Prognostic and Surrogate Markers for Outcome in the Treatment of Pulmonary Tuberculosis

Patrick Peter John Phillips

patrick.phillips@ctu.mrc.ac.uk

London School of Hygiene and Tropical Medicine,
University of London

*Submitted for the degree of Doctor of Philosophy in
the field of Medical Statistics*

Supervised by Dr. Katherine Fielding and Prof. Abdel Babiker

September 2009

LONDON SCHOOL OF HYGIENE & TROPICAL MEDICINE

Statement of Own Work

All students are required to complete the following declaration when submitting their thesis. A shortened version of the School's definition of Plagiarism and Cheating is as follows (the full definition is given in the Research Degrees Handbook):

The following definition of plagiarism will be used:

Plagiarism is the act of presenting the ideas or discoveries of another as one's own. To copy sentences, phrases or even striking expressions without acknowledgement in a manner which may deceive the reader as to the source is plagiarism. Where such copying or close paraphrase has occurred the mere mention of the source in a biography will not be deemed sufficient acknowledgement; in each instance, it must be referred specifically to its source. Verbatim quotations must be directly acknowledged, either in inverted commas or by indenting. (University of Kent)

Plagiarism may include collusion with another student, or the unacknowledged use of a fellow student's work with or without their knowledge and consent. Similarly, the direct copying by students of their own original writings qualifies as plagiarism if the fact that the work has been or is to be presented elsewhere is not clearly stated.

Cheating is similar to plagiarism, but more serious. Cheating means submitting another student's work, knowledge or ideas, while pretending that they are your own, for formal assessment or evaluation.

Supervisors should be consulted if there are any doubts about what is permissible.

Declaration by Candidate

I have read and understood the School's definition of plagiarism and cheating given in the Research Degrees Handbook. I declare that this thesis is my own work, and that I have acknowledged all results and quotations from the published or unpublished work of other people.

Signed:.. Date: 22/09/2009

Full name: Patrick Peter John Phillips .... (please print clearly)

2

Abstract

Phase III trials for new tuberculosis treatment regimens require large numbers of participants and can take over five years to complete. A surrogate marker for poor outcome (failure at end of treatment or recurrence following successful treatment), the established endpoint in such trials, could shorten trial duration and reduce trial size. Culture results after two months of treatment have shown the most promise but, prior to this research, no formal evaluation had been performed.

In this thesis, culture results during treatment are evaluated as prognostic and surrogate markers for poor outcome using data on 6974 patients from twelve tuberculosis treatment randomised controlled multi-arm trials conducted in East Africa and East Asia.

A strong association was found between culture results during treatment and poor outcome. Nevertheless, culture results were not good patient-specific predictors of poor outcome with low sensitivities and specificities.

Existing meta-analytic methods for evaluating surrogate markers are not wholly suited to this setting of multi-arm trials with binary true and surrogate endpoints. Extending these methods, the two month culture was found to be a good surrogate marker using data from Hong Kong trials and the three month culture was found to be a good surrogate marker using data from East African trials. These results are an indication that cultures during treatment do capture some of the treatment effect. Further work is needed in understanding the differences between the Hong Kong and East African trials.

The meta-analytic methods for evaluating surrogate markers in this thesis included a graphical representation that permitted a clear visual evaluation of the surrogate. Methods developed in this thesis for modelling the relationship between the treatment effects on the true and surrogate endpoints were not satisfactory. The deficiencies were not overcome with the two extensions proposed. Further work is needed in developing a more appropriate model.

Acknowledgements

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Tuberculosis: Preventable, Curable and Unrelenting

Tuberculosis (TB) is the world's oldest infectious disease and over the centuries has been responsible for more mortality, morbidity and human suffering that any other (Youmans, 1979). The bacilli causing TB, *Mycobacterium tuberculosis*, has been detected in Andean Mummies dating from 140 AD (Konomi et al., 2002), in Egyptian Mummies dating from 1000 BC (Zink et al., 2003), in bones from an infant and a woman in the Eastern part of the Mediterranean from 7000 BC (Hershkovitz et al., 2008) and in the bones of extinct North American bison dating from 15000 BC (Rothschild et al., 2001). In his *Aphorisms*, Hippocrates (ca. 460 BC - ca. 370 BC) identifies tuberculosis (*phthisis* in Greek) as a widespread disease that was almost always fatal, particularly if 'the hairs of the head fall off'[1]. There are two probable references to TB in the Old Testament of the Bible during the period when the Israelites lived among the Egyptians around 1440 BC (Daniel and Daniel, 1999). Many specimens of ancient drawings and pottery from around the world show people with physical deformities characteristic of TB.

More recently, in the United States of America, in the middle of the nineteenth century as many as a quarter of all deaths reported were caused by diseases of the lung (Dubos and Dubos, 1996) and even in 1930, TB was the

---

[1]Hippocrates. Aphorisms. eBooks @ Adelaide. 2007. http://ebooks.adelaide.edu.au /h/hippocrates/aphorisms/. Retrieved 23 Apr 2009.

most frequent cause of death or disability for those aged between 15 and 45. TB was also the single biggest cause of death and disease in England in the nineteenth century with mortality at times over 400 per 100,000 of population per year (Dubos and Dubos, 1996).

In 1882 *M. tuberculosis* was discovered by Robert Koch. Deaths due to TB dropped in both England and the United States over the first half of the twentieth century, largely due to improvements in living conditions and nutrition following the end of the industrial revolution (Dubos and Dubos, 1996).

At the beginning of the 20th century, a character in a play by George Bernard Shaw, *A Doctor's Dilemma*, described the treatment for TB in England at that time as "a huge commercial system of quackery and poison" (Iseman, 2002). Philip D'Arcy Hart's review of chemotherapy for TB published in 1946 states that 'principal measures recommended at this time [the turn of the century] were rest, sometimes tempered with exercise, and plentiful diet, often to the point of overfeeding'. The author also notes 'It was a time when a drug was seldom dropped, although it was a routine procedure to add one' (Hart, 1946). It was not until 1944 that Streptomycin, the first truly effective treatment against tuberculosis (Hart, 1946) was discovered in the United States leading to great celebration (including patients apparently dancing in the wards in a New York Hospital, Reichman (1999)). Following a small study of 24 patients showing benefit (Hinshaw and Feldman, 1945), a small quantity of streptomycin was made available to the British Medical Research Council (MRC) on the other side of the Atlantic. This lead to what is generally regarded as the first randomised controlled trial (Medical Research Council, 1948) using the methods developed by Sir Austin Bradford Hill (Hill, 1937), and the first controlled trial of streptomycin for pulmonary tuberculosis (PTB) in man. Thus begun a series of studies based largely around this tool of the randomised controlled trial that took place over the next forty years to determine the most effective regimen to treat TB. For the first time in history, TB became a curable disease and, by the 1970s, a highly effective six month regimen was identified. The development of *short-course chemotherapy* (described as such to contrast with the 18-month regimens in use in the 1960s and 1970s) is largely a result of the work of the MRC (Iseman, 2002; Murray, 2004; Raviglione and Pio, 2002).

In 1986, the MRC's highly influential TB research units closed down and attention shifted to other disease areas as it was widely believed that there was no need for new anti-tuberculosis agents—all that was required to defeat

tuberculosis was the correct implementation of the recommended treatment regimens. In 1999, an entire 50-page supplement to the International Journal of Tuberculosis and Lung Disease, written by three of the key scientists involved—a statistician, a biochemist and a microbiologist, was devoted to the work of these units and made this point:

> 'When the [MRC TB] units were closed in 1986, *all of the measures necessary* for successful programmes for the control of tuberculosis had been delineated... These tools were then available to national organisations and to international organisations such as the World Health Organisation (WHO) and the International Union Against Tuberculosis and Lung Disease (IUATLD), to implement in control programmes.'

Fox et al. (1999)

Unfortunately, the tools available at the time were insufficient to prevent the resurgence of the TB epidemic.
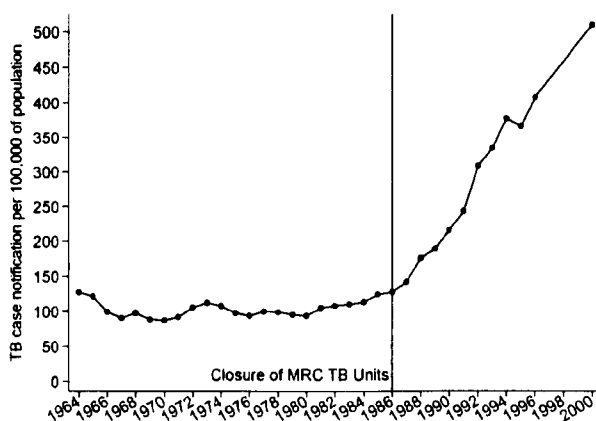
Figure 1.1: New TB case notification rates per 100,000 in Zambia from 1964 to 2000. Data from Mwaba et al. (2003).

Figure 1.1 shows new TB case notification rates in Zambia from 1964 to 1996. Notification rates had levelled during the 1960s, 70s and 80s and it is the introduction and spread of HIV in the 1980s in Zambia that caused the steep increase in the late 1980s (Maartens and Wilkinson, 2007).

History has shown that the aforementioned *control of tuberculosis* has not been achieved:

13

'The eradication of TB has proven to be an elusive goal, often in individual patients as well as in large populations'

(Wallis et al., 1999)

1.2   The Need for New Anti-Tuberculosis Drugs

Today almost one third of the world's population are infected with TB and approximately 8 million of these will develop active TB. There are 9.2 million new cases and 1.7 million people die every year from TB despite a cure being available for as little at US $10. The emergence of drug resistant strains of *M. tuberculosis* and the HIV epidemic have fuelled the increase with TB being the leading cause of death among people infected with HIV (World Health Organization, 2008; Stop TB Partnership, 2006; World Health Organization, 2006).

In response to this renewed threat, the WHO declared TB a global emergency as far back as April 1993. It has been shown that cure rates of up to 95% can be achieved in a clinical trial setting (Fox et al., 1999; World Health Organization, 2008) and at this time it was recognised that to effectively control tuberculosis, well-organised tuberculosis diagnostic and treatment services were required. To meet this need, WHO and its partners introduced the directly observed therapy, short course (DOTS) strategy (World Health Organisation, 1994), replacing this with an expanded DOTS framework in 2002 (World Health Organisation, 2002). This expanded framework has five elements:

1. Sustained political commitment to increase resources,

2. Access to quality-assured TB sputum microscopy for case detection,

3. Standardised short-course chemotherapy to all cases of TB under proper case-management including direct observation of treatment,

4. Uninterrupted supply of quality-assured drugs,

5. Recording and reporting of treatment outcome.

Standardised short-course chemotherapy (as recommended by World Health Organization (2003)) consists of four medications taken daily for two months

14

followed by two medications taken for a further four months. DOTS is labour-intensive for health staff and, while six months of treatment is significantly short than the eighteen or twenty-four month regimens recommended before the introduction of short-course chemotherapy (British Medical Research Council, 1962), it is still a significant burden on the patient (particularly as patients can be taking up to eleven separate pills every day (Camp et al., 2006)) and is difficult to implement where health services are poorly accessible.

An average treatment success rate of 84.5% was observed in the WHO DOTS 2005 cohort (patients diagnosed worldwide with active TB in 2005, treated using DOTS and followed up into 2006), but this figure was only 74% in the African Region dropping below 60% for some African countries (World Health Organization, 2008). A treatment success is defined as a patient who has completed at least six months of treatment with a negative smear at the end of treatment in addition to a negative smear before the end of treatment or who has not had a positive smear after the fifth month of treatment (World Health Organization, 2008). National TB programmes are judged by their treatment success rates.

Low success rates may be due to poor adherence or limited access to health care facilities. Symptoms of TB usually disappear within the first few months of treatment and therefore many patients can be reluctant to continue to take their treatment for the full course of six months (Munro et al., 2007) although six months of treatment is necessary to reduce rates of relapse and acquired drug resistance. A simpler, shorter regimen is needed to improve compliance and improve treatment outcomes. It has been demonstrated that this is not possible with the existing drugs in use today (Fox et al., 1999). New drugs are also needed to combat increasing rates of drug resistance and the complex co-infection of TB and HIV.

With the WHO as the leading partner, the Stop TB Partnership was established in 2000 to realise the goal of the elimination of TB by 2050. To raise awareness of TB, the Stop TB Partnership launched World TB Day in 2000 on the 24th of March which has since been repeated annually. The Partnership produced the Global Plan to Stop TB 2006-2015, building on the Partnership's first plan for 2001-2005, which sets out activities and goals for all levels of political and community involvement to bring TB under control.

Responding to the need for new drugs as expressed in the Global Plan and by other authors (O'Brien and Nunn, 2001), the Global Alliance for TB Drug Development (GATDD) was also formed in 2000 with the mission to 'acceler-

ate the discovery, development and equitable distribution of new drugs' primarily to 'shorten the duration of TB treatment or otherwise facilitate its successful completion' (Global Alliance for TB Drug Development, 2001). Their aim is to register such a new compound by 2010 (The Working Alliance for TB Drug Development, 2000). One such candidate drug is Moxifloxacin which is being investigated in a Phase III clinical trial (REMoxTB) coordinated by UCL and the MRC Clinical Trials Unit under the auspices of the GATDD with recruitment having started in the middle of 2008. The trial uses a non-inferiority design.

