observed failure time, T_i, is in fact the time of the first type of failure to occur amongst a mostly unobserved set of failure times,

$$T_i = \min\{T_{i1}, T_{i2}, \ldots, T_{iJ}\},$$

and if these failure times are assumed to be independent, the marginal survivor function can be written as

$$\begin{aligned} S_i(t_i) &= P(T_i > t_i) \\ &= P(T_i > t_i | \varphi_i = 1) P(T_i > t_i | \varphi_i = 2) \ldots P(T_i > t_i | \varphi_i = J) \\ &= \prod_j S_{ij}(t_i) \\ &= \prod_j \exp\left(- \int_0^{t_i} \lambda_{ij}(t)\, dt \right) \\ &= \exp\left(- \int_0^{t_i} \sum_j \{\lambda_{ij}(t)\}\, dt \right), \end{aligned}$$

so that the marginal hazard function for the i^{th} individual is

$$\lambda_i(t_i) = \sum_j \{\lambda_{ij}(t)\}.$$

9.1.1.2 Markov Process

An alternative specification of the competing risks problem is to consider it as a Markov process[130]. That is, if there are J types of failure, then at any time t, an individual can be in one of J+1 states, where state 0 indicates being "alive", and states j=1, 2, ..., J indicate failure of type j. The general Markov process can be modelled in terms of a (J+1)×(J+1) matrix of transition intensities, $\lambda(t)$, where $\lambda_{jk}(t)$ is the hazard (or transition intensity) for an individual in state j moving into state k at time t. In the competing risks setting, all individuals begin in state 0, and from there can move into any other state, but when failure has occurred, the individual cannot move out of that state.

To express this algebraically, let **P**(s,t) be the (J+1)×(J+1) transition probability matrix for the process, so that $P_{jk}(s,t)$ is the probability that an individual in state j at time s will be in state k at time t. By definition, $P_{00}(0,0)=1$. The transition intensity matrix can be written as

$$\lambda(t) = \begin{bmatrix} 1 - \sum_j \lambda_{0j}(t) & \lambda_{01}(t) & \cdots & \lambda_{0J}(t) \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix},$$

since the rows of $\lambda(t)$ must sum to 1. The overall survivor function under this framework is simply the transition probability from state 0 to state 0 between times 0 and t,

$$S(t) = P_{00}(0,t) = \exp\left(- \int_0^t \{1 - \lambda_{00}(u)\} du \right) = \exp\left(- \int_0^t \sum_j \lambda_{0j}(u) du \right),$$

which can be estimated by the Kaplan-Meier curve for any cause of failure.

It is not sensible, however, to consider the survivor functions for specific types of failure. By constructing the competing risks problem in this way, there has been no need to construct an imaginary set of survival times, all but one of which will be unobserved, and to assume that these survival times are independent, an assumption that is highly tenuous in the context of medical data. Rather than view the competing risks problem in terms of multiple survivor functions, it is more sensible to estimate the probabilities over time that individuals will suffer each of the failure types, or the cumulative incidence of each failure type.

The cumulative incidence of the j^{th} failure type can then be defined as

$$P_{0j}(0,t) = \int_0^t P_{00}(0,t)\lambda_{0j}(u) du,$$

which can be estimated from the cause-specific baseline hazard estimate, and the all-cause survivor function estimate.

9.1.1.3 Model Fitting

A set of competing risks data can be written as

$$\{T_i, \delta_{ij} : i = 1,2,...,n; j = 1,2,...,J\},$$

where $\delta_{ij}=1$ if the j^{th} type of failure is observed to occur for the i^{th} individual and $\delta_{ij}=0$ if it is not. Notice that exactly one of the δ_{ij} will be equal to 1, and all others will be zero.

The log likelihood of the data is then

$$\sum_i \sum_j - \Lambda_{ij}(t_i) + \sum_i \sum_j \delta_{ij} \log \lambda_{ij}(t_i). \qquad \textbf{(Eq. 9.1)}$$

The first term of (Eq. 9.1) reflects the fact that the i^{th} individual survives all J types of failure until time t_i, whilst the second term reflects the fact that a type j failure occurs at time t_i, and no other type of failure occurs.

The traditional approach to fitting models for independent competing risks is by separating the model into its J component parts, and fitting each submodel individually, with models for the j^{th} failure type being fitted by treating other types of failure as censored data. This is valid, and is a convenient technique given the wide availability of software for fitting survival models to censored data. Covariate information can be incorporated into the model in exactly the same way as with univariate survival data, for example through application of the Cox proportional hazards regression model. However, it is not straightforward to compare the effects of a covariate across different types of failure or to test for differences between the J baseline hazard functions.

To allow these features, the model can be fitted in its entirety by the method of data duplication[131]. The survival times and covariate data are replicated J times, so that each individual contributes J sets of 'observations', which are indexed by a variable representing the J types of failure. The event indicators for the observations are all zero, except for the one type of failure that was observed to have occurred, if any. Thus if an individual experiences none of the events considered in the study, than all of their observations are censored at the survival time. However, if an individual does suffer an event, then (s)he contributes the information that a particular type of event occurred at their survival time, and at that time all other types of event were censored.

The standard Cox model can then be fitted to these data. Separate baseline hazard functions can be fitted for each failure type, or hazard functions could be related between some or all types of failure through a constant of proportionality. Similarly, the effects of covariates can be modelled as being equal for all or some types of failures, or they could be different for each failure type. In fact, the covariates could be included in the model in such a way that different subsets of the predictor variables apply to each failure type.

Given the baseline hazard function estimates from any model, it is relatively straightforward to estimate the overall survivor function and the cause-specific cumulative incidence functions. Though standard statistical software packages are not usually designed to accommodate competing risks data, the survival analysis elements of these packages will provide the user with cause-specific estimates of survival, from which the baseline hazard function estimates can be derived.

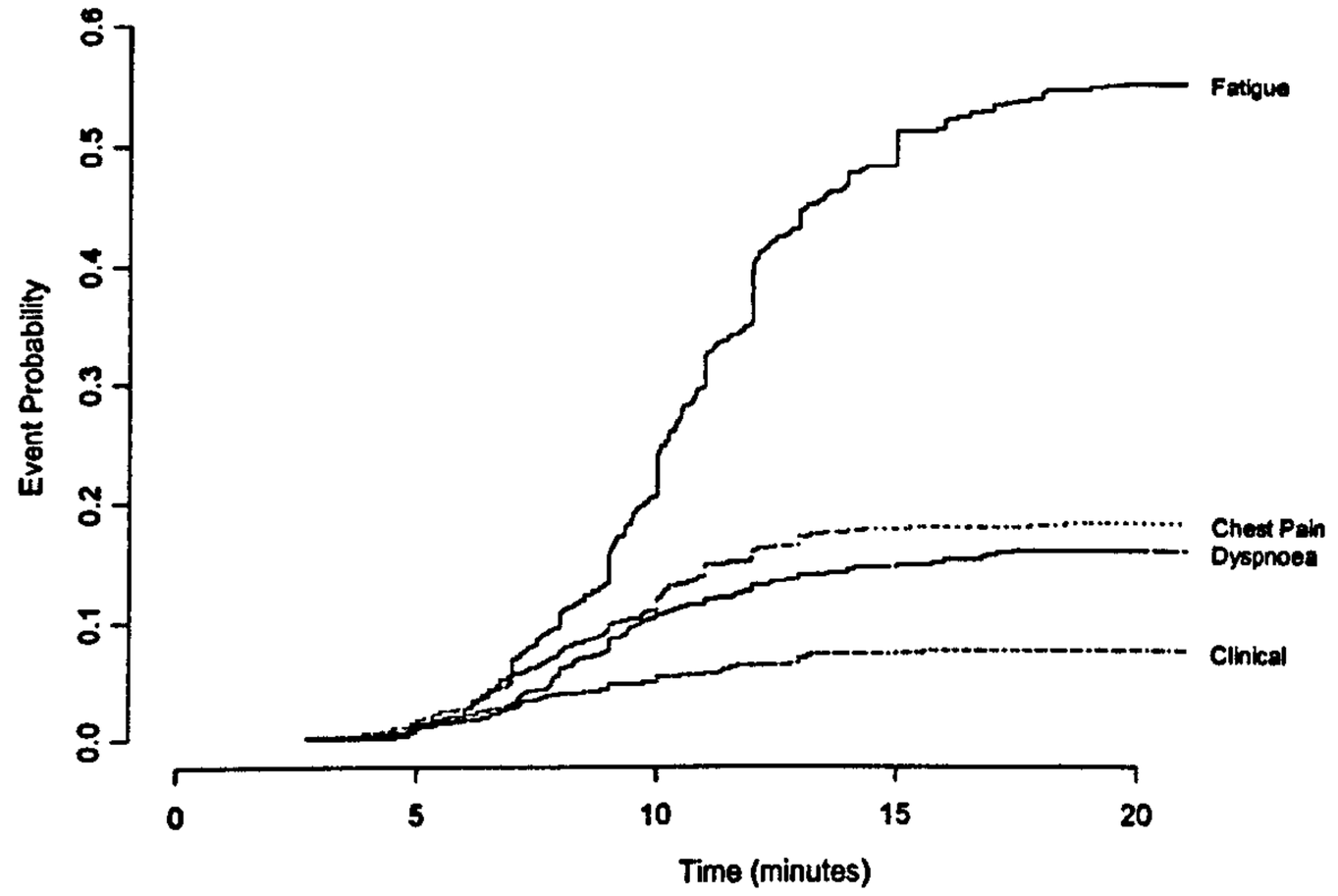


Figure 9.1 Estimated cumulative incidence functions for times to end of exercise subject to competing risks of fatigue, chest pain, dyspnoea and clinical signs (severe ST-segment depression, cardiac dysrhythmia or sudden fall in SBP).

Example 9.1 Independent Competing Risks in the TIBET Study

Six possible reasons for stopping the exercise test were recorded, and of the 611 subjects included in this analysis, 350 stopped due to fatigue, 114 to chest pain, 100 to dyspnoea, 38 to severe ST-segment depression, 2 to cardiac dysrhythmia and 7 to a sudden fall in systolic blood pressure (SBP). As there were so few tests stopped due to dysrhythmia or a fall in SBP, these were included with severe ST-segment depression since these three reasons for stopping the test were based on clinical decisions rather than the decision of the patient. The analysis modelled the total exercise time in the face of these four competing risks. The chest pain component of this model is not the same as previous analyses looking at the times to anginal pain, since many subjects that experienced anginal pain during the test continued to exercise, and did not necessarily stop exercising due to their chest pain

For this analysis, to increase the sample size, the 119 individuals for whom weight was not recorded had the predicted value based on a linear model of weight on age and sex substituted for their weight.

After duplicating the dataset as described in Section 9.1.1.2, a null model was fitted in order to determine the four cumulative incidence functions, and Figure 9.1 shows the corresponding estimates. Individuals are clearly most likely to withdraw from exercise due to fatigue, with more than 50% of tests ending in this way. There are

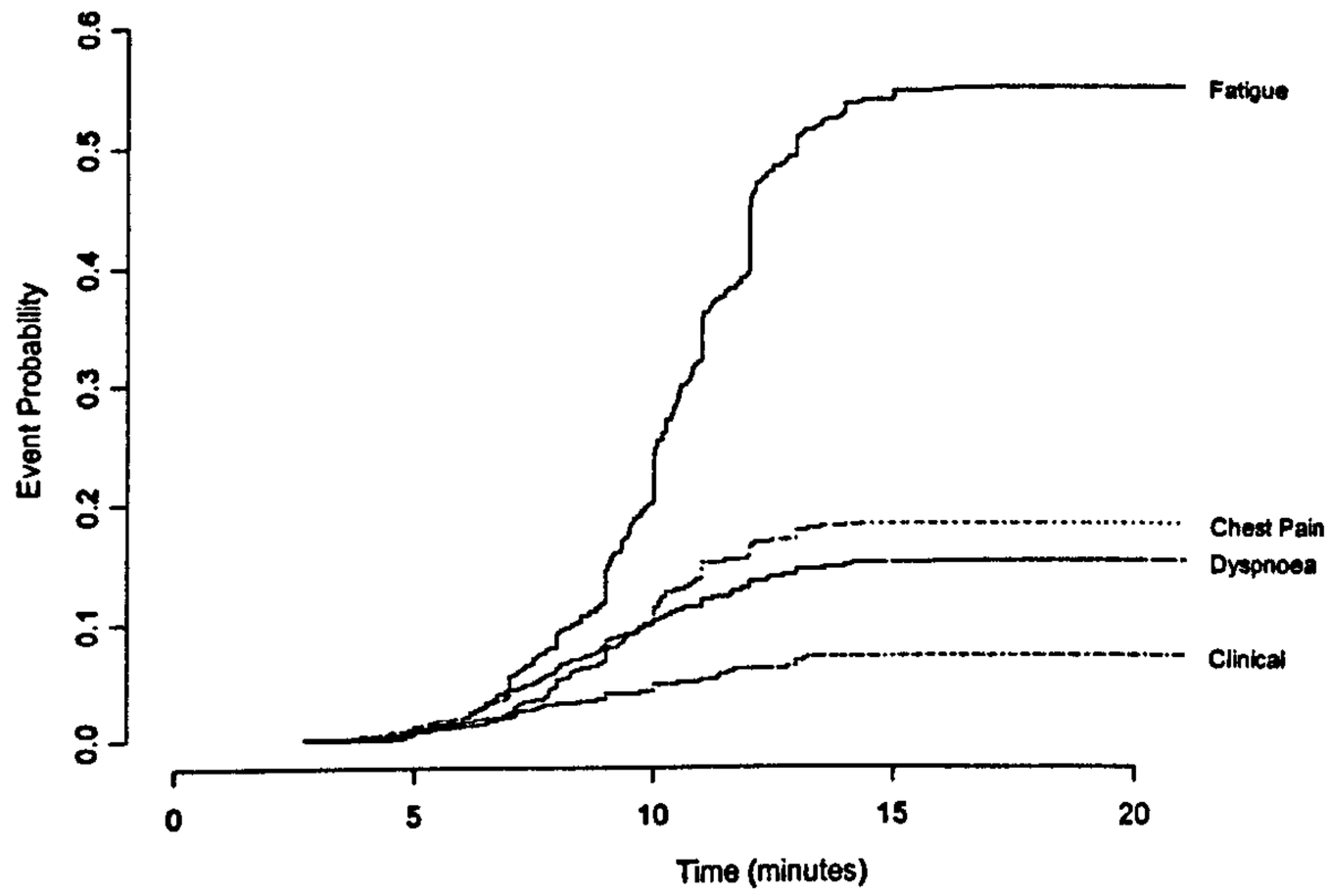


Figure 9.2 Estimated cumulative incidence functions for times to end of exercise subject to competing risks of fatigue, chest pain, dyspnoea and clinical signs adjusted for mode of exercise, age, gender and weight.

visible "jumps" in the incidence of fatigue at one-minute intervals, particularly over the range from 7 to 15 minutes.

Covariate effects were then added for exercise type, age, sex and weight. Exercise type was included as a proportional hazards effect, rather than as a stratifying variable, to reduce the number of strata in the model and allow a simpler presentation of the results. Assuming these covariates to have equal effects on the four hazard functions, all these variables were found to have significant effects, with shorter exercise times being experienced by those exercising on a treadmill, by older patients and by women. Weight was found to have differential effects dependent upon the type of exercise being performed; with treadmill exercise greater weight was associated with shorter exercise times, whilst on a bicycle heavier patients exercised for longer.

Figure 9.2 shows the estimated cumulative incidence functions from this model. The curves are derived from the baseline hazard function increments as estimated in a stratified Cox proportional hazards model. These functions are the estimated curves for imaginary individuals whose covariate values are equal to the population mean values. Compared to Figure 9.1, the curves have very similar limiting values. Just visible in the curve for fatigue, the predominant reason for stopping, is an increased curvature in the cumulative incidence curve, with lower incidence early on, but an increased probability of withdrawal over the second part of the timescale. This is a result of the model identifying those most likely to suffer an event and "explaining" early withdrawals and

	Hazard Ratio	95% CI	p-value	Heterogeneity Test p-value
Exercise Type (Bicycle : Treadmill)	0.24	(0.20, 0.29)	<0.0001	0.0086
Age (/10 years)	1.49	(1.33, 1.67)	<0.0001	0.69
Gender (Female : Male)	2.93	(2.25, 3.82)	<0.0001	0.0015
Weight (Treadmill) (/10 kg)	1.19	(1.03, 1.37)	0.016	0.69
Weight (Bicycle) (/10 kg)	0.72	(0.63, 0.81)	<0.0001	0.61

Table 9.1 Covariate effect estimates, with 95% CIs and p-values, from competing risks model of total exercise time assuming effects are equal across the four reasons for stopping exercise. Also shown is p-value for test of heterogeneity of effects upon different failure types.

long survivals. The figure is constrained to have the same limiting incidence levels, since the curves correspond to the average participant. The effect estimates from this model are listed in Table 9.1. Also shown are the p-values from likelihood ratio tests comparing this model to the extensions allowing each covariate to have differential effects on the four endpoints. There is good evidence that exercise type and gender have different effects on the incidence of stopping the test for different reasons.

A more complex model was then fitted, allowing exercise type and gender to show heterogeneity between the four competing risks. The resulting estimates of the cumulative incidence functions are shown in Figure 9.3, and the effect estimates from this model are shown in Table 9.2. It was not possible to estimate an effect of gender on the time to withdrawal from exercise due to clinical indications, since all 47 of these withdrawals including 38 due to severe ST-segment depression, were experienced by men.

Gender had other drastic influences on the reasons for stopping the tests, with hazard ratios between women and men of 4 and 4.5 for stopping due to fatigue or dyspnoea. The hazard ratio for stopping due to chest pain was not seen to be associated with gender, so that whilst men and women will stop testing due to angina at roughly equal rates, women will withdraw due to non-anginal reasons at much higher rates, so there will be more withdrawals due to angina amongst men; in the data as a whole, 20% of men and 12% of women stopped exercising for this reason.