There are currently seven compounds in clinical development for the treatment of TB (Casenghi and von Schoen-Angerer, 2006; Spigelman, 2007), four of which are new drug candidates with novel mechanisms of action (Rivers and Mancera, 2008), as well as a number of compounds in the pre-clinical and discovery stages . The possibility of using a high dose of a compound from the rifamycin family may also lead to shorter, intermittent regimens (Rosenthal et al., 2007).

It is clear that there will be an increasing number of late phase clinical trials being conducted over the next few years as more and more compounds enter the clinical stage of drug development.

1.3   The Need for Surrogate Markers

In the scientific blueprint for TB drug development produced by the GATDD in 2001, several barriers to TB drug development are explored, highlighting particularly the lack of infrastructure and experience in phase III clinical trials. They identify surrogate markers for long-term response as a way to streamline phase III trials (Global Alliance for TB Drug Development, 2001). A *surrogate marker* (synonymous in this thesis for *surrogate endpoint*, see discussion in section 2.2) is one that is used as a substitute for the usual final endpoint in a clinical trial and must therefore capture any effect of the trial intervention on this final endpoint.

Tuberculosis is unique among infectious diseases in that failure to culture bacilli in a sputum sample is not necessarily indicative of cure—a patient needs to remain smear or culture negative for a period of several consecutive months to be declared cured (World Health Organization, 2008). In a TB clinical trial, patients need to be followed up for from twelve to twenty-

four months after the end of treatment to be assessed for final endpoint of *poor outcome* to treatment (failure during treatment or relapse after the end of treatment). This period of extensive follow-up is costly and can mean that a TB trial could take five or more years to complete (e.g. Jindani et al., 2004). Recurrence rates under clinical trial conditions are often less than 5% (Fox et al., 1999) and therefore large numbers of patients are required in a trial, even to show non-inferiority (Nunn et al., 2008). It has been suggested that the total time required to develop a new TB drug from discovery to regulatory approval would be 20 years[2]. This is all because 'there are no accepted surrogate markers for efficacy' (Spigelman, 2007).

A properly validated surrogate marker for response to anti-tuberculosis drugs would result in trials without the need for extensive follow-up for relapse which are therefore shorter in duration and cheaper. Such a marker would ultimately speed the drug development process aiding the Stop TB partnership and the GATDD in their aims.

1.4 The Scope of this Thesis

The aim of this research is to identify and evaluate prognostic and surrogate markers for poor outcome of treatment for TB. As is demonstrated in Chapter 3, the markers that have shown most promise historically and in recent years are culture results during treatment. Much has been written about the use of culture results to predict relapse (Global Alliance for TB Drug Development, 2001; Desjardin et al., 1999; Wallis et al., 2000) or in some cases substitute for relapse in a clinical trial (Sirgel et al., 2000) but the evidence for these claims is ambiguous. This research project will focus primarily on the formal evaluation of culture results during treatment as prognostic and surrogate markers for poor outcome.

The statistical literature on evaluating prognostic and surrogate markers is reviewed in Chapter 2. Chapter 3 begins with an overview of the diagnosis, treatment and progression of TB disease before reviewing the literature on possible prognostic and surrogate markers for poor outcome. The data used for the analyses in this thesis are introduced and described in Chapter 4, and the results of some exploratory analyses presented. Culture positivity during treatment is evaluated as a prognostic marker in Chapter 5 and evaluated as a

---

[2]Personal communication from Amina Jindani.

surrogate marker in Chapter 6. In Chapter 7, methods for evaluating surrogate markers are extended to better account for some of the specific complexities of the data used in this thesis. These methods are also compared using a simulation study. All of these results are brought together in the discussion in Chapter 8 and conclusions drawn.

Appendix A contains some additional figures and tables of detailed results where only summary tables have been presented in the main body of the thesis. Appendix B gives the derivation of the reliability ratio in different situations to be used in analyses in Chapter 7. Appendix C contains a glossary of terms used in this thesis and Appendix D lists the mathematical notation used to describe the methods used in this thesis.

Chapter 2

The Evaluation of Prognostic and Surrogate Markers

2.1 Introduction

In this chapter, the statistical literature on evaluating prognostic and primarily surrogate markers is summarised and reviewed with particular focus on those methods which are likely to be appropriate for use in this thesis.

Any discussion of surrogate endpoints must involve a review of the 'controversial issues arising when surrogate endpoints are used as study outcomes' (Fleming et al., 1998). The benefits promised by the idealised concept of a surrogate endpoint—shorter, cheaper, smaller trials—are not in question; the controversy relates to the actual process of the evaluation of surrogate endpoints and the subsequent use of apparently 'validated' markers.

The literature is littered with examples of failed surrogates. These are markers that were used as surrogates in clinical trials on the basis of poor evidence, to demonstrate the efficacy of an experimental regimen. They were subsequently found not to be surrogates and the results from the clinical trials therefore of little use. It is the use of surrogate markers before they have been thoroughly evaluated as such that has fuelled the negative perception of surrogate markers for long term response.

To circumvent this controversy, the guidelines covering Statistical Principles for Clinical Trials (ICH E9, International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals For Human Use,

1998) prepared by the International Conference on Harmonisation of Techni-
cal Requirements for Registration of Pharmaceuticals for Human Use (ICH) in
1998 gives the following three criteria which are necessary for the evaluation
of a surrogate endpoint:

1. 'The biological plausibility of the relationship,

2. the demonstration in epidemiological studies of the prognostic value of
   the surrogate for the clinical outcome, and

3. evidence from clinical trials that treatment effects on the surrogate cor-
   respond to effects on the clinical outcome.'

International Conference on Harmonisation of Technical Requirements for
Registration of Pharmaceuticals For Human Use (1998)

These three requirements are distilled from much of the literature de-
scribed below, but form the basis of any statistical validation of a surrogate
endpoint and will be referred back to in subsequent chapters of this thesis. A
similar list of criteria or 'provisos' is given in Boissel et al. (1992).

There is now regulation by the US Food and Drug Administration (FDA)
that can grant accelerated marketing approval for drugs on the basis of tri-
als using surrogate endpoints and this was used to quickly approve new an-
tiretroviral (ARV) drugs for HIV for which the efficacy was later shown using
relevant clinical outcomes (Bucher et al., 1999).

In section 2.2, clear definitions of the various terms in use are given. Fol-
lowing these definitions, section 2.3 surveys some examples of surrogates that
have been used in other disease areas, both correctly and incorrectly. Section
2.4 reviews methods for evaluating prognostic markers. Section 2.5 introduces
the seminal Prentice criteria for surrogate endpoints and reviews the literature
exploring these criteria.

The move away from the hypothesis tests of the Prentice criteria towards
quantifying the proportion of treatment effect captured by a surrogate marker
is summarised in section 2.6. Section 2.7 looks at one particular trial design
using a surrogate endpoint. Section 2.8 reviews the most recent body of statis-
tical work in this area by a group of Belgian statisticians and section 2.9 covers
related methods developed to evaluate surrogate markers in the disease area
of HIV.

2.2 Definitions

Most authors refer to one of these three definitions in any discussion of surrogate markers:

- *Prentice, 1989.* The first attempt to provide a clear statistical definition of a surrogate endpoint was made by Prentice (1989) in a paper published alongside three other papers summarising the use of surrogate endpoints in clinical trials in three separate disease areas. The introduction to this series of four papers provides a typically weak definition of a surrogate endpoints as 'one that an investigator deems as correlated with an endpoint of interest but that can perhaps be measured at lower expense or at an earlier time than the endpoint of interest' (Herson, 1989). Other definitions in this same series suggest that a surrogate endpoint is one that is 'sufficiently well correlated' (Ellenberg and Hamilton, 1989) or that 'relates in some way' (Hillis and Seigel, 1989) to the clinical endpoint. In actuality, each of these definitions describe a *prognostic marker*. Prognosis is the 'estimation of the relative probability of the various possible outcomes of a disease' (Walter, 1998) whatever the treatment given.

  Prentice argued that a surrogate endpoint should be considered in the context of the treatment comparison, as this is the purpose of an intervention clinical trial, and redefined a surrogate endpoint to be 'a response variable for which a test of the null hypothesis of no relationship to the treatment groups under comparison is also a valid test of the corresponding null hypothesis based on the true endpoint' (Prentice, 1989). This definition has become widely accepted and is the most common starting point for a discussion on the evaluation of surrogate endpoints. Prentice was careful to highlight the fact that a surrogate is only defined with respect to a particular treatment comparison.

- *Temple, 1995.* Temple (1995) defined a surrogate endpoint as 'a laboratory measurement or a physical sign used as a substitute for a clinically meaningful endpoint that measures directly how a patient feels, functions or survives. Changes induced by a therapy on a surrogate endpoint are expected to reflect changes in a clinically meaningful endpoint.' This definition emphasises the need to have a clearly defined final endpoint for which the surrogate will be a substitute for, and that it is the changes

due to the therapy in the surrogate, rather than absolute values, that should be related to changes in the final endpoint.

- *Biomarkers Definitions Working Group, 2001.* A working group of the US National Institute of Health on Biomarkers and Surrogate Endpoints defines a *biomarker* as 'a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention' and defined a surrogate endpoint as 'a biomarker that is intended to substitute for a clinical endpoint' (Biomarker Definitions Working Group, 2001). They go on: 'A surrogate endpoint is expected to predict clinical benefit (or harm or lack of benefit or harm) based on epidemiologic, therapeutic, pathophysiologic, or other scientific evidence.' Under this definition it is clear that a *biomarker* is a more general term and must meet additional conditions to qualify as a surrogate endpoint. A biomarker is merely an *indicator* of some treatment induced response, but a surrogate must *predict* some meaningful clinical benefit.

Some authors use the term *surrogate marker* as a synonym for surrogate endpoint, while some authors discourage this terminology in the context of trials and treatment comparisons as it detracts from the role of a surrogate as a substitute for a hard clinical endpoint (Biomarker Definitions Working Group, 2001; Burzykowski et al., 2005). In this thesis, these two terms will be considered to be synonymous.

Freedman et al. (1992) describe *intermediate endpoints* as 'biological markers or events that may be assessed or observed prior to the clinical appearance of the disease, and that bear some relationship to the development of that disease.' With this definition, they are effectively proposing the term 'intermediate endpoints' to refer to candidate surrogate endpoints that have not yet been validated. Boissel et al. (1992) define an intermediate endpoint as 'a response variable which is statistically correlated with the clinical endpoint' which can then 'qualify' as a surrogate 'if it can be used as an appropriate alternative to a clinical endpoint in a clinical trial'. This is in contrast with an *auxiliary endpoint* which is defined as one that cannot fully substitute for the final endpoint, but can perhaps augment the clinical event information thus strengthening the true endpoint analyses (Fleming et al., 1994).

Lassere et al. (2007b) links Biomarker Definitions Working Group (2001) and Temple (1995) by contrasting *patient outcomes* which directly reflect 'how a

patient feels, functions or survives' with *biomarkers* which are 'disease-centred variables of biological and pathological processes'. The authors propose a scoring system (ranging from 0 to 15) for evaluating surrogate endpoints concluding that a surrogate endpoint is effectively a biomarker that can also be described as a patient outcome.

A meeting was initiated jointly by the Special Programme for Research and Training in Tropical Disease (TDR) and the European Commission in Geneva, Switzerland in June 2008 to discuss and evaluate the role of biomarkers and surrogate endpoints in the management of patients with tuberculosis. The report from this meeting included a useful table giving a number of definitions used which is reproduced below.

- *Biomarker (biological marker).* A measurable characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathological processes, or physiological or pharmacological responses to a therapeutic intervention.

- *Biosignature.* A group of biomarkers used together that form highly multiplexed biosignatures.

- *Surrogate end-point.* A biomarker that is intended to substitute for a clinical end-point based on epidemiological, therapeutic, pathophysiological or other scientific evidence.
  Predicts clinical outcome in terms of benefit, or harm or lack of benefit.

- *Clinical end-point.* A characteristic or variable that reflects the final outcome of disease in terms of function, effect, progress, recovery, survival or death.

- *Surrogates of protection.* Validated markers of correlates of protection.

- *Correlates of protection.* Measurable sign(s) in a host in response to an infectious agent indicating whether the individual is being protected against becoming infected and/or developing disease.

Reproduced with permission from Zumla et al. (2008)

2.2.1 Defining, Evaluating, Qualifying or Validating?

Prentice (2005) is clear, that he proposed his original criteria to give a means of *defining* a surrogate marker and that '*evaluating* whether a certain biomarker, or short-term clinical outcome can reasonably serve as a replacement for a true endpoint T, is quite another matter'. Many authors talk about *surrogate marker validation* (e.g. Freedman et al., 1992; De Gruttola et al., 1997; Buyse and Molenberghs, 1998; Lassere et al., 2007b) or *qualification* (e.g. Boissel et al., 1992), although the single published textbook on surrogate endpoints (Burzykowski et al., 2005) is entitled 'The *Evaluation* of Surrogate Endpoints'.