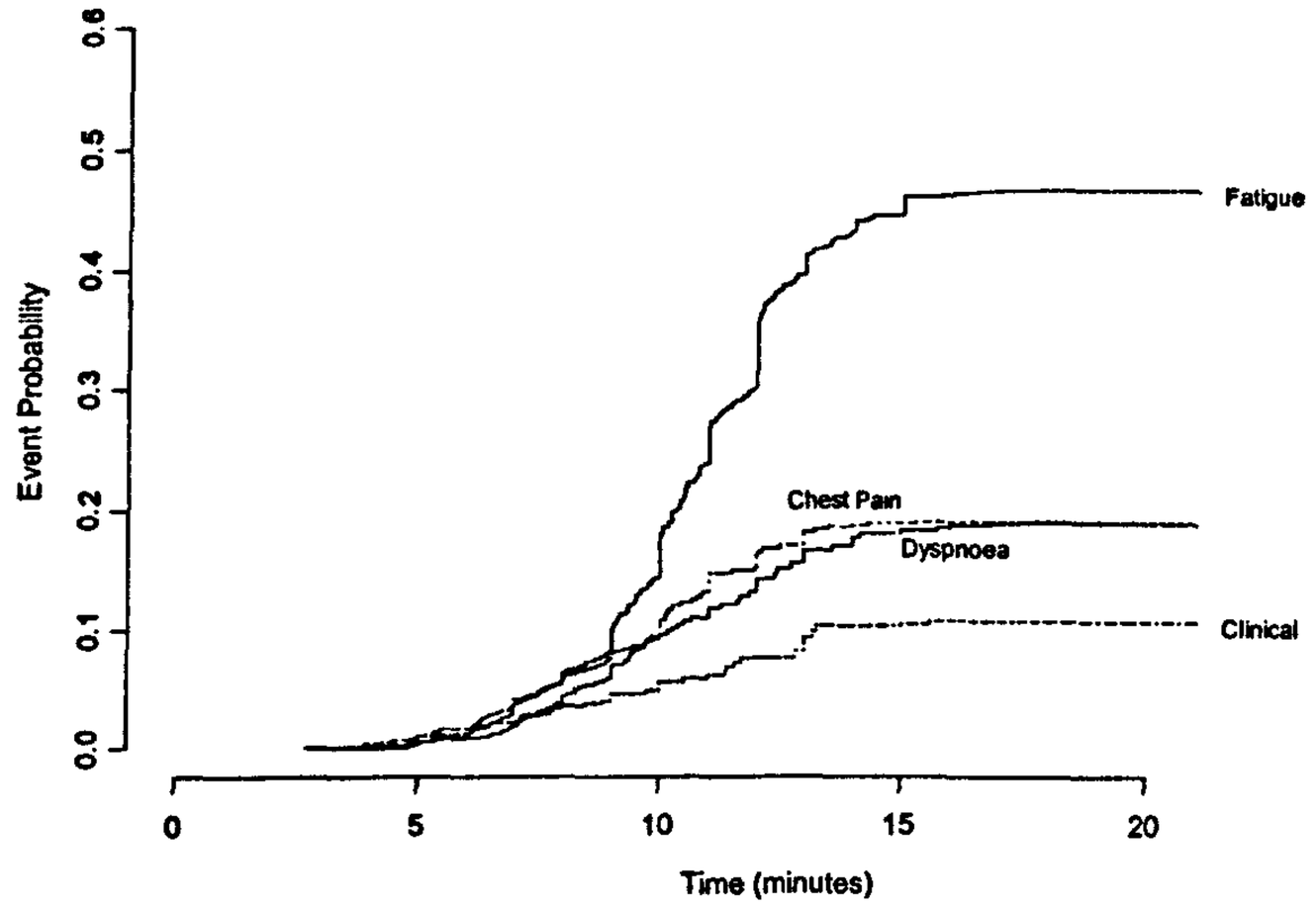


Figure 9.3 Estimated cumulative incidence functions for times to end of exercise subject to competing risks of fatigue, chest pain, dyspnoea and clinical signs adjusted for cause-specific effects of mode of exercise, age, gender and weight.

Exercise using a bicycle was associated with much lower hazard for all endpoints, as expected, though whilst the hazard ratio for ending exercise due to fatigue was 0.32 between bicycle and treadmill exercise, it was only 0.24 for ending exercise due to chest pain and as low as 0.1 that exercise should end due to dyspnoea or be interrupted due to clinical indications. Thus, though the hazard for all endpoints is lower under bicycle exercise, fatigue will be a relatively more common reason for stopping the test. This is borne out in the raw figures, since 70% of bicycle tests end in fatigue, compared to 45% of treadmill tests. Likewise, 33% of treadmill tests were stopped due to anginal pain or adverse clinical signs, compared to only 19% of bicycle tests.

There was no evidence of heterogeneity of the effects of age or weight between different endpoints. A 10-year difference in age was associated with 50% greater hazard for all endpoints. Under treadmill exercise, a 10 kg increase in weight was associateed with nearly 20% greater hazard for ending a test, though during bicycle exercise the same weight difference offered a 28% reduction in hazard.

Finally, treatment effects were considered. Overall, there was no evidence that treatment had any influence on times to first events (likelihood ratio test statistic, 2.5 on 2 df, p=0.29). When the effects of treatment on each endpoint were considered individually, there was some evidence that treatment group was associated with time until stopping exercise due to fatigue (likelihood ratio test statistic, 6.3 on 2 df, p=0.044), with the suggestion that those using Combination therapy were most likely to

	Reason for Stopping Exercise Test			
	Fatigue	Chest Pain	Dyspnoea	Clinical
	Haz. Ratio (95% CI) p	Haz. Ratio (95% CI) p	Haz. Ratio (95% CI) p	Haz. Ratio (95% CI) p
Exercise Type (Bicycle : Treadmill)	0.32 (0.25, 0.41) <0.0001	0.24 (0.16, 0.36) <0.0001	0.11 (0.07, 0.18) <0.0001	0.10 (0.05, 0.22) <0.0001
Age (/10 years)	1.50 (1.34, 1.67) <0.0001			
Gender (Female : Male)	3.99 (2.90, 5.51) <0.0001	1.42 (0.74, 2.71) 0.29	4.46 (2.61, 7.64) <0.0001	-
Weight (Treadmill) (/10 kg)	1.19 (1.03, 1.36) 0.017			
Weight (Bicycle) (/10 kg)	0.72 (0.64, 0.82) <0.0001			

Table 9.2 Covariate effect estimates, with 95% CIs and p-values, from competing risks model of total exercise time adjusted for cause-specific effects of mode of exercise, age, gender and weight.

withdraw from the test for this reason; Atenolol : Combination hazard ratio estimate, 0.76 (95% CI 0.58, 0.98), p=0.034; Nifedipine : Combination estimate 0.62 (0.47, 0.81), p=0.0005. However, since this is an effect of borderline significance found amongst four possible effects, without evidence of a global treatment effect, it must be viewed with some scepticism.

9.1.2 Dependent Failure Types

The competing risks model was derived in two ways, firstly by assuming the existence of a set of independent survival times, only one of which is observed, and then by constructing the model within the framework of a Markov model, which does not depend upon these assumptions.

Within the first conceptual framework, it would be natural to assume that times to failure from different causes within the same individual would be correlated. There are also clear parallels between the multivariate survival models explored in Section 7.4 and the way that the data are replicated to facilitate the fitting of competing risks models, producing multiple "observations" of survival data in the same individual. An obvious approach to modelling dependent competing risks data would therefore be to employ the methods for marginal and frailty models of correlated multivariate survival data.

There are, however, crucial differences between competing risks data and multivariate survival data. With competing risks data all survival times will be equal, and there will be precisely one event observed for each individual, with all other "observations" on that individual censored at the same time. With multivariate survival data, survival times will in general be different, and within any cluster of observations, any number of events could occur.

The fact that the same specification for a competing risks model is reached from two different perspectives, one of which assumes independent survival times whilst the other does not, indicates the fundamental problem with modelling dependent survival times in a competing risks setting. For any distribution of dependent survival times, it is possible to construct an independent distribution of survival times with the same marginal distribution for the time to the first event[132]. Thus the competing risks problem when constructed in terms of a set of dependent survival times is inherently unidentifiable.

9.2 Haemodynamic and Electrocardiographic Response

In a clinical setting, when exercise tests are used to evaluate cardiac patients, the physician will evaluate the total response of the patient to the test. This will involve monitoring the patient for signs of discomfort or fatigue, as well as their heart rate, blood pressure and electrocardiographic response. Whilst some of these will be subjective observations, interpreted based on experience and knowledge of the patient, some of these are objective measurements that are recorded as part of a standard exercise protocol.

Standard analyses of these measures often involve drastic reduction of the data, so that ECG response is summarised as the time until the first recorded occurrence of ≥1mm ST-segment depression, or heart rate response as the time until the patient reaches their age-, sex- and weight-predicted maximal heart rate, or a percentage thereof. A more detailed analysis of these responses to exercise could be achieved by considering a complete set of observations as repeated measures data. Random effects models could be applied to study changes in outcomes throughout the exercise test, either one outcome at a time, or as a group of correlated responses.

9.2.1 Repeated Measures

Repeated measures data can arise in medical studies for a number of reasons. It may be a planned part of a clinical trial to measure participant outcomes at a number of time points in order to demonstrate treatment efficacy over a specified length of time. In studies of pharmacokinetics the concentrations of drug metabolites will be measured on several occasions after taking a preparation in order to model the dispersion of the agent in the body. In retrospective studies involving case note review, the outcomes of a number of contacts with the health service might be recorded.

In clinical trials that involve exercise tests, levels of ST-segment depression, heart rate and blood pressure will be recorded on several occasions. This is partly to determine pre-specified endpoints such as the time until ≥1mm ST-segment depression or the heart rate at maximal exercise. There is also an element of safety monitoring, since a subject will be withdrawn from the test upon signs of severe ST-segment depression, dysrhythmia or a sudden fall in systolic blood pressure. Thus the fact that a number of data series are recorded is incidental to the study, since a per-protocol analysis might only investigate specific parts of the data. However, the fact that repeated measurements of the cardiac response have been recorded under increasing levels of exercise allows an opportunity for further analysis.

The distinguishing feature of repeated measures data is that the same measurement is made on several occasions on the same individual. In general, this will allow more precise estimation of within- and between-subject variability, and therefore more precise estimation of associations between patient outcomes and exposures of interest. If the measurements are made under changing conditions, such as increasing levels of exercise during an exercise test protocol, this allows more precise estimation of individual responses to these conditions.

9.2.2 Within-Subject Correlation

The main analytical problem with repeated measures data is that multiple measures of the same quantity made on several occasions in the same individual will be correlated. In order to gain the full benefit of repeated measures data and to obtain improved precision of estimates of between-subject differences will require the best possible estimation of between- and within-subject variability. This will entail the modelling of the within-subject correlation structure.

It would be expected that responses measured on the same subject would be more closely related than those measured on different subjects, so that there would be a positive correlation between repeated measurements. Furthermore, when measurements are made at short intervals such as during an exercise test, it would seem likely that measurements made at consecutive time points would be more closely related than those made several minutes apart. These two aspects of the correlation structure should be incorporated into the analysis.

Example 9.2 Repeated Measurements of ST-Segment Depression and Heart Rate during Exercise in the TIBET Study

To illustrate the application of repeated measures analysis to exercise test data, the following example models ST-segment depression and heart rate (HR) increases during exercise. The response variables used in this analysis were the changes from baseline ST-segment and HR levels (prior to the start of exercise), with the sign of ST-segment changes reversed, so that depression relative to baseline was expressed as positive values. For simplicity, the analysis was restricted to those subjects that performed exercise using a treadmill. The analysis models both responses simultaneously, allowing for the likely correlation between the two responses, the correlation between repeated measurements of responses in the same individual, and the autocorrelation between responses over time.