Wagner et al. (2007), from a regulatory perspective, distinguishes between *assay method validation* as 'determining the range of conditions under which the assay will give reproducible and accurate data' and *biomarker qualification* as the 'evidentiary process of linking a biomarker with biological processes and clinical endpoints'. On the basis of these definitions, the authors argue that the term qualification should be used in preference to validation. The authors state that a biomarker that is to be used as a surrogate marker will require substantial qualification. They break the process of qualification into four stages naming the final stage as 'surrogacy'.

Berger (2004) prefers the term 'validation', although the author notes that 'it is common to assume that the Prentice criteria imply that the use of the now *validated surrogate endpoint* will lead to *valid inference* for the true clinical endpoint'. The author unsuccessfully attempts to discover an inconsistency in Prentice's work, but makes an important point with regard to the use of the term *validation*. To avoid these connotations, the word *evaluation* has been preferred in this thesis.

2.3 Examples

There many examples of surrogate endpoints that are used in different disease areas, a number of which have subsequently been shown to be very poor surrogates in later research. A selection of these surrogates are detailed here.

- *Ventricular arrhythmia as a surrogate for death following myocardial infarction*. It is known that ventricular arrhythmia is associated with an almost four-fold increase in the risk of death related to cardiac complications. Three drugs (encainide, flecainide and moricizine) were found to sup-

press arrhythmia and were approved by the FDA and more than 200,000 persons per year took these drugs in the US. Beginning enrolment in 1987, the Cardiac Arrhythmia Suppression Trial (CAST) evaluated these three drugs in patients who had had a myocardial infarction (The Cardiac Arrhythmia Suppression Trial (CAST) Investigators, 1989; Ruskin, 1989). Importantly, this trial did not use the accepted surrogate endpoint, but the final clinical endpoint (death) to evaluate these therapies. The trial was stopped early after finding an increased risk of death in *all three treatment arms*. Ventricular arrhythmia was accepted as a surrogate marker on the basis of statistical correlation without a proper evaluation of whether it was a valid surrogate.

- *Bone Mineral Density as a surrogate for bone fracture in osteoporosis.* Bone mineral density was proposed as a surrogate endpoint for bone fracture in osteoporosis in post-menopausal women, with higher bone mineral density reflecting fewer fractures. A randomised controlled trial conducted in 202 women (Riggs et al., 1990) showed that treatment with sodium fluoride increased bone mineral density by 35% over five years. The trial also showed that the same treatment resulted in 50% more vertebral and non-vertebral fractures when compared to placebo. The conclusions were that sodium fluoride increases bone mineral density, but also causes the bones to become more brittle leading to more fractures.

- *Blood pressure as a surrogate for a coronary event.* Phase III trials of new therapies for the treatment of hypertension use the surrogate endpoint of blood pressure as the endpoint. Blood pressure in itself is not a an endpoint that directly reflects how a patients 'feels, functions or survives' (see section 2.2) and, in fact, it is intended that these therapies will prevent episodes of congestive heart failure and other coronary events (Berger, 2004; Temple, 1999).

2.4 The Evaluation of Prognostic Markers

2.4.1 Introduction

Before surveying the literature on the statistical evaluation of surrogate markers, this section examines approaches to evaluate markers for use in prognosis. As described in section 2.2, a prognostic marker is used for predicting disease

outcome. In the context of this thesis, the aim is to evaluate markers for long-term response to treatment for tuberculosis. These markers will therefore be used to classify individuals as having a *fair outcome* to treatment or as having a *poor outcome* to treatment. In this section, we use the binary variable T to denote the final treatment outcome, the true endpoint:

$$T = \begin{cases} 0 & \text{for fair outcome;} \\ 1 & \text{for poor outcome.} \end{cases} \qquad (2.1)$$

Much of this theory comes from the Pepe (2003), Feinstein (2002) and Erdreich and Lee (1981). Appendix D gives a list of notation used throughout the thesis.

2.4.2 The Evaluation of binary prognostic markers

The simplest type of prognostic marker is one, denoted by S, that is binary:

$$S = \begin{cases} 0 & \text{for a negative marker result;} \\ 1 & \text{for a positive marker result.} \end{cases} \qquad (2.2)$$

2.4.2.1 Statistics to quantify prognostic value

There are three different pairs of statistics, with each statistic denoted by a three-letter acronym, that can be used to describe a binary prognostic marker.

2.4.2.1.1 True positive and false positive fractions
The *true positive fraction* (*TPF*) and the *false positive fraction* (*FPF*) are defined as follows:

$$\text{TPF} = P[S = 1 | T = 1], \qquad (2.3)$$

$$\text{FPF} = P[S = 1 | T = 0]. \qquad (2.4)$$

The FPF is the proportion with a positive result among those who have a fair outcome and the TPF is the proportion with a negative result among those who have a poor outcome. Without loss of generality, in this thesis, a positive marker value is intended to indicate a poor outcome and a negative marker value is intended to indicate a fair outcome. Therefore, a case where $S = 1$ and $T = 1$ is a *true positive* and conversely, a case where $S = 1$ and $T = 0$ is a *false positive.*

A good prognostic marker will be one for which the TPF is high and the FPF is low. The TPF is also known as the *sensitivity* and $1 - FPF$ as the

specificity.

2.4.2.1.2 Positive and negative predictive values The *positive predictive value* (*PPV*) and the *negative predictive value* (*NPV*) are defined as follows:

$$PPV = P[T = 1|S = 1], \qquad (2.5)$$

$$NPV = P[T = 0|S = 0]. \qquad (2.6)$$

The *PPV* is the fraction of those with a positive result that go on to have a poor outcome to treatment and the *NPV* is the fraction of those with a negative result that are go on to have a fair outcome. A good prognostic marker will have a high *PPV* and a high *NPV*.

It is important to note that, unlike *FPF* and *TPF*, the predictive values are affected by the proportion of the total that have a poor outcome, that is $\rho = P[T = 1]$. This be seen by looking at a useless marker, S^*, that is completely independent of the true endpoint, ($P[T = t|S^* = s] = P[T = t]$). For such a marker $PPV = P[T = 1|S^* = 1] = P[T = 1] = \rho$ and $NPV = P[T = 0|S^* = 0] = P[T = 0] = 1 - \rho$. The *PPV* will always be bounded below by the overall probability of a poor outcome. Consider a very good prognostic marker with $TPF=P[S = 1|T = 1] = 0.98$ and $FPF=P[S = 1|T = 0] = 0.05$ but the proportion of those that fail treatment is low, $\rho = 0.02$. Then:

$$PPV = P[T = 1|S = 1], \qquad (2.7)$$

$$= P[S = 1|T = 1]\frac{P[T = 1]}{P[S = 1]}, \text{ by Bayes' theorem,} \qquad (2.8)$$

$$= TPF\frac{\rho}{\rho TPF + (1 - \rho)FPF} \qquad (2.9)$$

$$= 0.29, \qquad (2.10)$$

$$NPV = P[S = 0|T = 0]\frac{P[T = 0]}{P[S = 0]} \qquad (2.11)$$

$$= (1 - FPF)\frac{1 - \rho}{\rho(1 - TPF) + (1 - \rho)(1 - FPF)} \qquad (2.12)$$

$$= 0.999. \qquad (2.13)$$

A low *PPV* could be the result of a small probability of a poor outcome or a poor prognostic marker (in this example it was the former). A high

PPV could be the result of a large probability of a poor outcome or a good prognostic marker.

In spite of this drawback, *PPV* and *NPV* answer a more practical question than the *TPF* and the *FPF* for a clinician treating a patient having just received the result of the test. Given a positive marker result, the *PPV* is the probability that this patient will go on to have a poor outcome and given a negative marker result, the *NPV* is the probability that this patient will go on to have a fair outcome.

2.4.2.1.3 Positive and negative diagnostic likelihood ratios Likelihood ratios are another way of describing a prognostic marker. The *positive diagnostic likelihood ratio* (*DLR*$^+$) and the *negative diagnostic likelihood ratio* (*DLR*$^-$) are defined are follows:

$$DLR^+ = \frac{TPF}{FPF} = \frac{P[S = 1|T = 1]}{P[S = 1|T = 0]}, \tag{2.14}$$

$$DLR^- = \frac{1 - TPF}{1 - FPF} = \frac{P[S = 0|T = 1]}{P[S = 0|T = 0]}. \tag{2.15}$$

These two statistics are the ratio of the likelihood of the marker being positive or negative in those who have a poor outcome versus those who are a fair outcome. Unlike the two previous pairs, these two ratios can take values in the interval $(0, \infty)$. A perfect prognostic marker has $DLR^+ = \infty$ and $DLR^- = 0$. $DLR^+ > 1$ indicates that a positive marker result is more likely in an individual who will have a poor outcome than one who will have a fair outcome and this is clearly desirable. Similarly, $DLR^- \leq 1$ indicates that a negative marker result is more likely in an individual who will have a fair outcome than one who will have a poor outcome, and this is also clearly desirable. These two likelihood ratios are functions of the *TPF* and the *FPF* alone and therefore, unlike the predictive values, do not depend on the overall probability of a poor outcome, p.

Define the odds of a poor outcome prior to knowledge of the prognostic marker, $\pi(T)$, and the odds of a poor outcome given the prognostic marker,

$\pi(T|S)$, as:

$$\pi(T) = \frac{P[T=1]}{P[T=0]} \text{ and} \qquad (2.16)$$

$$\pi(T|S=s) = \frac{P[T=1|S=s]}{P[T=0|S=s]}. \qquad (2.17)$$

It follows that:

$$\pi(T|S=1) = DLR^+ \pi(T) \text{ and} \qquad (2.18)$$

$$\pi(T|S=0) = DLR^- \pi(T), \qquad (2.19)$$

and therefore DLR^+ and DLR^- quantify the change in odds and are therefore Bayes factors relating the prior and the posterior distributions.

These likelihood ratios are not used extensively in practice and are more appropriate for case-control studies (Pepe, 2003). They will therefore not be considered further in this thesis.

2.4.2.2 Empirical estimation of prognostic statistics

If, considering the data in question, the proportion of the response to treatment has not been fixed in advance (this proportion would be fixed in a case-control study), these quantities can be estimated *empirically* from the following table:

	$T=0$	$T=1$	
$S=0$	a	b	$n_0^S = a+b$
$S=1$	c	d	$n_1^S = c+d$
	$n_0^T = a+c$	$n_0^T = b+d$	$N = a+b+c+d$

Here, for example, a patients have negative marker result and a fair outcome to treatment (these are true negatives) and b patients also have a negative marker result but have a poor outcome to treatment (these are false negatives). The statistics can be estimated as follows:

$$\widehat{TPF} = \frac{d}{b+d}, \quad \widehat{FPF} = \frac{c}{a+c}, \qquad (2.20)$$

$$\widehat{PPV} = \frac{d}{c+d}, \quad \widehat{NPV} = \frac{a}{a+b}, \qquad (2.21)$$

where, for example, $\widehat{TPF}$ denotes an estimate of the statistic TPF.

2.4.2.3 Regression based estimators

Regression modelling is a more sophisticated method for estimating the statistics to quantify prognostic value. It allows for the adjusting for covariates and, importantly for the application in this thesis, for clustering within groups.

To estimate the TPF and the FPF, a model can be fit as follows:

$$g\left(P[S = 1|T]\right) = \alpha + \beta T, \tag{2.22}$$

where $g()$ is a link function of choice and α and β are the model parameters to be estimated. The logit link function is the usual link function chosen and it will be used in this thesis from this point. The parameter estimates from this model give:

$$\text{logit } P(\widehat{TPF}) = \hat{\alpha} + \hat{\beta}, \tag{2.23}$$

$$\text{logit } P(\widehat{FPF}) = \hat{\alpha}. \tag{2.24}$$

Similarly, to estimate the PPV and the NPV, the following model can be fit:

$$g\left(P[T = 1|S]\right) = \gamma + \delta S, \tag{2.25}$$

where γ and δ are the model parameters to be estimated. The parameter estimates from this model give:

$$\text{logit } P(\widehat{PPV}) = \hat{\gamma} + \hat{\delta}, \tag{2.26}$$

$$\text{logit } P(1 - \widehat{FPF}) = \hat{\gamma}. \tag{2.27}$$

Additional covariates, X, can be included in the models to give estimates of these statistics adjusted for these covariates. Clustering effects can also be incorporated into the regression model using random effects.

2.4.2.3.1 Comparing prognostic markers
Given two prognostic markers S_A and S_B, if $TPF(S_B) > TPF(S_A)$ and $FPF(S_B) < FPF(S_A)$, then S_B is clearly superior to S_A. It is not so straightforward to compare prognostic markers if only one of these two conditions hold.

30

Pepe (2003) show that the statements (i) and (ii) below are equivalent and therefore a marker that is superior in TPF and FPF is also superior in PPV and NPV.

(i) $TPF(S_B)$ $> TPF(S_A)$ and $FPF(S_B)$ $< FPF(S_A)$,

(ii) $PPV(S_B)$ $> PPV(S_A)$ and $NPV(S_B)$ $> NPV(S_A)$.

Prognostic markers can be compared using a regression model. Let:

$$I_S = \begin{cases} 0 & \text{for test A and} \\ 1 & \text{for test B,} \end{cases} \tag{2.28}$$

and fit the model:

$$g\left(P[S = 1|T]\right) = (\alpha_A + \alpha_B I_S) + (\beta_A + \beta_B I_S)T, \tag{2.29}$$

where S is the combined results from the two tests ($S_A = S|I_S = 0$ and $S_B|I_S = 1$). The parameter estimates from this model give:

$$\ln(oTPF(B, A)) = \hat{\alpha}_B + \hat{\beta}_B, \tag{2.30}$$

$$\ln(oFPF(B, A)) = \hat{\beta}_B, \tag{2.31}$$

where $oTPF(B, A)$ is the odds ratio comparing $TPF(S_B)$ with $TPF(S_A)$ and $oFPF(B, A)$ is the odds ratio comparing $FPF(S_B)$ with $FPF(S_A)$:

$$oTPF(B, A) = \frac{TPF(S_B)(1 - TPF(S_A))}{TPF(S_A)(1 - TPF(S_B))}, \tag{2.32}$$

$$oFPF(B, A) = \frac{FPF(S_B)(1 - FPF(S_A))}{FPF(S_A)(1 - FPF(S_B))}. \tag{2.33}$$

2.4.3 The Evaluation of continuous and categorical prognostic markers

In this section we consider evaluating a continuous prognostic marker, S. Given a point of dichotomy, c, a binary test can be defined as follows:

$$\text{positive if } S \geq c, \tag{2.34}$$

$$\text{negative if } S < c. \tag{2.35}$$

The TPF and FPF can therefore be defined as:

$$TPF(c) = P[S \geq c | T = 1] \qquad (2.36)$$

$$FPF(c) = P[S \geq c | T = 0]. \qquad (2.37)$$

2.4.3.1 Receiver Operating Characteristic (ROC) Curve

Which point of dichotomy, c, leads to the most effective binary prognostic marker? The superior point of dichotomy will be that which maximises $TPF(c)$ while minimising $FPF(c)$. The point which maximises $TPF(c)$ will not necessarily be the same as that which minimises $FPF(c)$, but the *Receiver Operating Characteristic (ROC) curve* is used to explore the effect of varying the point of dichotomy (Erdreich and Lee, 1981). The ROC curve is the set of all possible values of TPF and FPF across the range of S:

$$ROC = \{(FPF(c), TPF(c)), c \in (-\infty, \infty)\}. \qquad (2.38)$$

The following results follow from the definitions of $TPF(c)$ and $FPF(c)$:

$$\lim_{c \to -\infty} TPF(c) = \lim_{c \to -\infty} FPF(c) = 1, \qquad (2.39)$$

$$\lim_{c \to \infty} TPF(c) = \lim_{c \to \infty} FPF(c) = 0. \qquad (2.40)$$

Let $ROC(t)$ be the function that maps $t = FPF(c)$ onto $TPF(c)$. Both $TPF(c)$ and $FPF(c)$ are monotone increasing functions of c and therefore the function $ROC(t) = TPF(c)$ is an increasing function of $t = FPF(c)$ with domain and range on the unit interval with $ROC(0) = 0$ and $ROC(1) = 1$.

For a wholly uninformative prognostic marker, S^*, $TPF(c) = FPF(c)$ and the ROC curve for this marker will by the line $ROC(t) = t$. This diagonal line is called the *chance line*. The ROC curve of any prognostic marker can therefore be compared to this line since it must satisfy $ROC(t) \geq t \forall t \in [0,1]$[1]. Figure 2.1 shows a ROC curve for a simulated prognostic marker.

One of the key advantages with the ROC curve is that no assumptions are made about the distribution of the continuous marker, S. The ROC curve is invariant to a strictly increasing transformation of S. The ROC curve is a plot of TPF against FPF. A similar curve can be defined for PPV and NPV, but

---

[1]This inequality, that $TPF(c) \geq FPF(c)$, follows from the assumption that a positive marker is intended to indicate a poor outcome to treatment and therefore the probability of a poor outcome given a positive result is greater than the probabilitiy of a poor outcome given a negative result.
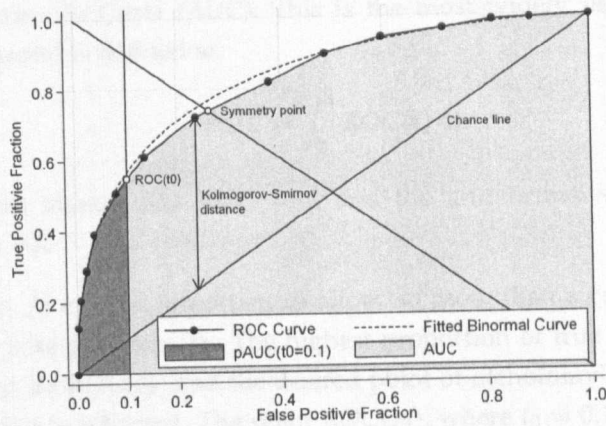
Figure 2.1: Example receiver operating characteristic curve evaluating a simulated prognostic marker.

this has not been found to be particularly insightful when it has been explored (Pepe, 2003).

Despite being a useful tool for displaying the value of a prognostic marker, there is still no definitive measure for identifying the point of dichotomy that maximises the prognostic value of the marker or for comparing markers, although there are a number of useful summary measures (see Figure for graphical representation).

- *Kolmogorov-Smirnov distance.* The maximum vertical distance between the ROC curve and the chance line is a useful summary measure for a prognostic marker. This distance is defined as:

$$KS = \max_{t \in [0,1]} |ROC(t) - t| = \max_{c \in (-\infty, \infty)} |TPF(c) - FPF(c)|. \qquad (2.41)$$

A perfect marker has $KS = 1$ and the uninformative marker has $KS = 0$. A good choice of the point of dichotomy could be the point at which this maximum distance is achieved, that is c^* where $|TPF(c^*) - FPF(c^*)| = KS$.

- *Symmetry point, t_{sym}.* The symmetry point is defined as the point on the ROC curve where sensitivity is equal to specificity, $ROC(t_{sym}) = 1 - t_{sym}$. A perfect prognostic marker will have $t_{sym} = 0$ and for an uninformative marker, $t_{sym} = 0.5$.

33

- *Area Under the Curve (AUC)*. This is the most widely used summary measure and is defined as:

$$AUC = \int_0^1 ROC(t)\, dt. \tag{2.42}$$

A perfect marker has $AUC = 1$ and the uninformative marker has $AUC = 0.5$.

- *$ROC(t_0)$*. It may be important to allow no more than a certain proportion of false positives, t_0. The highest proportion of true positives will therefore be $ROC(t_0)$ and the desired point of dichotomy will be that at which this is achieved. The point $ROC(t_0)$, where $t_0 = 0.1$, is shown on the graph.

- *$t_1 \mid ROC(t_1)=r_1$*. In a situation of low prevalence, it may be important to have a TPF of no less than a certain value, r_1. In this case, the lowest TPF will be t_1 where $ROC(t_1) = r_1$ and the desired point of dichotomy will be that at which this is achieved.

- *Partial Area Under Curve ($pAUC(t_0)$)*. If only markers with a false positive fraction of less than a certain value, t_0, are of interest, the partial AUC is a useful measure for summarising the behaviour of the marker in this region. It is defined as:

$$pAUC(t_0) = \int_0^{t_0} ROC(t)\, dt. \tag{2.43}$$

This area is shaded on the graph for $t_0 = 0.1$.

2.4.3.2 Binormal Curve

The *binormal curve* is the parametric model for *ROC* curves. Assuming the prognostic marker is normally distributed conditional on the true endpoint,

34

$(S|T = 0) \sim N(\mu_0, \sigma_0^2)$ and $(S|T = 1) \sim N(\mu_1, \sigma_1^2)$, then:

$$TPF(c) = P[S \geq c|T = 1] = \Phi\left(\frac{\mu_0 - c}{\sigma_0}\right), \tag{2.44}$$

$$FPF(c) = P[S \geq c|T = 0] = \Phi\left(\frac{\mu_1 - c}{\sigma_1}\right), \tag{2.45}$$

$$ROC(t) = TPF(c) = \Phi\left(\frac{\mu_0 - \mu_1 + \sigma_1\Phi^{-1}(t)}{\sigma_0}\right) \tag{2.46}$$

$$= \Phi\left(a + b\Phi^{-1}(t)\right), \tag{2.47}$$

where $a = (\mu_0 - \mu_1)/\sigma_0$ and $b = \sigma_1/\sigma_0$. The binormal curve is defined as

$$ROC(t) = \Phi\left(a + b\Phi^{-1}(t)\right). \tag{2.48}$$

The binormal curve provides a good approximation to a wide range of ROC curves that can occur in practice (Pepe, 2003). Since the ROC curve is invariant to strictly increasing functions of S, the *binormal assumption* states that there exists a strictly increasing transformation that transforms the distribution of $S|T$ into a normal distribution. This is a fairly weak assumption explaining the wide-ranging use of the binormal curve and shows that the ROC displays the relationship between distributions $S|T = 1$ and $S|T = 0$, rather than the distributions themselves (Pepe, 2003).

The area under the binormal curve AUC_b can be calculated from the estimates of the parameters of the binormal curve:

$$AUC_b = \Phi\left(\frac{a}{\sqrt{1 + b^2}}\right). \tag{2.49}$$

2.4.3.3 The ROC curve for ordinal prognostic markers

The ROC curve can be defined for a ordinal prognostic marker, S, with ordinal categories $s_1, s_2, \ldots s_k$:

$$TPF(c) = P[S \geq c|T = 1] \tag{2.50}$$

$$FPF(c) = P[S \geq c|T = 0]. \tag{2.51}$$

$$ROC = \{(FPF(s_i), TPF(s_i)), i \in \{1, 2, \ldots k\}\}. \tag{2.52}$$

One of the key advantages in using the ROC curve for an ordinal marker

is that no 'value' is assigned to categories of the marker and therefore no assumption is made about the relative 'distance' between the categories. It doesn't matter whether, for example, the categories are *none, slight, moderate* and *high* or *poor, unsatisfactory, satisfactory,* and *excellent,* the ROC curve will be the same.

This can be seen by defining the binormal curve for an ordinal marker. Let Y be the unobserved continuous latent variable for which only thresholds, S, are observed. That is

$$S = s_i \implies Y = y \in (s_{i-1}, s_i], \qquad (2.53)$$

where s_0 is nominally $-\infty$ and s_{k+1} is nominally $+\infty$. Given the binormal assumption, $Y|T$ can be transformed to have a normal distribution and therefore a binormal curve can be fit to the data. Therefore, the binormal curve for an ordinal marker S will be the same whatever the underlying unobserved variable actually is. Assuming that the binormal curve fits the data, AUC_b is an estimate of the area under the curve for the unobserved continuous variable which could therefore be more useful than the empirical AUC calculated from the categorical marker.

2.4.4 Explained variability: Pseudo-R^2 Statistics

In ordinary least squares (OLS) regression used for a continuous dependent variable, prognostic markers are commonly evaluated using the R^2 statistic. This is a measure of the proportion of variation in the dependent variable that is explained by the model and is interpreted as the combined prognostic strength of the explanatory variables in the model. This statistic does not translate unambiguously to a logistic regression model where the dependent variable is binary. A number of authors have proposed *pseudo-R^2* statistics that measure the proportion of variation explained by the model. Unlike the OLS R^2 statistic, there is no convincing evidence that any one of these statistics definitively measures the proportion of explained variation and should instead be used as scales with which to compare different models (Long, 1997; Long and Freese, 2006). Four pseudo-R^2 statistics have been selected on the basis of their relative ease of interpretation and acceptability for use in this thesis. For each of these statistics, as with positive and negative predictive values, the probability of $T = 1$ is modelled conditional on the surrogate, S,

(whether binary or categorical) using logistic regression.

- *McFadden's R^2*. Also known as the *likelihood-ratio index*, McFadden's R^2 compares the log-likelihood of the full model with all parameters with that of the null model with only the intercept. It is defined to be

$$R^2_{McF} = 1 - \frac{\ln \hat{L}(M_{Full})}{\ln \hat{L}(M_{Null})}.$$

In the logistic regression models used in this thesis, the trial effect is including in the model to control for between-trial variation but is not considered to be a prognostic factor. A more useful statistic, denoted by $R^{2\prime}_{McF}$, would be comparing the log-likelihood of the full model with that of the model containing only the trial effect:

$$R^{2\prime}_{McF} = 1 - \frac{\ln \hat{L}(M_{Full})}{\ln \hat{L}(M_{Trial})}.$$

This can be interpreted as a measure of the improvement in the strength of prediction resulting from the addition of the culture result.