The time course of both response variables was modelled as linear for the purpose of this analysis, though in principle non-linear associations could be fitted. The time trend in both responses was also be considered as a random effect, allowing for the possibility that the response to exercise of one or both outcomes may differ between individuals. The effects of age, gender, weight and treatment were estimated as fixed effects. In other words,

$$\left(\Delta ST_{i1},\ldots,\Delta ST_{iJ_i},\Delta HR_{i1},\ldots,\Delta HR_{iJ_i}\right)^T=\left(X_i,X_i\right)\begin{pmatrix}\beta_{i1}\\ \beta_{i2}\end{pmatrix}+\begin{pmatrix}\varepsilon_{i1}\\ \varepsilon_{i2}\end{pmatrix}$$

where ΔST_{ij} and ΔHR_{ij} are the response variables and X_i is the design matrix for subject i, including a column of 1s to model the intercept and a column taking values $(1,2,\ldots,J_i)$ to model the dependence of each response on the time spent exercising. Notice that the intercept term in this model corresponds to the change from baseline ST-segment depression or HR when time is 0, which when viewed as the immediate response to starting exercise, need not necessarily be zero. The residual vectors ε_{i1} and ε_{i2} are i.i.d.

$N(0,\sigma_1^2)$ and $N(0,\sigma_2^2)$ respectively. The regression coefficients were also, in general, considered to be random variables

$$\begin{pmatrix}\beta_{i1}\\ \beta_{i2}\end{pmatrix} \sim N\left(\begin{bmatrix}b_1\\ b_2\end{bmatrix}, \begin{bmatrix}\Sigma_{11} & \Sigma_{21}\\ \Sigma_{12} & \Sigma_{22}\end{bmatrix}\right).$$

The model was fitted as a multilevel model, using the statistical package MLwiN[133]. The model was fitted over three levels, with individual being the highest level, time point the middle level and type of response (ST-segment depression or increase in HR) the lowest level. The lowest level in the hierarchy is included to allow the model to estimate correlations between the two types of response, even though none of the effects included in the models were considered as random at this level.

The data are arranged as a single vector of response variables. By including indicator variables for the two types of response variable and separate variables for the time trends of each response as random effects at the individual level, the model fits separate intercepts and slopes for each individual for the associations between response variables and the time spent exercising. These intercepts and slopes are random in the population and may be correlated amongst themselves. The variation in the intercept coefficients between time points can be interpreted as residual variation. Residuals for the ΔST and ΔHR response variables may be correlated. Thus the multilevel structure of the model implies a general correlation between repeated measures on the same individual, without directly specifying a given correlation structure.

To allow for serial correlation between responses, fixed effects were included of the previous pair of measurements, so that the last recording of ST-segment depression was seen to influence the current level of ST-segment depression, and similarly for the previous measurement of HR. The coefficients for these terms can be interpreted as estimates of a first-order autocorrelation, so that the correlation between pairs of measurements decays as the distance between them (in time) becomes larger.

Table 9.3 shows the effect estimates obtained from fitting a model with these effects as the only terms in the model. Significant variation was found between individuals in terms of their ST-segment depression and HR intercepts, and in their gradients over time in these variables.

For example, changes from baseline ST-segment depression levels were seen to have a mean intercept of –0.12 mm, so that immediately upon starting exercise, ST-segment levels become elevated by an average of 0.12 mm. These intercepts were not

Estimate (95% CI)	ST-segment depression (mm)		Heart rate (b/min)	
	Intercept	Slope	Intercept	Slope
Mean	-0.12 (-0.17, -0.07)	0.14 (0.12, 0.16)	15.3 (13.8, 16.8)	3.9 (3.6, 4.1)
Standard deviations				
Between-subject (population variation)	0.19 (0.13, 0.24)	0.10 (0.08, 0.11)	10.4 (9.3, 11.5)	1.3 (1.2, 1.5)
Within-subject (residual variation)	0.36 (0.34, 0.37)	-	4.4 (4.3, 4.6)	-
Repeated measures correlations				
ST-segment depression	0.23 (0.13, 0.32)	-	0.35 (0.27, 0.44)	-
Heart rate	0.35 (0.27, 0.44)	-	0.85 (0.82, 0.88)	-0.25 (-0.37, -0.13)
Autocorrelations	0.41 (0.36, 0.46)	-	0.13 (0.09, 0.16)	-

Table 9.3 Random effects estimates from initial model of repeated measurements of ST-segment depression and heart rate during treadmill exercise, allowing for subject-specific average levels and reactions to exercise, as well as first-order serial correlation between responses

the same for every individual, with a standard deviation of 0.19 mm (95% CI 0.13 – 0.24 mm) over the population, so that a 95% range in this population for the initial response to exercise of the ST-segment could be estimated as between 0.5 mm of elevation and 0.26 mm of depression. After an initial response, ST-segment depression was seen to increase by on average 0.14 mm/min, but this slope varied between individuals with a standard deviation of 0.10 mm/min (0.08 – 0.11 mm/min).

Based on the estimated between- and within-subject variance-covariance matrices, estimates of the correlations between repeated measures of each response can be derived (i.e. correlations between responses without taking account of autocorrelation). The variance-covariance matrix for these correlations can be estimated by the delta method, based on the variance-covariance matrix of the variance components. The autocorrelation between successive responses is estimated from the estimate of the regression coefficient for the previous response.

Correlation between repeated measures of ST-segment depression has a strong serial component, though correlation between measurements more than 2 minutes apart would seem to be influenced mainly by individual effects. Correlation between repeated HR measurements was mainly a result of variation between individuals, though there

Estimate (95% CI)		ST-segment depression		Heart rate	
		Intercept (mm)	Slope (mm/min)	Intercept (b/min)	Slope (b/min/min)
Mean		-0.32 (-0.45, -0.19)	0.15 (0.13, 0.16)	8.5 (4.6, 12.4)	3.8 (3.6, 4.1)
Standard deviations					
	Between-subject	0.17 (0.11, 0.22)	0.10 (0.09, 0.12)	10.2 (9.1, 11.2)	1.3 (1.2, 1.5)
	Within-subject	0.35 (0.34, 0.36)	-	4.4 (4.3, 4.6)	-
Repeated measures correlations					
	ST-segment depression	0.20 (0.10, 0.30)	-	0.28 (0.19, 0.37)	-
	Heart rate	0.28 (0.19, 0.37)	0.15 (0.01, 0.28)	0.84 (0.81, 0.87)	-0.30 (-0.42, -0.18)
Autocorrelations		0.38 (0.33, 0.42)	-	0.13 (0.10, 0.16)	-

Table 9.4 Random effects estimates from final model of repeated measurements of ST-segment depression and heart rate during treadmill exercise, allowing for subject-specific average levels and reactions to exercise, first-order serial correlation and fixed effects of age, gender, weight and treatment

was a correlation of –0.25 between underlying HR and the HR reaction to exercise, such that individuals with less of an initial increase in HR upon starting exercise would tend to show greater increases in HR during exercise.

This model was then extended to include fixed effects for the age, gender and weight of each individual, as well as for differences between the three treatment groups. Table 9.4 shows the random effects estimates from this model. The random effects are similar to those obtained from the initial model, though there is some evidence of a positive correlation between the initial HR response to exercise and the gradient of ST-segment depression with time (estimated correlation 0.15; 95% CI, 0.01 – 0.28). The positive correlation might indicate that less physically fit patients (with greater initial reactions of HR to exercise) tend to have reduced cardiac fitness (greater increase in ST-segment depression with continuing exercise).

Table 9.5 lists the fixed effects estimates from this model. Age was not seen to be associated with either ST-segment depression or HR. Female gender was associated with initial reactions to exercise that were 0.18 mm greater for ST-segment depression (95% CI 0.04, 0.32 mm) and 5.9 b/min higher for HR (1.8, 10.1 b/min). Weight was

Estimate (95% CI)	ST-segment depression (mm)	Heart Rate (b/min)
Covariate effects		
Age (/10 years)	0.01 (-0.05, 0.07)	0.3 (-1.5, 2.0)
Gender (Female – Male)	0.18 (0.04, 0.32)	5.9 (1.8, 10.1)
Weight (/10 kg)	0.03 (-0.02, 0.07)	1.9 (0.6, 3.4)
Treatment effects		
Atenolol – Combination	0.11 (0.00, 0.21)	1.2 (-2.0, 4.5)
Nifedipine – Combination	0.21 (0.11, 0.32)	5.5 (2.3, 8.8)

Table 9.5 Estimated fixed effects of age, gender, weight and treatment from the final model of repeated measurements of ST-segment depression and heart rate during treadmill exercise, allowing for subject-specific average levels and reactions to exercise, and first-order serial correlation

seen to affect HR but not ST-segment depression, with a 10 kg greater body weight being associated with an increased initial reaction to exercise of 1.9 b/min (0.6, 3.4 b/min).

Relative to those on combination therapy, there was some evidence that the initial response of ST-segment depression is 0.11 mm (0.00, 0.21 mm) greater for those treated with atenolol, but there was no evidence that the initial change from resting HR was affected. Those treated with nifedipine, however, showed clearly higher changes from rest relative to the combination therapy group, with the initial changes in ST-segment depression and HR increased by 0.21 mm (0.11, 0.32 mm) and 5.5 b/min (2.3, 8.8 b/min) respectively.

9.3 Summary

The methods outlined in this section have gone beyond what is often done in practice, where the data are reduced to the time until a pre-specified event occurs. Attempts have been made to consider a broader picture, and to model the effects of covariate and treatment information within more complex frameworks.

The competing risks model investigated the effects of variables on total exercise time, where exercise could be stopped due to a number of reasons. In a standard

survival analysis model, the assumption of non-informative censoring might be violated, and observed relationships between treatment groups and the time to an ischaemic event could arise due to an imbalance in the likelihood of stopping exercise for some other reason. The competing risks method attempts to disentangle the many causes of an exercise test stopping, in order to determine the true nature of covariate effects.