- *Efron's R^2*. Defining $\hat{\pi} = \hat{P}(T = 1|x)$ as the predicted probability of a poor outcome given the explanatory variables, x, Efron's R^2 is defined to be

$$R^2_{Ef} = 1 - \frac{\sum_i (T_i - \hat{\pi}_i)^2}{\sum_i (T_i - \bar{T})^2}.$$

This expression is comparable to that for the OLS R^2 statistic in that the model residuals are squared, summed, and divided by the total variability in the dependent variable. Model residual from a logistic regression model are very different since the dependent variable is binary and so this comparison is limited.

- *McKelvey and Zavoina's R^2*. Considering the binary dependent variable, T, as the realisation of a continuous latent variable $y^* = x\beta + \varepsilon$, McKelvey and Zavoina's R^2 is defined to be

$$R^2_{M\&Z} = \frac{\widehat{Var}(\hat{y}^*)}{\widehat{Var}(\hat{y}^*) + \widehat{Var}(\varepsilon)}$$

The variance of the error term ($\widehat{Var}(\varepsilon)$) cannot be calculated since y^*

is not observed and is therefore assumed to be $\pi^2/3$ as recommended by Long and Freese (2006) as the variance of the logistic distribution. Using simulations to give a result that is perhaps not surprising, it has been shown that McKelvey and Zavoina's R^2 most closely approximates the R^2 obtained by fitting the linear regression model on the underlying latent variable (Hagle, 1992; Windmeijer, 1995).

• *Count R^2.* Treating individuals with a predicted probability of 0.5 or greater as having a predicted outcome $T = 1$ and those with a predicted probability of less than 0.5 as having a predicted outcome $T = 0$, the count R^2 is simply the proportion of individuals for whom their predicted outcome matches their actual outcome. The model with no explanatory variables will always have a count R^2 equal to the proportion of positive or negative outcomes in the population, whichever is greater (the model will either predict all individuals with a positive or all with a negative outcome). The count R^2 must therefore be compared to that from the model with no explanatory variables. It is defined to be

$$R_C^2 = \frac{\text{Number Correct}}{\text{Total}}.$$

2.5 The Prentice Criteria

Having presented some methods for the evaluation of prognostic markers, subsequent sections in this chapter are devoted to the more complex and more controversial methods for the evaluation of surrogate markers. The first area for consideration, both logically and chronologically, must be *The Prentice Criteria*.

In addition to providing a clear definition described in section 2.2, Prentice (1989) proposes three operational criteria for identifying surrogate endpoints for a time-to-event clinical endpoint, T.

1. He introduces the notion that a surrogate endpoint should fully capture the effect of the treatment on the true endpoint and expresses this mathematically:

$$\lambda_T(t; S, Z) \equiv \lambda_T(t; S), \tag{2.54}$$

where T is the true time-to-event endpoint, λ_T is the hazard function, S is the surrogate endpoint and Z is an indicator variable corresponding

to the treatment. The hazard function conditional on S and Z should be equal to the hazard function conditional on S only for all t.

He provides two additional restrictions:

2. The surrogate must, in some way, be associated with the true endpoint:

$$\lambda_T(t; S) \neq \lambda_T(t), \tag{2.55}$$

and

3. The effect of the treatment on the distribution of the surrogate must alter the average endpoint risk:

$$E_S[\lambda_T(t; S)|Z] \neq E_S[\lambda_T(t; S)] \forall t > 0, \tag{2.56}$$

with the expectation taken with respect to S.

Prentice shows that if these three conditions hold, then the null hypothesis that there is no treatment effect on the true endpoint is equivalent to the null hypothesis that there is no treatment effect on the surrogate endpoint. These three conditions are commonly known collectively as the *Prentice Criteria* (although the exact number and the order that they are given can differ between authors). The third condition is sometimes omitted when discussing the Prentice criteria (for example, the entry on Surrogate Endpoints in the Encyclopedia of Biostatistics, Fleming et al. (1998)); indeed, Prentice himself describes it as 'innocuous'. It is nevertheless important and 'necessary to avoid pathological relationships for (non-binary) S in which [criterion 1] and [criterion 2] hold, but the dependence of the hazard rate on S does not effect the marginal hazard for T (averaged over S) at any rate of T ' (Prentice, 2005)

Though the criteria are given in terms of a time-to-event true endpoint, T, these criteria can be restated for a binary endpoint, T, using probabilities:

1. $P(T = 1|S = s, Z = z) \equiv P(T = 1|S = s)$,

2. $P(T = 1|S = s) \neq P(T = 1)$,

3. $E[P(T = t|S = s, Z = z)] \neq E[P(T = t|S = s)]$

39

2.5.1 Exploration of the Prentice Criteria

2.5.1.1 The Difference between Correlates and Surrogates

Baker and Kramer (2003) explore the relationship between correlation and surrogacy. They use a simple graphical method to show for Gaussian endpoints that even when the surrogate is perfectly correlated with the true endpoint within each of the two treatment groups, the difference in means of the surrogate between treatment groups can be in the opposite direction to the difference in means on the true endpoint resulting in conclusions based on the surrogate endpoint contradicting conclusions based on the true endpoint. A simulated example of this is given in Figure 2.2 on the next page.

In this example, the candidate surrogate and true endpoints are perfectly correlated within each of the treatment and control groups respectively. Despite this, the mean surrogate outcome in the treatment group (equal to 267 and marked on the graph) is larger than the mean surrogate outcome in the control group (154) while the mean true outcome in the treatment group (160) is smaller than the mean true outcome in the control group (256), leading to conflicting conclusions for the effect of the experimental treatment. The experimental treatment appears to lead to a decrease in the true endpoint and an increase in the surrogate endpoint, despite these two being perfectly correlated within treatment group.

The authors conclude that only when the relationship between the surrogate and the true endpoint is the same within each treatment group do the null hypotheses based on the surrogate endpoint and the true endpoint correspond. This is equivalent to the first Prentice criterion that the surrogate marker captures the full effect of the treatment on the true endpoint.

As perfect correlation is not a sufficient condition, neither it is necessary, as the second Prentice criterion requires only that the surrogate is in some way associated with the true endpoint.

Fleming and DeMets (1996) also point out that a commonly held misconception is that if an outcome correlates well with the true endpoint then it can be used as a valid surrogate. They perform a systematic review of the use and evaluation of surrogates in several disease areas concluding that in practice, the Prentice Criteria are rarely met for surrogates that are in use and, despite this danger, it is rare for a surrogate endpoint to be carefully validated. They note that a surrogate marker that captures only 50% of the treatment effect is as uninformative as tossing a coin. They illustrate some examples where a
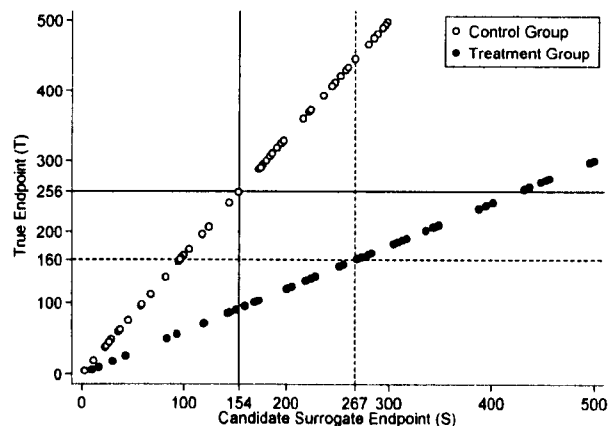
Figure 2.2: A simulated example showing a candidate surrogate marker that correlates perfectly with the true endpoint, but would be a disastrous surrogate. Highlighted are the mean surrogate in the control group (154) and the treatment group (267) and the mean true endpoint in the control group (256) and the treatment group (160).

marker correlated with the true endpoint would fail as a surrogate, using the language of *disease causal pathways*. A disease may have several causal pathways that result in the clinical endpoint. The surrogate may be in the pathway of only a subset of these and the intervention may act on a different subset such that the surrogate does not fully capture the treatment effect. The intervention may also have unintended or unrecognised mechanisms of action on the clinical endpoint independent of the disease process and the surrogate endpoint. The authors show that the ideal setting for a valid surrogate occurs when the surrogate is on the only causal pathway of the disease process and the entire effect of the intervention is mediated through the surrogate. To identify such a situation requires considerable understanding of the disease and the intervention.

2.5.1.2 Criticism of the Prentice Criteria

Despite Prentice's work frequently and rightly being described as 'seminal' (e.g., Taylor and Wang, 2002; Weir and Walley, 2005) or 'landmark' (Buyse and Molenberghs, 1998), there has been much criticism.

41

Freedman et al. (1992) extend the Prentice Criteria for a binary rather than time-to-event clinical outcome using logistic regression to model the probability of the true endpoint occurring conditional on the surrogate and the treatment group. They propose a two-step procedure testing first for an interaction between the surrogate endpoint and the treatment group. If the interaction is found to be statistically significant then the performance of the surrogate marker is dependent on the treatments being compared and there is therefore evidence invalidating the surrogate.

An interaction between the marker and the treatment is sometimes desirable. It may be the case that an experimental treatment has an important effect on the true endpoint when compared to the control treatment in those with a positive marker result, but there is no effect in those with a negative marker result. In this scenario, the marker could be used to tailor the treatment strategy for individual patients (those with a positive marker result respond to the experimental treatment, but those with a negative marker result do not). This marker, however, cannot be used as a valid surrogate.

If there is no evidence of interaction, the second step involves fitting the model with no interaction term and testing for a treatment effect. If there is a significant treatment effect then there again is evidence invalidating the surrogate. A statistically significant result in a test of a null hypothesis provides evidence against the null hypothesis, but failure to achieve a statistically significant result does not constitute evidence in favour of the null hypothesis. Since the Prentice Criteria are based on hypothesis tests, the authors observe that this procedure of testing the Prentice Criteria only provides a means of dismissing undesirable intermediate endpoints rather than identifying valid surrogate endpoints. Prentice defends his criteria pointing out that they are for defining when an intermediate endpoint can serve as a surrogate and that evaluation is a completely different matter (Prentice, 2005). Indeed, 'the main challenge with validating surrogate endpoints in a hypothesis testing framework is evaluating the surrogate endpoint when the Prentice Criterion cannot be rejected' (Baker, 2006a).

The use of hypothesis testing in the Prentice criteria does mean that only failed surrogates can be identified; there is no framework for accepting a true surrogate (as failure to reject the null hypothesis could be due to insufficient information rather than it necessarily being correct).

Buyse and Molenberghs (1998) suggest that a surrogate endpoint would be useful only if there was no evidence to reject the hypothesis test evaluating

the first Prentice criterion in a large number of studies and that the Prentice Criteria are 'necessary and sufficient to establish the validity of binary surrogate endpoints, but not of more complex surrogate endpoints'. Prentice refutes this stating that the authors have 'evidently overlooked' the third Prentice criterion (Prentice, 2005) without identifying the specific mistake in their algebra. He argues that the three Prentice criteria together do ensure equality of hypothesis tests for both binary and non-binary surrogates.

Berger (2004) neatly summarises the process required in evaluating surrogate markers: 'If improvement in a [candidate] surrogate endpoint does not itself confer patient benefit, then consideration must be given to the extent to which improvement in a surrogate endpoint implies improvement in the true clinical endpoint of interest'. The author proposes a 'validity criterion' which corresponds to Prentice's definition of a surrogate endpoint (see section 2.2), that 'a valid between-group analysis of the surrogate endpoint constitutes also a valid analysis of the true clinical endpoint'. Following Buyse and Molenberghs (1998), he claims that the Prentice Criteria do not imply the 'validity criterion'. He attempts to construct a counter-example where the Prentice criteria are met but the 'validity criterion' is not. He succeeds, but manages to do so by ignoring the third Prentice criterion which his counter-example does not satisfy. This appears to be because, rather than using Prentice's original paper, he is using Molenberghs et al. (2001) for his list of the Prentice criteria where the third criterion is not included.

It is clear that the condition of a surrogate fully capturing the treatment effect on the true endpoint is difficult to satisfy. Fleming (1994) instead suggest the Prentice criteria are restrictive and should rather be 'an ideal to be kept in mind'. He suggests that when the candidate for surrogacy does not fully capture the treatment effect on the true endpoint, but does capture some of the treatment effect, the candidate endpoint should be termed an *auxiliary endpoint* and propose that information on this endpoint can be used to add weight to the standard analysis or to aid imputation of missing data. Cox (1983) proposed a similar idea, before any of this discussion around surrogate endpoints, in the context of right-censored survival data.

2.6 The Proportion of Treatment Effect and Other Summary Measures

2.6.1 The Proportion of Treatment Effect

Following their criticisms, Freedman et al. (1992) attempt to overcome the problems with the Prentice Criteria and quantify surrogacy by defining *Proportion of Treatment Effect*, *PTE*, that is an estimate of the proportion of the effect of the treatment on the true endpoint that is captured by the surrogate. They fit the following two models:

$$T_k = a_1 + b_1 Z_k + \varepsilon_k$$
$$T_k = a_2 + b_2 Z_k + c_2 S_k + \varepsilon_k,$$

where T_k is the true clinical endpoint, S_k is the surrogate endpoint and Z_k is the treatment indicator for the kth patient. T and S are both Gaussian normal. They define:

$$\text{PTE} = \frac{\hat{b}_1 - \hat{b}_2}{\hat{b}_1}.$$

Conceptually, the PTE is the proportion of the treatment effect on the true endpoint that is 'explained' or 'removed' on adjusting for the surrogate. If the surrogate captures the entire treatment effect, it would be expected that $\hat{b}_2$ were very small and therefore PTE$\simeq$ 1. If the surrogate fails to capture any of the treatment effect, it would be expected that $\hat{b}_2 \simeq \hat{b}_1$ and therefore PTE$\simeq$ 0.