The application of methods for the analysis of repeated measures is a broad and complex subject, and the example shown in Section 9.2 was undoubtedly over-simplified. The correlation structure employed was basic, and determined to a large degree by the limitations of the software used. However, by differentiating between effects that act on the initial and continuing responses to exercise, it provided important insights into the mechanism of treatment and covariate effects.

Repeated measures data are naturally prone to missing values, particularly in this case where the previous observation was used as an independent variable, in order to model the autocorrelation in the data. For this model, observations with missing data were excluded, so that a single missing response would result in two missing observations. The use of imputation techniques might well give improved performance with this model.

Further improvements could be made by the use of more appropriate covariance structures, by the inclusion of information regarding systolic blood pressure, or by incorporating non-linear associations between response variables and time. For instance, the time course of ST-segment depression could well be non-linear, with the gradient increasing when ischaemia begins. Also, with continuing improvements in computing performance and statistical software, it would seem likely that models for analysing continuously recorded ECG output and other responses during exercise could be developed.

CHAPTER 10 Concluding Remarks

This thesis has addressed the statistical analysis of exercise test data, with particular attention paid to the analysis of times spent exercising until the occurrence of ischaemic events such as anginal pain or the first observation of ≥1mm ST-segment depression. Various techniques have been applied to data from the Total Ischaemic Burden European Trial (TIBET) and their results compared. Simulation studies have been used to evaluate the performance of some of these methods with respect to particular features of exercise test data, namely interval censoring and repeated survival times. Some novel approaches to analysing exercise times and other data that are accrued during exercise tests have also been explored.

To conclude, this Chapter will briefly discuss the findings of this work, and make recommendations for which methods should be used to analyse exercise test data in practice.

10.1 Survival Analysis

In clinical trials of anti-anginal therapies, it is a regulatory requirement that exercise testing be employed to evaluate patients[48]. The main outcome variables from exercise tests used in this setting are the times spent exercising until the occurrence of anginal chest pain or significant ischaemia as determined by depression of the ST-segment of the ECG trace. These variables can be censored, though in practice, analyses are carried out that ignore this fact, despite the observation that the failure to use appropriate statistical methods could be inefficient or misleading[52].

The fact that inappropriate methods continue to be used may reflect a lack of awareness of the censored nature of these data or an intransigence amongst those conducting or publishing the results of these studies to accept novel analytical approaches. This is not a criticism; since methods for uncensored data have traditionally been used, it is natural for researchers to adopt similar methods for analysis, as it is natural that reviewers of journal articles would question the use of alternative methods.

In most cases, the use of survival analysis as opposed to an alternative method that treats the data as uncensored will not affect the conclusions of a study, and those who attempt to use more appropriate statistics may find that the gain from being more correct in their analysis is outweighed by the difficulty in presenting the results to clinical colleagues and those who would publish their findings.

From the standpoint of a statistician, it is preferable to use survival techniques to analyse exercise times, despite the difficulties in presenting this to non-statistical collaborators. Many analyses involve a trade-off between statistical correctness and ease of comprehension by the target audience. However, part of the responsibility of the statistician is to push the boundary of this compromise, by finding better ways of presenting data and of explaining the methods used. For example, in addition to reporting hazard ratios associated with particular treatments, estimated differences in median exercise time could be reported from a Cox proportional hazards model. In Section 4.1, it was found that the time to anginal pain could be well represented by a Weibull distribution, and it would be relatively straightforward to report differences in mean exercise time based on such a parametric model. In this way a more appropriate method is used to analyse the data, whilst the results are presented in a format familiar to others.

10.2 Interval Censoring

The time until the occurrence of significant ST-segment depression will often be restricted to take one of a set of distinct values, since the ECG trace is recorded at intervals only. For an observed occurrence, the actual time of the first occurrence is known only to lie in the interval since the previous recording.

Methods exist for the analysis of such interval censored data, though the results of simulations presented in Section 6.3 would suggest that it is the use of survival methods that is most important for the validity of the analysis. Standard methods for survival analysis, rather than methods designed for this type of data perform sufficiently well, and would be recommended for use in practice.

10.3 Repeated Exercise Times

The fact that exercise tests may be repeated is often made advantage of in clinical studies. Baseline exercise tests are performed, and study results reported in terms of the mean change in exercise times. This would be expected to improve the power of a study

to detect treatment effect differences. Also, in studies where it may be unethical to include a placebo arm, since the treatments being compared have previously been shown to be of benefit, the use of baseline data allows improvements in exercise tolerance to be reported for each intervention, if only to quantify the scale of treatment effects.

The use of survival analysis methods with repeated data is a relatively new field, and until recently not supported by standard statistical software packages. The results of the simulation study presented in Section 8.3 indicate that when exercise times follow a frailty model, there can be significant advantages to using this model for the analysis. It is not clear, however, whether the frailty model performs as well when this model is a misspecification of the underlying process. Nonetheless, for the analysis of repeated exercise times in practice, such models would appear to offer the best initial option, though the use of a fully parametric baseline hazard function, such as a Weibull distribution, might allow greater flexibility in terms of reporting the results.

10.4 Other Methods

Competing risks analysis of exercise times can be used to assess the effects of treatments and other covariates on the multiple potential reasons for stopping an exercise test, and can be used to gain additional insight into the process of maximal exercising. However, the primary endpoints of an exercise test would be the first occurrences of chest pain or significant ST-segment depression, so such an analysis would most likely form part of a secondary analysis.

The use of mixed effects models to analyse the wealth of ECG and haemodynamic data that are collected could be used as part of a main analysis of exercise test data. The levels of these variables at specific points, such as at maximal exercise or at the occurrence of an ischaemic event, are often reported alongside exercise times. The analysis shown in Section 9.2 is somewhat basic, but this type of analysis offers much potential for extracting important information from studies using exercise tests. To use such methods in practice would require both development of the models and their interpretation. In the same way as the use of survival methods for the analysis of exercise times, there would need to be a considerable effort to persuade the clinical audience for the results the method added value to an analysis. However, there would perhaps be less resistance to accepting these mixed effects models for data that have in general not been analysed in detail before, than to the use of survival models for

data that have traditionally been analysed by other methods. Also, since the results can be viewed in terms of mean changes in outcomes over time, it might be easier for the target audience to interpret the results of such an analysis.

10.5 Conclusion

In the analysis of exercise test data, is has been recognised that times spent exercising before the occurrence of ischaemic events should be analysed using methods appropriate for survival data, though these methods do not appear to have been widely adopted in clinical trials of anti-anginal therapies. Given the wealth of electrocardiographic and haemodynamic data collected during exercise tests, there is scope for the development of more complex mixed effects models that would provide additional insight into factors influencing exercise tolerance.

Appendix A Parametric Form for Difference in Survival Times

Section 7.3 derived the likelihood function (Eq. 7.3) for a model in which the differences between pairs of survival times were assumed to follow a Normal distribution. If the data are written as

$$\{T_{i1}, T_{i2}, \delta_{i1}, \delta_{i2}, \mathbf{z}_i : i = 1,2,...,n\}$$

where the Ts are observed survival times, the δs are failure indicators and $\mathbf{z}_i$ is a vector of covariates for the i^{th} individual. By defining the difference in observed exercise times to be $\Delta_i = T_{i2} - T_{i1}$, and

$$u_i = \frac{\Delta_i - \mathbf{z}_i\beta}{\sigma},$$

the contribution to the log likelihood for the i^{th} individual is

$$\begin{aligned} l_i &= \delta_{i1}\delta_{i2}\log\phi(u_i) + \delta_{i1}(1-\delta_{i2})[1-\log\Phi(u_i)] + (1-\delta_{i1})\delta_{i2}\log\Phi(u_i) - \delta_{i1}\delta_{i2}\log\sigma \\ &= A_i - \delta_{i1}\delta_{i2}\log\sigma \end{aligned},$$

where ϕ and Φ are the p.d.f. and c.d.f. of a Standard Normal distribution, respectively and the negative log likelihood function can be written as

$$-l(\beta, \sigma | \Delta_i, \delta_{i1}, \delta_{i2}) = -\sum_i l_i = -\sum_i \{A_i - \delta_{i1}\delta_{i2}\log\sigma\}.$$

Maximisation of the likelihood function, in order to find estimates for the model parameters, can be achieved using the statistical software package S-Plus[134] using the minimisation function **nlminb**. For optimal performance, the **nlminb** function requires code to calculate both the first and second derivatives of the negative log likelihood with respect to the model parameters. The matrix inverse of the second derivative of the negative log likelihood evaluated at the minimum then provides an estimate of the variance-covariance matrix of the parameter estimates.

Some useful identities when evaluating the derivatives of the negative log likelihood are

$$\frac{\partial}{\partial u_i}\phi(u_i) = -u_i\phi(u_i),\ \frac{\partial}{\partial u_i}\Phi(u_i) = \phi(u_i),\ \frac{\partial}{\partial \beta_j}u_i = -\frac{z_{ij}}{\sigma} \text{ and } \frac{\partial}{\partial \sigma}u_i = -\frac{u_i}{\sigma},$$

so that

$$\frac{\partial}{\partial \beta_j}\phi(u_i) = \frac{z_j u_i}{\sigma}\phi(u_i),\ \frac{\partial}{\partial \sigma}\phi(u_i) = \frac{u_i^{\,2}}{\sigma}\phi(u_i),$$

$$\frac{\partial}{\partial \beta_j}\Phi(u_i) = -\frac{z_j}{\sigma}\phi(u_i) \text{ and } \frac{\partial}{\partial \sigma}\Phi(u_i) = -\frac{u_i}{\sigma}\phi(u_i).$$

Defining

$$B_i = \frac{\partial}{\partial u_i}A_i$$

$$= -\delta_{i1}\delta_{i2}u_i - \delta_{i1}(1-\delta_{i2})\frac{\Phi(u_i)}{1-\Phi(u_i)} + (1-\delta_{i1})\delta_{i2}\frac{\Phi(u_i)}{\Phi(u_i)}$$

and

$$C_i = \frac{\partial}{\partial u_i}B_i$$

$$= -\delta_{i1}\delta_{i2} - \delta_{i1}(1-\delta_{i2})\Phi(u_i)\frac{\Phi(u_i) - u_i[1-\Phi(u_i)]}{[1-\Phi(u_i)]^2} - (1-\delta_{i1})\delta_{i2}\Phi(u_i)\frac{\Phi(u_i) - u_i\Phi(u_i)}{\Phi(u_i)^2},$$

the first and second derivatives of the negative log likelihood can be written as

$$-\frac{\partial l}{\partial \beta_j} = -\sum_i -\frac{z_{ij}}{\sigma}A_i,$$

$$-\frac{\partial l}{\partial \sigma} = -\sum_i -\frac{1}{\sigma}\{u_i A_i + \delta_{i1}\delta_{i2}\},$$

$$-\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} = -\sum_i \frac{z_{ij}z_{ik}}{\sigma^2}B_i,$$

$$-\frac{\partial^2 l}{\partial \beta_j \partial \sigma} = -\sum_i \frac{z_{ij}}{\sigma^2}\{A_i + u_i B_i\},$$

$$-\frac{\partial^2 l}{\partial \sigma^2} = -\sum_i \frac{1}{\sigma^2}\{2u_i A_i + u_i^{\,2}B_i + \delta_{i1}\delta_{i2}\}.$$

Appendix B Gamma Frailty Model with Weibull Baseline Hazard

Example 7.4 derives the log likelihood for a Gamma frailty survival regression model with a Weibull baseline hazard function as a sum over all observations of the form (in terms of the negative log likelihood, ignoring subscripts and summation signs)

$$-l(\theta,\beta,\alpha)=\left(\frac{1}{\theta}+\delta\right)\log\left(1+\theta t^{\alpha}\exp(z\beta)\right)-\delta\left(\log\alpha+(\alpha-1)\log t+z\beta\right),$$

so that the first derivatives can be written as

$$-\frac{\partial l}{\partial\beta_j}=\left(\frac{1}{\theta}+\delta\right)\frac{\theta t^{\alpha}z_j\exp(z\beta)}{\left(1+\theta t^{\alpha}\exp(z\beta)\right)}-\delta z_j,$$

$$-\frac{\partial l}{\partial\alpha}=\left(\frac{1}{\theta}+\delta\right)\frac{\theta t^{\alpha}\log(t)\exp(z\beta)}{\left(1+\theta t^{\alpha}\exp(z\beta)\right)}-\delta\left(\frac{1}{\alpha}+\log(t)\right),\text{ and}$$

$$-\frac{\partial l}{\partial\theta}=\left(\frac{1}{\theta}+\delta\right)\frac{t^{\alpha}\exp(z\beta)}{\left(1+\theta t^{\alpha}\exp(z\beta)\right)}-\frac{1}{\theta^2}\log\left(1+\theta t^{\alpha}\exp(z\beta)\right).$$

The second derivatives are

$$-\frac{\partial^2 l}{\partial\beta_j\partial\beta_k}=\left(\frac{1}{\theta}+\delta\right)\frac{\theta t^{\alpha}z_jz_k\exp(z\beta)}{\left(1+\theta t^{\alpha}\exp(z\beta)\right)^2},$$

$$-\frac{\partial^2 l}{\partial\beta_j\partial\alpha}=\left(\frac{1}{\theta}+\delta\right)\frac{\theta t^{\alpha}z_j\log(t)\exp(z\beta)}{\left(1+\theta t^{\alpha}\exp(z\beta)\right)^2},$$

$$-\frac{\partial^2 l}{\partial\beta_j\partial\theta}=\left(\frac{1}{\theta}+\delta\right)\frac{t^{\alpha}z_j\exp(z\beta)}{\left(1+\theta t^{\alpha}\exp(z\beta)\right)^2}-\frac{t^{\alpha}z_j\exp(z\beta)}{\theta\left(1+\theta t^{\alpha}\exp(z\beta)\right)},$$

$$-\frac{\partial^2 l}{\partial\alpha\partial\theta}=\left(\frac{1}{\theta}+\delta\right)\frac{t^{\alpha}\log(t)\exp(z\beta)}{\left(1+\theta t^{\alpha}\exp(z\beta)\right)^2}-\frac{t^{\alpha}\log(t)\exp(z\beta)}{\theta\left(1+\theta t^{\alpha}\exp(z\beta)\right)},$$

$$-\frac{\partial^2 l}{\partial \alpha^2} = \left(\frac{1}{\theta} + \delta\right)\frac{\theta t^{\alpha}(\log t)^2 \exp(z\beta)}{\left(1+\theta t^{\alpha}\exp(z\beta)\right)^2} - \frac{\delta}{\alpha^2}, \text{ and}$$

$$-\frac{\partial^2 l}{\partial \theta^2} = -\frac{t^{\alpha}\exp(z\beta)}{\left(1+\theta t^{\alpha}\exp(z\beta)\right)^2}\left\{\frac{2}{\theta^2} + \frac{2}{\theta}t^{\alpha}\exp(z\beta) + \delta t^{\alpha}\exp(z\beta)\right\} + \frac{2}{\theta^3}\log\left(1+\theta t^{\alpha}\exp(z\beta)\right).$$

The maximum likelihood estimates for the fit can be found by minimising the negative log likelihood, for example using the **nlminb** routine in the statistical package S-Plus[134]. The resultant matrix of 2nd derivatives can be inverted to estimate the variance-covariance matrix of the parameter estimates. Since the first column of **z** is **1**, the parameter β_1 is related to the scale parameter of the baseline hazard function by

$$\gamma = \exp\left(\beta_1 / \alpha\right),$$

and the variance of the estimate of γ can be derived by the delta method from the original variance-covariance matrix.

REFERENCES

[1] Froelicher VF, Myers J, Follansbee WP, Labovitz AJ. *Exercise and the Heart, 3rd Edn.* St. Louis, Mosby, 1993.

[2] Gibbons RJ, Balady GJ, Bricker JT, et al. 'ACC/AHA 2002 guideline update for exercise testing: a report of the AmericanCollege of Cardiology/American Heart Association Task Force on Practice Guidelines (Committee on Exercise Testing)', http://www.acc.org/clinical/guidelines/exercise/exercise_clean.pdf. 2002.

[3] Stuart RJ, Ellestad MH. 'National survey of exercise stress testing facilites', Chest 1980; **77**: 94-97.

[4] Gibbons L, Blair SN, Kohl HW, Cooper K. 'The safety of maximal exercise testing', Circulation 1989; **80**: 846-852.

[5] Master AM, Rosenfeld I. 'The two step exercise test brought up-to-date', NY J Med 1961; **61**: 1850-.

[6] Bruce RA. 'Exercise testing of patients with coronary heart disease', Ann Clin Res 1971; **3**: 3223-332.

[7] Patterson JA, Naughton J, Pietras RJ, Gumar RN. 'Treadmill exercise in assessment of patients with cardiac disease', Am J Cardiol 1972; **30**: 757-762.

[8] Ellestad MH. *Stress Testing, 2nd Edn.* Philadelphia, FA Davis, 1980.

[9] Sullivan M, McKirnan MD. 'Errors in predicting functional capacity for postmyocardial infarction patients using a modified Bruce protocol', Am Heart J 1984; **107**: 486-492.

[10] Webster MWI, Sharpe DN. 'Exercise testing in angina pectoris: the importance of protocol design in clinical trials', Am Heart J 1989; **117**: 505-508.

[11] Panza JA, et al. 'Prediction of the frequency and duration of ambulatory myocardial ischemia in patients with stable coronary artery disease by determination of the ischemic threshold from exercise testing: importance of the exercise protocol', J Am Coll Cardiol 1991; **17**: 657-663.

[12] Redwood DR, Rosing DR, Goldstein RE, Beiser GD, Epstein SE. 'Importance of the design of an exercise protocol in the evaluation of patients with angina pectoris', Circulation 1971; **43**: 618-628.

[13] Will PM, Walter JD. 'Exercise testing: improving performance with a ramped Bruce protocol', Am Heart J 1999; **138**: 1033-1037.

[14] Loeppky JA, Greene ER, Hoekenga DE, Caprihan A, Luft UC. 'Beat-by-beat stroke volume assessment by pulsed Doppler in upright and supine exercise', J Appl Physiol 1981; **50**: 1173-1182.

[15] Rowell LB. *Human Circulation Regulation During Physical Stress.* New York, Oxford University Press, 1986.

[16] Wasserman K, Hansen JE, Sue DY, Whipp BJ, Casaburi R. *Principles of Exercise Testing and Interpretation*, 2nd Edn. Malvern, PA, Lea & Febiger, 1994.

[17] Nelson RR, Gobel FL, Jorgensen CR, Wang K, Wang Y, Taylor HL. 'Hemodynamic predictors of myocardial oxygen consumption during static and dynamic exercise', Circulation 1974; **50**: 1179-1189.

[18] MacRae HSH, Allen PJ. 'Automated blood pressure measurement at rest and during exercise: evaluation of the motion tolerant CardioDyne NBP 2000', Med Sci in Sports and Exercise 1998; **30**: 328-331.

[19] White WB, Berson AS, Robbins C, Jamieson MJ, Prisant LM, Roccella E, Sheps SG. 'National standard for measurement of resting and ambulatory blood pressures with automated sphygmomanometers', Hypertension 1993; **21**: 504-509.

[20] Detry J-MR, Piette F, Brasseur LA. 'Haemodynamic determinants of exercise ST-segment depression in coronary patients', Circulation 1970; **42**: 593-599.

[21] Weiner DA, McCabe C, Hueter DC, Ryan TJ, Hood WB. 'The predictive value of anginal chest pain as an indicator of coronary disease during exercise testing', Am Heart J 1978; **96**: 458-462.

[22] Cole JP, Ellestad MH. 'Significance of chest pain during treadmill exercise: correlation with coronary events', Am J Cardiol 1978; **41**: 227-232.

[23] Julian, D. J. *Angina Pectoris*, 2nd Edn. New York, Churchill Livingstone, 1985.

[24] The Scottish Executive Department of Health. *The Scottish Health Survey 1998.* http://www.show.scot.nhs.uk/ scottishhealthsurvey/index.htm. 2000.