An obvious drawback with this *proportion* is that it will not necessarily be less than one, nor will it necessarily be positive and so is not strictly a 'proportion'. Buyse and Molenberghs (1998) shows the treatment effect on the surrogate and true endpoint must be in the same direction otherwise the PTE will lie outside the interval [0,1]. Hughes (2002) observes that if the treatment has adverse effects on the true endpoint that are not mediated through the candidate surrogate endpoint, the PTE will be artificially inflated and a poor surrogate will appear to capture a large proportion of the treatment effect.

Freedman et al. (1992) suggest that a surrogate could be deemed important if the lower limit of this confidence interval of the PTE is greater than a critical value such as 0.5 or 0.75. Using Fieller's theorem to calculate standard errors, they show that, to make meaningfully precise estimates of the PTE,

unadjusted treatment effects on the true endpoint must be more than 4 times their standard errors.

Lin et al. (1997) extend this idea, defining the PTE for a time-to-event outcome using the proportional hazards model and constructing appropriate confidence intervals. A letter in response to this paper (Flandre and Saidi, 1999) reviews several papers where the PTE has been used to evaluate surrogate endpoints. They observe a large range of values for the PTE and very wide confidence intervals, highlighting several examples where even the point estimate of the PTE lies outside the range [0,1]. The authors recommend that the use of the PTE should be discontinued. Table 2.1 gives a number of examples of the use of the PTE showing the huge variability and large confidence intervals.

True Endpoint	Surrogate Endpoint	Treatment	PTE (95% CI)
Coronary Heart Disease	Serum Cholesterol at 1 year[a]	Choles--tyramine	0.50 (0.07,5.91)
Progression to AIDS	75% drop in HIV1-RNA[b]	AZT	0.59 (0.13,1.12)
	HIV1-RNA with CD4[b]	AZT	0.79 (0.27,1.45)
	Net CD4+%[c]	AZT	0.74
	RNA levels at week 16[d]	AZT+ddI	1.83 (0.79,2.90)
	RNA levels at week 16[d]	AZT+ddC	2.49 (0.83,4.16)

[a]Freedman et al. (1992)
[b]O'Brien et al. (1996)
[c]Choi et al. (1993)
[d]Delta Coordinating Committee and Virology Group (1999)

Table 2.1: Some examples of the use of the Proportion of Treatment Effect (PTE)

Bycott and Taylor (1998) follow on from the work of Lin et al. (1997) evaluating the statistical properties of the PTE using Monte Carlo simulations showing that the PTE has 'tremendous variability' and may in fact have considerable bias towards zero. The authors also show that the case when PTE has reasonable precision, when the treatment effect is strong, is the very situation when it seems most unlikely that a surrogate endpoint will capture all of the treatment effect!

De Gruttola et al. (1997) look at the calculation of the PTE when the treatment has many intended and unintended mechanisms of action. In this situation it is unlikely that this surrogate would be in all of the causal pathways (as

earlier discussed by Fleming and DeMets (1996)). They show that a surrogate endpoint that captures only a small fraction of the change in the true endpoint induced by the treatment may appear to capture a proportion close to 1 if there are unintended and unrecognised treatment effects on the true endpoint not mediated via the surrogate. They conclude by saying that, for the PTE to be identifiable and even for a surrogate to be reliable 'it is required that the biology of disease and all important effects of this intervention (including adverse effects) be fundamentally understood' (De Gruttola et al., 1997).

2.6.2   Extensions of the PTE

Li et al. (2001) also criticise the PTE, highlighting the fact that the numerator and denominator are estimated from two separate models that are unlikely to hold simultaneously. They propose a method for estimating the overall treatment effect and that explained by the surrogate basing both estimates on one model only. They decompose the risk of failure $(T_i = 1)$ to a patient prior to taking treatment, showing that it can be explained by three components: (i) the reduction in risk captured by the surrogate, (ii) the reduction in risk not captured by the surrogate and (iii) the remaining risk to the patient after treatment. They then propose a summary measure, quantifying the reduction in risk captured by the surrogate, rather than the proportion of treatment effect. Using a log-linear model for a binary true endpoint, the *Risk Reduction* is the estimate of the reduction in risk captured by the surrogate divided by the estimate of the total reduction in risk. Consider the model:

$$\log(P\,(T_k = 1)) = \beta_1 + \beta_2 Z_k + \beta_3 S_k, \tag{2.57}$$

where S_k is the surrogate endpoint, T_k the true endpoint and Z_k is the treatment indicator for patient k. The authors show that the proportion of the reduction in risk explained by the surrogate is:

$$RR = \frac{1 - \exp\left(\beta_3 E(\Delta_S)\right)}{1 - \exp\left(\beta_2 + \beta_3 E(\Delta_S)\right)}, \tag{2.58}$$

where $E(\Delta_S)$ is the expectation of the difference in the surrogate marker between the treatment and the placebo groups in the trial. This method uses estimates from the same model and the authors give an explanation for the proportion if it lies outside the interval [0,1] (if the direction of the treatment

46

effect and the surrogate effect on the risk are in the same direction the ratio will be less than 1) but note that intended and unintended effects of the treatments are not distinguished.

Chen et al. (2003) propose another alternative to compute the PTE again using estimates from only one model. They use Cox regression models, allowing for time-varying covariates, and state that the procedure proposed by Li et al. (2001) using log-linear models allowing only for time-independent covariates is too restrictive and not as applicable in the evaluation of surrogate endpoints.

Qu and Case (2006) extend these methods to quantify the treatment effect mediated through multiple surrogates acting on multiple disease pathways, accounting for the causal relationships between these surrogates, although the same authors later state that this method has 'similar disadvantages' to the PTE and those described above (Qu and Case, 2007).

Wang and Taylor (2002) propose another alternative based on a similar idea to Li et al. (2001). They propose a measure, F, as the reduction in risk due to the change in the surrogate induced by the treatment divided by the total reduction in risk between treatment groups, but estimating these risks differently to Li et al. (2001). F is estimated based on the distribution of the true endpoint given the treatment and the surrogate endpoint and the distribution of the surrogate endpoint given the treatment. They suggest that this measure has better properties than the PTE and is more generalisable as it is not tied to a linear model and so can be estimated with more flexibility and fewer assumptions. For binary true and surrogate endpoints, S and T, where Z is the treatment indicator, their measure is:

$$F = [P(T = 1|Z = 0, S = 1) - P(T = 0|Z = 0, S = 0)]$$
$$\cdot \frac{P(S = 1|Z = 0) - P(S = 1|Z = 1)}{P(T = 1|Z = 0) - P(T = 1|Z = 1)}, \tag{2.59}$$

and a complementary form:

$$F' = [P(T = 1|Z = 1, S = 1) - P(T = 0|Z = 1, S = 0)]$$
$$\cdot \frac{P(S = 1|Z = 0) - P(S = 1|Z = 1)}{P(T = 1|Z = 0) - P(T = 1|Z = 1)}, \tag{2.60}$$

Taylor and Wang (2002) consider the PTE and joint models for longitudinal and survival data. They suggest that the PTE described by Lin et al. (1997)

47

is generally not appropriate since it is calculated from estimates from two models that are unlikely to hold simultaneously and often the PTE is greater than one or negative. They propose another PTE based on estimates from only one joint model of a longitudinal surrogate with a failure time true endpoint.

Sarkar and Qu (2007) propose the *excess relative odds* as a replacement for the PTE using regression calibration in the presence of measurement error to remove some of the bias.

Ditlevsen et al. (2005) propose the *Mediation Proportion* as a parallel concept to the PTE for epidemiological observation studies. The formula given is basically the same (and therefore retains the same drawbacks as the PTE) except that it is the proportion of the effect of an *exposure* rather than a *treatment* than is being estimated.

Beyond the PTE, none of these measures have been extensively taken up and used in surrogate marker evaluation. A great deal of work has been done in developing measures that purport to summarise the strength of a surrogate marker, but setbacks have been frequent and a question mark hangs over this whole area.

The final words in this section surveying the plethora of attempts to quantify the value of a surrogate are from Laurence Freedman, writing nearly ten years after first proposing the PTE:

> 'It would be misleading to conclude this article without expressing strong reservations regarding the use of [the PTE], as an ultimate test of the validity of a surrogate endpoint for a new clinical trial; …the method can be usefully applied only in limited situations, for how often do we encounter treatment effects of 5 standard errors or more?
>
> …In general, I feel that the methods described here will find more useful application in identifying intermediate endpoints in epidemiology than in placing a stamp of approval on surrogate endpoints for clinical trials.'

(Freedman, 2001)

2.7 Trial Design based on the use of a Surrogate Endpoint

Day and Duffy (1996) use surrogate endpoints in the design of a trial to compare the effect of different breast screening frequencies on breast cancer mortality rate. The authors state that it had been demonstrated in previous trials that screening by mammography reduces breast cancer mortality in women over 50 years of age. In this trial, the intention was to compare different breast screening frequencies. They comment that, since the purpose of the trial is 'to resolve subsidiary issues', it would be 'perverse' to use the standard primary endpoint of breast cancer mortality and instead design the trial using a surrogate endpoint as the primary endpoint of efficacy. This is a rare example of the use of a surrogate endpoint as a complete substitute for the true endpoint in a clinical trial in practice and led to some interesting discussion.

The surrogate endpoint used is a combination of tumour size, malignancy grade and lymph node status (expressed as a predicted mortality) and the authors show, using data from a previous clinical trial of frequency of screening undertaken in Sweden, that survival from cancers is independent of screening frequency after adjusting for these three variables and therefore that the surrogate endpoint satisfies the Prentice criteria for this particular trial. The authors show that the use of predicted mortality is more powerful than the use of observed mortality in estimating the hazard ratio between interventions since the observed mortality is estimated from the binary outcome of death whilst the predicted mortality is based on the continuous probability of death calculated from three categorical variables. They show that, using this surrogate as the primary endpoint, the trial is reduced in size by a factor of between 2.5 and 3 and the trial duration is reduced from 20 years to 5 years.

Begg and Leung (2000a) criticise Day and Duffy (1996) citing two resulting 'conundrums'. Firstly, the trial design suggests that it is always better to use predictions based on surrogates rather than the true endpoint that is being studied and secondly, the degree by which the predicted endpoint leads to more powerful estimates is inversely proportional to the square of the correlation between the true endpoint and the surrogate, implying that better surrogates are those which are only weakly correlated with the true endpoint. The authors consider these two results counter-intuitive and propose two 'principles' that should be kept in mind in the evaluation of surrogate endpoints:

firstly that the 'best attainable inference' for comparing treatments is based on their effect on the true endpoint and secondly, that the results of a trial using the surrogate endpoints should be 'concordant' with those that would have been found if the true endpoint had been used. The authors also conclude that the Prentice criteria are not a useful basis for evaluating surrogate endpoints.

The authors Day and Duffy and also Prentice respond to this paper (Day and Duffy, 2000; Prentice, 2000). Day and Duffy emphasise two critical assumptions. Firstly, that the Prentice criteria is satisfied for these variables[2] and secondly that the effect of the treatment on breast cancer mortality acts only though the variables on which the surrogate endpoint of predicted mortality is based and therefore the surrogate which is based on three categorical variables will have greater power. They also point out that a surrogate can never be generally applicable and is only valid for the treatments in question—a fact that they suggest Begg and Leung have overlooked. In his comments, Prentice notes that if the Prentice criteria hold then all of the treatment effect on the true endpoint is captured by the surrogate endpoint and any further variability in the true endpoint not explained by the surrogate is simply 'noise'. Therefore, if the Prentice criteria do truly hold, when the surrogate and the true endpoint are only weakly correlated, then there is much noise in the true endpoint and the increase in power in using the surrogate over using the true endpoint is great as shown by Begg and Leung.

Begg and Leung respond to these comments (Begg and Leung, 2000b) rejecting the proposition that, in a 'gold standard' analysis, a true clinical endpoint can have a component of 'noise'. They argue that the situation where only the true endpoint and not the surrogate endpoint is affected by noise, and the situation where there is no treatment effect on the surrogate which is not mirrored in the true endpoint are unrealistic.

Berger (2004) supports Begg and Leung in their criticism. The author refers to any precision that a surrogate endpoint offers over the true endpoint that does not reflect patient benefit as 'pseudo-precision'.The authors link this with their 'validity criterion' (see section 2.5.1): 'The idea that pseudo-precision could be treated as actual precision led us to consider the possibility that per-

---

[2]The Prentice Criteria had been satisfied on these variables in a different clinical trial. This exchange shows that Prentice Criteria can never be tested in a clinical trial where a surrogate endpoint has been substituted for the true endpoint (since the true endpoint has not been measured) and therefore the applicability of any surrogate from one trial to the next must always be taken with some degree of trust.

haps the Begg and Leung criticism was in fact a criticism not of the relevance of the Prentice criteria but rather if the statement that the Prentice criteria imply the validity criterion' (Berger, 2004).