[25] Rose GA, Blackburn H. *Cardiovascular Survey Methods.* WHO, Geneva, 1968.

[26] General Register Office for Scotland. *Vital Events Reference Tables.* http:// www.gro-scotland.gov.uk/grosweb/ grosweb.nsf/pages/reftabs. 2002.

[27] Mukerji V, Beitman BD, Alpert MA. 'Chest pain and angiographically normal coronary arteries. Implications for treatment', Texas Heart Institute J 1993; **20**: 170-179.

[28] McGill HC. *The Geographic Pathology of Atherosclerosis.* Baltimore, Williams and Wilkins, 1968.

[29] Yasue H, Kugiyama K. 'Coronary spasm: clinical features and pathogenesis', Internal Med 1997; **36**: 760-765.

[30] vanderLoo B, Martin JF. 'A role for changes in platelet production in the cause of acute coronary syndromes', Arteriosclerosis Thrombosis and Vascular Biology 1999; **19**: 672-679.

[31] Cianflone D, Lanza GA, Maseri A. 'Microvascular angina in patients with normal coronary arteries and with other ischaemic syndromes', Eur Heart J 1995; **16 (Suppl. I)**: 96-103.

[32] Scottish Intercollegiate Guidelines Network. 'Management of Stable Angina: A national clinical guideline', http://www.show.scot.nhs.uk/sign/pdf/sign51.pdf. 2001

[33] Gibbons RJ, Abrams J, Chatterjee K, et al. 'ACC/AHA 2002 Guideline Update for the Management of Patients with Chronic Stable Angina', http://www.acc.org/clinical/guidelines/stable/stable.pdf. 2002.

[34] Wood D, Durrington P, McInnes G, Poulter N, Rees A, Wray R, on behalf of the British Cardiac Society, British Hyperlipidaemia Association and British Hypertension Society. 'Joint British recommendations on prevention of coronary heart disease in clinical practice', Heart 1998; **80 (supplement 2)**: S1-S29.

[35] Psaty BM, Smith NL, Siscovick DS, *et al.* 'Health outcomes associated with antihypertensive therapies used as first-line agents: a systematic review and meta-analysis', JAMA 1997; **277**: 739-745.

[36] Tatti P, Pahor M, Byington RP, *et al.* 'Outcome results of the Fosinopril Versus Amlodipine Cardiovascular Events Randomized Trial (FACET) in patients with hypertension and non-insullin dependent diabetes mellitus', Diabetes Care 1998; **21**: 597-603.

[37] Taira N. 'Similarity and dissimilarity in the mode and mechanism of action between nicorandil and classic nitrates: an overview', J Card Pharmacol 1987; **28**: 1-9.

[38] The IONA Study Group. 'Effect of nicorandil on coronary events in patients with stable angina: the Impact of Nicorandil in Angina (IONA) randomised trial', Lancet 2002; **359**: 1269-1275.

[39] Agbo-Godeau S, Joly P, Lauret P, Szpirglas R, Szpirglas H. 'Association of major aphthous ulcers and nicorandil', Lancet 1998; **352**: 1598-1599.

[40] Watson A, Al Ozairi O, Fraser A, Loudon M, O'Kelly T. 'Nicorandil associated anal ulceration', Lancet 2002; **360**: 546-547.

[41] Task Force on the Management of Stable Angina Pectoris. 'Management of stable angina pectoris: Recommendations of the Task Force of the European Society of Cardiology', Eur H J 1997; **18**: 394-413.

[42] Faxon, D. P. 'Myocardial revascularization in 1997: angioplasty versus bypass surgery', American Family Physician, **56**, 1409-1418 (1997).

[43] Prinzmetal, M., Ekmekci, A., Kennamer, R., Kwoczynski, J. K., Subin, H., Toyoshima, A. 'Variant form of angina pectoris', JAMA 1960; **174**: 1794-.

[44] Weiner DA, McCabe CH, Ryan TJ. 'Identification of patients with left main and three vessel coronary disease with clinical and exercise test variables', Am J Cardiol 1980; **46**: 21-27.

[45] Taylor CD, Bandura A, Ewart CK, *et al.* 'Exercise testing to enhance wives' confidence in their husbands' cardiac capability soon after clinically uncomplicated acute myocardial infarction', Am J Cardiol 1985; **55**: 635-638.

[46] Ewart CK, Taylor CB, Reese LB, Debusk RF. 'Effects of early postmyocardial infarction exercise testing on self perception and subsequent physical activity', Am J Cardiol 1983; **51**: 1076-1080.

[47] Sullivan M, Genter F, Savvides M, *et al.* 'The reproducibility of hemodynamic, electrocardiographic, and gas exchange data during treadmill exercise in patients with stable angina pectoris', Chest 1984; **86**: 375-382.

[48] Committee for Proprietary Medicinal Products (CPMP). *Note for guidance on the clinical investigation of anti-anginal medicinal products in stable angina pectoris (CPMP/EWP/234/95).* http://www.emea.eu.int/. 1996.

[49] Pocock SJ. *Clinical Trials: A Practical Approach*. Wiley, 1997.

[50] Senn, S. *Cross-over Trials in Clinical Research*. Wiley, 2002.

[51] *EMEA/CPMP position statement on the use of placebo in clinical trials with regard to the revised Declaration of Helsinki (EMEA/17424/01)*. 2001.

[52] Bristol DR, Castellana JV. 'Survival analysis techniques in angina pectoris trials', Stats in Med 1990; 9: 293-299.

[53] The TIBET Study Group. 'The total ischemic burden European trial (TIBET): design, methodology, and management', Card Drugs Ther 1992; 6: 379-386.

[54] Dargie HJ, Ford I, Fox KM, on behalf of the TIBET study group. 'Total Ischaemic Burden European Trial (TIBET): Effects of ischaemia and treatment with atenolol, nifedipine SR and their combination on outcome in patients with chronic stable angina', Eur Heart J 1996; 17: 104-112.

[55] Fox KM, Mulcahy D, Findlay I, Ford I, Dargie HJ, on behalf of the TIBET study group. 'The Total Ischaemic Burden European Trial (TIBET): Effects of atenolol, nifedipine SR and their combination on the exercise test and total ischaemic burden in 608 patients with stable angina', Eur Heart J 1996; 17: 96-103.

[56] Antiplatelet Triallists Collaboration. 'Collaborative overview of randomised trials of antiplatelet therapy, I: prevention of death, myocardial infarction and stroke by prolonged antiplatelet therapy in various categories of patients', BMJ 1995; 308: 81-106.

[57] The Heart Outcomes Prevention Evaluation Study Investigators. 'Effects of an angiotensin-converting-enzyme inhibitor, ramipril, on cardiovascular event in high-risk patients', NEJM 2000; 342: 145-153.

[58] Held C, Hjemdahl P, Rehnqvist N, et el. 'Fibrinolytic variables and cardiovascular prognosis in patients with stable angina pectoris treated with verapamil or metoprolol. Results from the Angina Prognosis study in Stockholm', Circulation 1997; 95: 2380-2386.

[59] The Medical Research Council's General Practice Research Framework. 'Thrombosis prevention trial: randomised trial of low-intensity oral anticoagulation with warfarin and low-dose aspirin in the primary prevention of ischaemic heart disease in men at increased risk', Lancet 1998; 351: 233-241.

[60] Melandri G, Semprini F, Cervi V, et al. 'Benefit of adding low molecular weight heparin to the conventional treatment of stable angina pectoris. A double-blind, randomized, placebo-controlled trial', Circulation 1993; 88: 2517-2523.

[61] 'Randomised trial of cholesterol lowering in 4,444 patients with coronary heart disease: the Scandinavian Simvastatin Survival Study (4S)', Lancet 1994; 344: 1383-1389.

[62] West of Scotland Coronary Prevention Study Group. 'Influence of pravastatin and plasma lipids on clinical event in the West of Scotland Coronary Prevention Study (WOSCOPS)', Circulation 1998; 97: 1440-1445.

[63] Hennekens CH, Albert CM, Godfried SL, Gaziano JM, Puring JE. 'Adjunctive drug therapy of acute myocardial infarction – evidence from clinical trials', NEJM 1996; 335: 1660-1667.

[64] Frishman WH, Hieman M, Soberman J, Greenberg S, Eff J. 'Comparison of celiprolol and propranolol in stable angina pectoris. Celiprolol International Angina Study Group', Am J Cardiol 1991; 67: 665-670.

[65] Ezekowitz MD, Hossack K, Mehta JL, et al. 'Amlodipine in chronic stable angina: results of a multicentre double-blind crossover trial', Am Heart J 1995; 129: 527-535.

[66] Chrysant SG, Glasser SP, Bittar N, et al. 'Efficacy and safety of extended-release isosorbide mononitrate for stable effort angina pectoris', Am J Cardiol 1993; 72: 1249-1256.

[67] Di Somma S, Liguori V, Verdecchia P, *et al.* 'A double-blind comparison of nicorandil and metoprolol in patients with effort stable angina', Card Drugs Ther, 1993; 7: 119-123.

[68] Swan Study Group. 'Comparison of the antiischaemic and antianginal effect of nicorandil and amlodipine in patients with symptomatic stable angina pectoris: the SWAN study', J Clin Basic Card 1999; 2: 213-217.

[69] Krepp HP. 'Evaluation of the antianginal and anti-ischaemic efficacy of slow-release isosorbide-5-mononitrate capsules, bupranolol and their combination, in patients with chronic stable angina pectoris', Cardiology 1991; 79 (Suppl. 2): 14-18.

[70] Dunselman P, Liem AH, Vardel G, Kragten H, Bosma A, Bernink P. 'Addition of felodipine to metoprolol versus replacement of metoprolol by felodipine in patients with angina pectoris despite adequate beta-blockade. Results of the Felodipine ER and Metoprolol CR in Angina (FEMINA) Study. Working Group on Cardiovascular Research, The Netherlands (WCN)', Eur Heart J 1997; 18: 1755-1764.

[71] Yusuf S, Zucker D, Peduzzi P, et al. 'Effect of coronary artery bypass graft surgery on survival: overview of 10-year results from randomised trials by the Coronary Artery Bypass Graft Surgery Trialists Collaboration', Lancet 1994; 344: 563-570 [published erratum appears in Lancet 1994; 344: 1446].

[72] 'Guidelines and indications for coronary artery bypass graft surgery. A report of the American College of Cardiology/American Heart Association Task Force on Assessment of Diagnostic and Therapeutic Cardiovascular Procedures (Subcommittee on Coronary Artery Bypass Graft Surgery)', J Am Coll Cardiol 1991; **17**: 543-589.

[73] Bucher HC, Hengstler P, Schindler C, Guyatt GH. 'Percutaneous transluminal coronary angioplasty versus medical treatment for non-acute coronary heart disease: meta-analysis of randomised controlled trials', BMJ 2000; **321**: 73-77.

[74] Hueb WA, Bellotti G, de Oliveira SA, et al. 'The Medicine, Angioplasty or Surgery Study (MASS): a prospective, randomized trial of medical therapy, balloon angioplasty or bypass surgery for single proximal left anterior descending artery stenoses', J Am Coll Cardiol 1995; **26**: 1600-1605.

[75] Serruys PW, Unger F, Sousa JE, et al. 'Comparison of coronary artery bypass surgery and stenting for the treatment of multivessel disease', NEJM 2001; **344**: 1117-1124.

[76] King SB III, Kosinski AS, Guyton RA, Lembo NJ, Weintraub WS. 'Eight-year mortality in the Emory Angioplasty versus Surgery Trial (EAST)', J Am Coll Cardiol 2000; **35**: 1116-1121.

[77] The BARI Investigators. 'Seven-year outcome in the Bypass Angioplasty Revascularization Intervention (BARI) by treatment and diabetic status', J Am Coll Cardiol 2000; **35**: 1122-1129.

[78] http://www.ncbi.nlm.nih.gov/entrez/query/static/overview.html

[79] Parker JO. 'Eccentric dosing with isosorbide-5-mononitrate in angina pectoris', Am J Cardiol 1993; **72**: 871-876.

[80] Foale RA. 'Atenolol versus the fixed combination of atenolol and nifedipine in stable angina pectoris', Eur Heart J 1993; **14**: 1369-74.

[81] Meinertz T, Kasper W, Meier R, et al. 'Alinidine in angina', Clin Pharmacol Ther 1983; **34**: 770-776.

[82] Fox KM, Deanfield J, Selwyn A, Krikler S, Wright C. 'Factors influencing the treatment of chronic stable angina pectoris with nifedipine', Postgrad Med J 1983; **59 Suppl 2**: 25-29.

[83] Cox DR. 'Regression models and life tables' (with discussion), JRSS B 1972; **34**: 187-220.

[84] Altman DG. 'Statistics in medical journals: Developments in the 1980s', Stats in Med 1991; **10**: 1897-1913.

[85] Efron B. 'The efficiency of Cox's likelihood function for censored data', J Am Stat Assoc 1977; **72**: 557-565.

[86] Hastie T, Tibshirani RJ. *Gnerealized Additive Models*. London, Chapman & Hall, 1990.

[87] Therneau TM, Grambsch PM, Pankratz VS. 'Penalized survival models and frailty', J Comp Graph Stats 2003; **12**: 156-175.

[88] Bowman AW, Azzalini A. Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-plus Illustrations. Oxford University Press, 1997.

[89] Andersen PK, Gill RD. 'Cox's regression model for counting processes: a large sample study', Ann Statist 1982; **10**: 1100-1120.

[90] Wei LJ. 'Testing goodness of fit for proportional hazards model with censored observations', J Am Stat Assoc 1984; **79**: 649-652.

[91] Koziol JA, Byar DP. 'Percentage points of the asymptotic distributions of one and two sample K-S statistics for truncated or censored data', Technometrics 1975; **17**: 507-510.

[92] Schoenfeld D. 'Partial residuals for the proportional hazards regression model', Biometrika 1982; **69**: 239-241.

[93] Grambsch PM, Therneau TM. 'Proportional hazards test and diagnostics based on weighted residuals', Biometrika 1994; **81**: 515-526. Correction; **82**: 668.

[94] Cox DR, Oakes D. *Analysis of Survival Data*. London, Chapman and Hall, 1984.

[95] Lee ET. Statistical methods for survival data analysis. New York, Wiley, 1992.

[96] Peto R, Peto J. 'Asymptotically efficient rank invariant procedures', JRSS A 1972; **135**: 185-207.

[97] Altman DG. 'Statistics and ethics in medical research: misuse of statistics is unethical', BMJ 1980; **281**: 1182-1184.

[98] Thompson Jr WA. 'On the treatment of grouped observations in life studies', Biometrics 1977; **33**: 463-470.

[99] Nelder JA, Wedderburn RWM. 'Generalized linear models', JRSS A 1972; **135**: 370-384.

[100] Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. 'A comparison of goodness-of-fit tests for the logistic regression model', Stats in Med 1997; **16**: 965-980.

[101] Solomon PJ. 'Effect of misspecification of regression models in the analysis of survival data', Biometrika 1984; **71**: 291-298.

[102] Aranda-Ordaz FJ. 'An extension of the proportional hazards model for grouped data', Biometrics 1983; 39: 109-117.

[103] Dorey FJ, Little RJA, Schenker N. 'Multiple imputation for threshold crossing data', Stats iun Med 1993; 12: 1589-1603.

[104] France LA, Lewis JA, Kay R. 'The analysis of failure time data in crossover studies', Stats in Med 1991; 10: 1099-1113.

[105] Gehan E. 'A generalized Wilcoxon test for comparing arbitrarily single censored samples', Biometrika 1965; 52: 203-223.

[106] Wei LJ. 'A generalized Gehan and Gilbert test for paired observations that are subject to arbitrary right censorship', J Am Stat Assoc 1980; 75: 634-637.

[107] Lam FC, Longnecker MT. 'A modified Wilcoxon rank sum test for paired data', Biometrika 1983; 70: 510-513.

[108] Albers W. 'Combined rank tests for randomly censored paired data', J Am Stat Assoc 1988; 83: 1159-1162.

[109] Dabrowska DM. 'Signed-rank tests for censored matched pairs', J Am Stat Assoc 1990; 85: 478-485.

[110] Woolfson RF, O'Gorman TW. 'A comparison of several tests for censored paired data', Stats in Med 1992; 11: 193-208.

[111] O'Brien PC, Fleming TR. 'A paired Prentice-Wilcoxon test for censored paired data', Biometrics 1987; 43: 169-180.

[112] Akritas MG. 'Rank transform statistics with censored data', Stats and Prob Letters 1992; 13: 209-221.

[113] Zeger, S. L. and Liang, K.-Y. (1986). 'Longitudinal Data Analysis for Discrete and Continuous Outcomes.' Biometrics 1986; 42: 121-30.

[114] Pinheiro J, Bates DM. *Mixed Effects Models in S and S-Plus*. New York, Springer-Verlag, 2000.

[115] Schweizer B, Sklar A. *Probabilistic Metric Spaces*. New York, North-Holland, 1983.

[116] Liang K-Y, Self SG, Bandeen-Roche KJ, Zeger SL. 'Some recent developments for regression analysis of multivariate failure time data', Lifetime Data Analysis 1995; 1: 403-415.

[117] Huster WJ, Brookmeyer R, Self SG. 'Modelling paired survival data with covariates', Biometrics 1989; 45: 145-156.

[118] Mantel N, Ciminera JL. 'Use of logrank scores in the analysis of litter-matched data on time to tumor appearance', Cancer Research 1979; 39: 4308-4315.

[119] Holt JD, Prentice RL. 'Survival analysis in twin studies and matched pair experiments', Biometrika 1974; 61: 17-30.

[120] Pickles A, Crouchley R. 'A comparison of frailty models for multivariate survival data', Stats in Med 1995; 14: 1447-1461.

[121] Vaupel JW, Manton KG, Stallard E. 'The impact of heterogeneity in individual frailty on the dynamics of mortality', Demography 1979; 16: 439-454.

[122] Dempster AP, Laird NM, Rubin DB. 'Maximum likelihood from incomplete data via the *EM* Algorithm' (with discussion), JRSS B 1977; 39: 1-38.

[123] Klein, J. P. 'Semiparametric estimation of random effects using the Cox model based on the EM algorithm', Biometrics 1992; 48: 795-806.

[124] Andersen PK, Klein JP, Knudsen KM, Tabanera y Palacios R. 'Estimation of variance in Cox's regression model with shared gamma frailties', Biometrics 1997; 53: 1475-1484.

[125] Li H, Thompson E. 'Semiparametric estimation of major gene and family-specific random effects for age of onset', Biometrics 1997; 53: 282-293.

[126] Gelfand AE, Smith AFM. 'Sampling-based approaches to calculating marginal densities', J Am Stat Assoc 1990: 85: 398-409.

[127] Hougaard P. 'Survival models for heterogeneous populations derived from stable distributions', Biometrika 1986: 73: 387-396.

[128] Gray RJ. 'Flexible methods for analyzing survival data using splines, with applications to breast cancer and prognosis', J Am Stat Assoc 1992; 87: 942-951.

[129] Yau KKW, McGilchrist CA. 'ML and REML estimation in survival analysis with time dependent correlated frailty', Stats in Med 1998; 17: 1201-1213.

[130] Andersen PK, Borgan O, Gill RD, Keiding N. *Statistical Models based on Counting Processes*. New York, Springer-Verlag 1993.

[131] Lunn M, McNeil D. 'Applying Cox regression to competing risks', Biometrics 1995; 51: 524-532.

[132] David HA, Moeschberger ML. *The theory of competing risks*. Lodon, Charles Griffin & Company Limited, 1978.

[133] Goldstein H, Rasbash J, Plewis I, et al. *A User's Guide to MLwiN*. Institute of Education: University of London, 1998.

[134] Venables WN, Ripley BD. *Modern Applied Statistics with S*. New York, Springer-Verlag, 2002.