2.8 The Belgian Paradigm

2.8.1 The Relative Effect and the Adjusted Association

Buyse and Molenberghs (1998) also criticise the PTE proposed by Freedman et al. (1992) pointing out that the denominator of the proportion is the estimate of the treatment effect on the true endpoint. It is likely that the imprecision of this estimate will be the motivation for the search for a surrogate and so, correspondingly, the PTE will generally be too poorly estimated and the confidence limits large as has been shown (see Table 2.1 on page 45). Following Freedman et al. (1992), the authors also emphasise that it is important to test for interaction between the treatment and the surrogate as the first step.

If the Prentice Criteria are proposed for *defining* a surrogate endpoint (Prentice, 2005), these authors propose criteria for *validating* or *evaluating* a surrogate endpoint. They initially consider the case of the true and surrogate endpoints being binary and fit the following two models:

$$\text{logit}(P(S_k|Z_k)) = \mu_{ZS} + \alpha Z_k$$
$$\text{logit}(P(T_k|Z_k)) = \mu_{ZT} + \beta Z_k,$$

where, as before, Z_k is the treatment given to the kth patient with T_k and S_k the corresponding true and surrogate endpoints. For this purpose, μ_{ZS} and μ_{ZT} are nuisance parameters estimated from the model.

The authors propose two new quantities: the *relative effect* (RE) and the *adjusted association* (AA). The RE is defined as β/α. The RE is simply the ratio of the log-odds ratio of true endpoint for the treatment effect and the log-odds ratio of the surrogate endpoint for the treatment effect. They propose that a surrogate should be called *valid at the population level* if RE is close to one or perfect if the RE is equal to one.

The AA is an estimate of the association between the true and surrogate endpoints adjusting for the treatment, equal to the log odds ratio of the true endpoints for the surrogate conditional on the treatment, or the coefficient γ_Z

51

from the following model

$$\text{logit}(P(T_k|S_k, Z_k)) = \mu_{ZST} + \gamma_Z S_k + \beta_S Z_k.$$

The authors suggest that a surrogate with a value of AA which is very large should be called *valid at the individual level* and perfect if the value of AA is infinite.

The RE captures the association between the surrogate and the true endpoint at the trial-level and so is of more interest to the trialist looking at an overall treatment effect whereas the AA captures the association at the subject-specific-level and so is of more interest to the clinician dealing with individual cases of the condition. Bringing these two measures together, the authors define a *perfect surrogate* as one for which the surrogate is perfect at both the subject-specific and population levels.

The authors also extend their methods appropriately to take account of surrogate and true endpoints that are continuous, fitting the two models:

$$S_k|Z_k = \mu_{ZS} + \alpha Z_k + \varepsilon_{S_k}$$
$$T_k|Z_k = \mu_{ZT} + \beta Z_k + \varepsilon_{T_k},$$

where ε_{S_k} and ε_{T_k} are zero mean correlated error terms with $\text{Var}(\varepsilon_{S_k}) = \text{Var}(\varepsilon_{T_k}) = 1$ and $\text{Cov}(\varepsilon_{S_k}, \varepsilon_{T_k}) = \rho$. Again, RE=$\beta/\alpha$ and the AA is the association between the two endpoints adjusting for the treatment, AA=ρ. The PTE proposed by Freedman et al. (1992) in this situation is therefore equal to $\rho(\alpha/\beta)$=AA/RE. The authors suggests that PTE is less desirable since it is a composite quantity.

This paper brought an important development in the statistical evaluation of surrogate endpoints, namely the separation into individual-level and trial-level surrogacy. If the trial-level surrogacy is of most interest then the individual-level parameter can be viewed almost as a nuisance parameter. The trial-level parameter would then be used rather than having to consider a composite parameter that combines the two resulting in a more noisy parameter as was the case earlier. It is clear that the RE depends greatly on the scale of the continuous variables S and T and is therefore not really a useful measure (Taylor and Wang, 2002), but the principles (giving rise to what, in this thesis, is called the *Belgian Paradigm*) that were the basis of the continuing work described below became very influential.

2.8.2 Meta-Analysis

These ideas are further explained in Buyse et al. (2000b) pointing out that the confidence limits for RE are smaller than for PTE since the denominator of the ratio is the treatment effect on the surrogate which will usually be measured with more precision than the treatment effect on the true endpoint, but admit that despite this, the confidence interval is still wide. In this paper, the authors suggest that, rather than using the RE as a criteria for validating a surrogate endpoint, it 'may be useful to predict the effect of treatment upon the true endpoint, having observed the effect of treatment upon the surrogate endpoint'. For a useful prediction, the RE must be estimated with precision requiring a large number of observations—a number that is not usually available in a single trial. It is necessary to evaluate a marker across a number of trials and treatment comparisons to have any confidence in the value of the marker. A marker that has been evaluated as a surrogate across a number of trials carries much greater weight than one which has only been validated in a single trial. These authors therefore recommend extending these ideas to the meta-analysis of several trials.

The notion of the RE is extended by Buyse et al. (2000a) for multiple trials. These authors propose extending the models above across $i = 1, \ldots, N$ trials each with $k = 1, \ldots, n_i$ subjects for Gaussian normal S and T:

$$S_{ik}|Z_{ik} = \mu_{S_i} + \alpha_i Z_{ik} + \varepsilon_{S_{ik}}$$
$$T_{ik}|Z_{ik} = \mu_{T_i} + \beta_i Z_{ik} + \varepsilon_{T_{ik}},$$

where μ_{S_i} and μ_{T_i} are trial-specific intercepts, α_i and β_i are trial-specific treatment effects for trial i and $\varepsilon_{S_{ik}}$ and $\varepsilon_{T_{ik}}$ are correlated normally distributed error terms with zero mean and covariance matrix:

$$\Sigma = \begin{pmatrix} \sigma_{SS} & \sigma_{ST} \\ \cdot & \sigma_{TT} \end{pmatrix}.$$

The authors proposed fitting a linear mixed model with random intercepts

and slopes assuming:

$$
\begin{pmatrix} \mu_{S_i} \\ \mu_{T_i} \\ \alpha_i \\ \beta_i \end{pmatrix} = \begin{pmatrix} \mu_S \\ \mu_T \\ \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} m_{S_i} \\ m_{T_i} \\ a_i \\ b_i \end{pmatrix},
$$

where the first term on the right-hand side are the fixed effects and the second term are the random effects are assumed to follow a mean zero normal distribution with covariance matrix

$$
D = \begin{pmatrix} d_{SS} & d_{ST} & d_{Sa} & d_{Sb} \\ & d_{TT} & d_{Ta} & d_{Tb} \\ & & d_{aa} & d_{ab} \\ & & & d_{bb} \end{pmatrix}.
$$

The purpose of validating a surrogate endpoint will be to predict the treatment effect on the true endpoint. The authors therefore derive the following measure:

$$
R^2_{trial} = R^2_{b_i | m_{S_i}, a_i} = \frac{\begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}}{d_{bb}}.
$$

They suggest a surrogate should be called *valid at the trial level* if R^2_{trial} is 'sufficiently close to one' and perfect if equal to one. In analogy to the treatment adjusted association (AA) defined above, the authors derive a second measure as the association between the endpoints conditional on the treatments given:

$$
R^2_{indiv} = R^2_{\varepsilon_{T_i} | \varepsilon_{S_i}} = \frac{\sigma^2_{ST}}{\sigma_{SS}\sigma_{TT}}.
$$

They suggest a surrogate should be called valid at the individual level if R^2_{indiv} is 'sufficiently close to one' and perfect if equal to one. The situation with $R^2_{trial} \simeq 1$, indicating that the surrogate and true endpoints were associated at the trial-level, and $R^2_{indiv} \ll 1$, would cast doubt over the biological plausibility of the relationship at the individual-level.

Molenberghs et al. (2001) extend this approach to consider the case when one of the surrogate and true endpoints is binary and the other continuous.

For the single-trial case, they propose formulations for the RE and the AA using several different models including a probit formulation considering the binary variable as a latent variable and one proposed by Dale (1986). Their methods involve the use of copulas to describe the joint distribution of the true and surrogate endpoints which add considerable complexity over the simple case where both the true and surrogate endpoints are Gaussian normal (as the joint distribution is simply multivariate normal). The authors extend this probit formulation to the multi-trial case deriving analogous measures R^2_{trial} and R^2_{indiv} for this combination of endpoints. Renard et al. (2002) consider the case when both endpoints are binary and define R^2_{ind} as the correlation between the two latent variables $\tilde{S}$ and $\tilde{T}$ using a bivariate probit model. In the context of longitudinal surrogate or true endpoints (or both), Alonso et al. (2003) propose the *variance-reduction factor* (*VRF*) and Alonso et al. (2004a) propose R^2_Λ as replacements for R^2_{ind}.

Gail et al. (2000) propose a similar meta-analytic approach, drawing on Daniels and Hughes (1997). They introduce methods for a more general class of models, not necessarily assuming both Gaussian endpoints, and use bootstrap methods to take into account the uncertainty in estimating variance components in the model. Hughes (2002) highlight the importance of joint models to further understand the individual-level association and commends the methods developed by Gail et al. (2000).

Molenberghs et al. (2002) summarise the single-trial and multi-trial approaches described in this section for a non-statistical audience applying the methods to five case studies. They identify a strong drawback to the RE that its calculation from data on a single trial requires the unverifiable assumption that it should be constant across a class of trials, that is the relationship between the treatment effects on the surrogate and true endpoint is multiplicative. They also show that the PTE is an combination of *three* different quantities—the RE, the AA and a nuisance parameter—and that therefore the PTE is ill defined except in trivial cases. The authors also encourage the use of meta-analysis, noting that it is possible to calculate the measure R^2_{trial} in most settings, but that the choice of an individual-level measure of agreement, R^2_{indiv}, is not universal and it depends on the type of joint model of the surrogate and true endpoint that is used.

Alonso et al. (2006) also echo this point stating that one the of the drawbacks of the meta-analytic methods based within the Belgian paradigm is that 'different settings required different definitions [of R^2_{ind}]'. The authors con-

tinue:

> 'Estimating individual-level surrogacy... has frequently been based in a variance-covariance matrix coming from the distribution of the errors. However, if we move away from the normal distribution it is not always clear how we can quantify the association between both endpoints after adjusting for treatment and trial effects and, as a result, several different parameters have been proposed showing a clear lack of a unified approach.'

(Alonso et al., 2006)

The authors then offer a unified approach and propose the *likelihood reduction factor*, LRF. Consider, for the ith trial, the two generalised linear models:

$$g_T \left\{ E\left(T_{ik}|Z_{ik}\right) \right\} = a_1 + b_1 Z_i$$
$$g_T \left\{ E\left(T_{ik}|Z_{ik}, S_{ik}\right) \right\} = a_2 + b_2 Z_i + c_2 S_i,$$

for some link function g_T. Let G_i^2 denote the log-likelihood ratio test statistic comparing these two models at trial i (minus twice the difference in the log likelihoods). Then:

$$LRF = 1 - \frac{1}{N} \sum_{i=1}^{N} \exp\left(-\frac{G_i^2}{n_i}\right). \tag{2.61}$$

Linking in with the work on information gain by Kent (1983), the authors explain that the LRF can be thought of as a generalisation of R_{ind}^2 based on the information gain in the true endpoint using the surrogate. The authors shows that (*i*) $LRF \in [0,1]$, (*ii*) $LRF = 0$ if S and T are independent in each trial, (*iii*) as $LRF \longrightarrow 1$, there is usually a deterministic bijection between S and T effectively implying perfect surrogacy, and (*iv*) LRF reduces to the maximum likelihood estimator of R_{ind}^2 when S and T are normally distributed. The authors note that LRF may be bounded above by $LRF_{max} < 1$ if T follows a discrete distribution and so propose $LRF_{adj} = LRF/LRF_{max}$ which will always be bounded above by 1. Here LRF_{max} is the best LRF value for the best possible fitted model.

The authors link the LRF back to the first of Prentice's criteria (equation 2.54 on page 38). The LRF is, in some way, quantifying the extent to

which the criterion is met and is therefore more appealing that the simple hypothesis test of whether or not the criterion is satisfied. In this way, the LRF bridges the gap between the Prentice criteria and the Belgian paradigm. Qu and Case (2007) review this work developing the LRF showing that it does reflect the association between T and S adjusted for the treatment linking back with the original single-trial measure the *adjusted association* proposed by Buyse and Molenberghs (1998) that was the basis for R^2_{ind}.

Alonso and Molenberghs (2007) observe that it is not clear what exactly the LRF is estimating how to interpret it in practice. They propose a new quantity, R^2_h, derived from an *information theory* foundation:

$$R^2_h = \frac{\mathrm{EP}(T) - \mathrm{EP}(T|S)}{\mathrm{EP}(T)}, \tag{2.62}$$

where $\mathrm{EP}(T)$ is the *entropy power* of T and $\mathrm{EP}(T|S)$ is the *conditional entropy power* of T given S:

$$\mathrm{EP}(T) = \frac{1}{(2\pi e)^n} \exp\left(2h(T)\right), \tag{2.63}$$

$$\mathrm{EP}(T|S) = \frac{1}{(2\pi e)^n} \exp\left(2h(T|S)\right), \tag{2.64}$$

where $h(T)$ is the *entropy* of T and $h(T|S)$ is the *conditional entropy* of T given S. For T discrete, the entropy and conditional entropy are defined as:

$$H(T) = E_T\left[\log P(T = t)\right], \tag{2.65}$$

$$H(T|S) = E_S\left[E_T\left[\log P(T = t|S = s)\right]\right]. \tag{2.66}$$

For T continuous, the entropy and conditional entropy are defined as:

$$h_d(T) = E_T\left[-\log f_T(t)\right], \tag{2.67}$$

$$h_d(T) = E_S\left[E_T\left[-\log f_{T|S}(t|S = s)\right]\right], \tag{2.68}$$

where T has probability density function, $f_T(t)$, and $T|S = s$ has probability density function $f_{T|S}(t|S = s)$. R^2_h is therefore a measure of the amount of uncertainty in T that is expected to be removed if S is known. R^2_h has the same properties as the LRF described above numbered (*i*) to (*iii*). The authors

extend the definition to a meta-analytic framework:

$$R_h^2 = 1 - \sum_{i=1}^{N} \alpha_i e^{-2I_i(S_i, T_i)},$$ (2.69)

where $I(S, T) = h(T) - h(T|S)$ is the *mutual information* and $\alpha_i > 0 \forall i$ are some scale parameters, $\sum_i \alpha_i = 1$. The exact values that the scale parameters take is undefined and is described as 'the subject of future research'. The authors show that the LRF is a consistent estimator of R_h^2 and therefore 'sharpens the interpretation of the LRF'. The authors go on to give an explanation of R_{indiv}^2 from an information theory perspective and, using *Fano's inequality*, that the quality of a surrogate depends on the power entropy of the true endpoint and implying that 'for some endpoints the search for a good surrogate can be a dead end street unless T and S are extremely closely related'. This application of information theory to surrogate marker evaluation is clearly a very exciting development, but has unfortunately arrived too late for use in this research project. This work has subsequently been summarised in Molenberghs et al. (2008) and Pryseley et al. (2007), but to quote Alonso and Molenberghs (2007): 'we believe more research will be needed to comprehensively map out the whole impact information theory can have on the validation of surrogate markers'. The application of information theory to surrogate marker evaluation will therefore not be considered further in this thesis.

Tilahun et al. (2007) discuss computational issues in evaluating surrogate endpoints across a number of trials using these methods. They note the lack of implementation of these methods in statistical software and include some SAS macros and R functions for fitting these models. Abrahantes et al. (2004) explore the effect of the choice of hierarchical units in evaluating surrogate endpoints across a number of trials that may themselves have data from several centres or several countries.

2.9 Meta-analysis using Trial-Level Summary Estimates in the context of HIV

Hughes et al. (1995) propose two different meta-analyses to investigate a surrogate. Firstly, the authors suggest quantifying the level of surrogacy of the potential surrogate within in each trial, perhaps using the PTE (this paper was

published before the widespread criticism of this measure), and calculating a weighted average across all trials using precision of the estimate as weights. The second method that the authors propose is a graphical approach: they suggest plotting the difference in average surrogate marker levels (the treatment effect on the surrogate, S) against the relative risk of clinical outcome (the treatment effect on the binary true endpoint, T) with a point plotted for each trial with error bars. This method provides a way to visually assess surrogacy by looking for a linear relationship between the treatment effect on the surrogate endpoint and the treatment effect on the true endpoint. The authors are also keen to stress the importance of defining the true endpoint before beginning to identify surrogate endpoints. This is particularly an issue in HIV since the definition of 'HIV-related disease progression' varies considerably. This is also discussed by Hughes (2005).

Daniels and Hughes (1997) build on this second approach by modelling this association across trials and assessing the reliability of such a model to predict a treatment difference on the true endpoint given a treatment difference on the surrogate endpoint. Given $\hat{\theta}_i$ and $\hat{\gamma}_i$ as estimates of the treatment differences on the true and surrogate endpoints respectively from the ith trial, the authors initially suggest the simple linear model

$$\theta_i|\gamma_i \sim N(\alpha + \beta\gamma_i, \tau^2).$$

$\beta = 0$ corresponds to the situation where the treatment difference in the surrogate in no way predicts the treatment effect on the true endpoint, thus invalidating it as a surrogate marker. If $\beta \neq 0$, $\tau^2 = 0$ corresponds to situation where the treatment difference in the true endpoint can be perfectly predicted from the treatment difference in the surrogate endpoint. This model also takes into account treatment effect mediated through a mechanism independent of the surrogate in the parameter α. If $\alpha \neq 0$, then there is a non-zero fraction of the treatment effect on the true endpoint that is not captured by the surrogate. The authors consider the treatment difference on the surrogate as a random effect in this model and use a Bayesian approach to fit the model. They highlight the importance of sensitivity analyses to assess the influence of each trial in the overall analysis. They emphasise the importance of Prentice's second criterion that the treatment has some effect on the surrogate endpoint and of having a large spread in magnitude of θ_i and γ_i leading to a better understanding of the relationship between the two. This point shows a conflict with

the Belgian paradigm where they suggest that a treatment need not have a real effect on the surrogate and that 'fluctuations around zero in individual trials can be very strongly predictive of the effect on the true endpoint' (Buyse et al., 2000a). In conclusion, Daniels and Hughes (1997) state that their proposed meta-analysis is relatively simple to apply and that it is particularly useful in situations where individual studies lack power to provide strong conclusions on their own.

Hughes et al. (1998) look to evaluate initial changes in CD4 count as a surrogate marker for progression to AIDS and death using the two meta-analysis methods proposed by Hughes et al. (1995), including data from 15 clinical trials involving 24 treatment comparisons. They first calculate a weighted average of the PTE. Whilst the confidence interval for this average is reasonably narrow, the authors observe that the estimates of the PTE were outside the range (0,1) and that some of the confidence intervals were very wide, in some cases completely containing the range $(-2,2)$, showing that some of these estimates have very poor precision. Secondly, the authors plot log hazard ratio for progression to AIDS or death for the treatment effect against difference in mean change in CD4 count for the treatment effect with a point plotted for each treatment comparison. They identified five comparisons that fell in the upper right-hand quadrant or lower left-hand quadrant and these indicated opposite effects on the surrogate and the true endpoint. They found that, for all five of these, there was no significant difference between the treatments in the hazard of clinical progression and so concluded that the lack of concordance merely reflected chance variation. The authors also fitted linear regression models to these data calculating predictions and 95% prediction intervals for the log hazard ratio of outcome given a particular difference in mean change in CD4 count.

The authors observe that the first approach is not likely to be valuable in individual studies due to the imprecision of the PTE within studies and favour the second approach noting that it allows clinical trials of similar sizes and similar numbers of events equal emphasis regardless of the size of the treatment effect whereas the first method tends to give greater weight to studies in which there is the greatest statistical significance for the treatment difference on the true endpoint. The authors suggested that the small PTE implied that there were other mechanisms of drug action that were not captured by CD4 count and concluded that CD4 cell count changes over 6 months are only a weak partial surrogate for clinical progression over 2 years. Another meta-

analysis of HIV trials (HIV Surrogate Marker Collaborative Group, 2000) also used this graphical method to assess surrogacy.

2.10 Discussion

The different approaches to surrogate marker evaluation are varied. Prentice (1989) provided a clear definition of a surrogate, and proposed three criteria that any candidate surrogate must satisfy. Freedman et al. (1992) proposed the PTE as a means of quantifying the proportion of treatment effect on the true endpoint explained by the surrogate—effectively measuring the extent to which the Prentice criteria are satisfied. This generated a lot of research activity and a large number of similar measures were proposed with varying success.

Buyse and Molenberghs (1998) and Buyse et al. (2000a) introduce two important paradigm shifts. These authors firstly propose the distinction between individual-level surrogacy and trial-level surrogacy, and secondly develop methods for moving the task of evaluation from a single trial to a meta-analytic approach. It had been previously stated by many authors that a surrogate marker must be evaluated across a number of treatment comparisons and a number of trials, but Buyse and Molenberghs were among the first to develop reliable methods for doing this. These methods were straight-forward for Gaussian true and surrogate endpoints, the situation that they concentrated on, but not for other types of endpoints. Other authors have since proposed a number of extensions for other types of endpoints that tend to involved more complex methods. Daniels and Hughes (1997) propose a different meta-analytic approach that allows greater flexibility, accommodating a many different types for the surrogate and true endpoints.

The evaluation of surrogate markers in this thesis will focus on multi-trial, meta-analytic methods. Single-trial methods will also be used, but the results from the meta-analytic approach of a number of trials will give a more reliable answer.

Chapter 3

Prognostic and Surrogate Markers for Poor Outcome after Treatment for Tuberculosis

3.1 Introduction

In Chapter 2, the statistical literature on the evaluation of prognostic and surrogate markers was reviewed and summarised. This chapter focuses on prognostic and surrogate markers for poor outcome after treatment for TB, identifying and reviewing those that have been suggested in the literature.

Before looking at the literature on predictors for response after treatment of tuberculosis, it is important to understand something of the nature and progression of TB disease. As noted in Chapter 2, before any discussion of markers, it is also necessary to examine and define a clear clinical endpoint for which a candidate surrogate will be a substitute. TB is unique among infectious diseases in that failure to culture the causative agent, *Mycobacterium tuberculosis*, in a specimen (a sputum sample for pulmonary TB) is not a sufficient condition for cure. It is therefore important to briefly look at the diagnosis of tuberculosis (section 3.2), the action of anti-tuberculosis drugs (section 3.3), and the clinical endpoint (section 3.4) before looking in more detail at

the published work on prognostic and surrogate markers for poor outcome to treatment (section 3.5).

Since sections 3.2 to 3.4 are general sections setting the scene, no systematic review was performed. They do not constitute exhaustive reviews of each topic, they include discussion of the important issues and reference the important papers relevant to the aim of this chapter, leading into section 3.5.

3.1.1 Latent Tuberculosis

It is estimated that one third of the world's population is infected with the *Mycobacterium tuberculosis*, but less than one percent (14.4 million people) of these were active cases (World Health Organization, 2008).

All persons who have inactive or latent infection are at continued risk for activation of the disease (Global Alliance for TB Drug Development, 2001)—an estimated 10% will go to develop active disease (Kato-Maeda and Small, 2000), but persons who have recently been infected or who have certain clinical conditions (such as HIV infection) are at increased risk of progression to active disease (Blumberg et al., 2005). While all infectious diseases involve a period of incubation or latency after exposure, tuberculosis is unlike most in that the delay between infection and disease is extremely varied and can range from a few weeks to a lifetime (Fine and Small, 1999). It is therefore important that diagnostic procedures distinguish between active and latent tuberculosis. In this thesis, discussion of tuberculosis disease will always refer to active disease unless otherwise stated.

3.1.2 Extra-pulmonary Tuberculosis

The most common form of tuberculosis is pulmonary tuberculosis (PTB), but *Mycobacterium tuberculosis* can also affect other parts of the body. Other forms include tuberculosis meningitis (tuberculosis of the central nervous system), tuberculosis of the lymph nodes, spinal tuberculosis and miliary tuberculosis (disseminated tuberculosis of the circulatory system). Chemotherapy is directed at the bacteria rather than the host and should therefore be effective against any form of the disease (Youmans, 1979). In practice, this is not necessarily the case and different variants of *extra-pulmonary tuberculosis* are commonly treated differently (Fox et al., 1999). Extra-pulmonary disease is therefore usually an exclusion criteria for PTB trials. This thesis focuses on

PTB and discussion of TB will refer to PTB unless otherwise stated. Cases of extra-pulmonary TB will be excluded from analyses where diagnosed.

3.2 Diagnosis of Tuberculosis

One of the five elements of the WHO-recommended DOTS strategy (see Chapter 1) relates to case detection emphasising the importance of good diagnostics for TB disease. While it is true that a *diagnostic marker* serves a different purpose that a *prognostic marker*, markers for diagnosis can also sometimes be used for prognosis. This is generally the case in TB, where diagnosis usually involves assessing whether *M. tuberculosis* bacilli is present in the lungs, and assessing prognosis involves monitoring the decline of bacilli in the lungs. Those markers which are used for diagnosis therefore form the pool from which to identify prognostic markers and ultimately surrogate markers.

3.2.1 Radiography

In the first few decades of the twentieth century, the most common method for the diagnosis of tuberculosis was with fluoroscopy: 'a dangerous procedure in which the patient stood so that an X-ray image of his chest appeared on a fluorescent screen without a film' (Mitchison, 2005). Images on film were introduced later allowing for the possibility of showing individual slices of cavities or other lesions.

It is common for previously unsuspected pulmonary tuberculosis to be detected on the inspection of a chest X-ray (Youmans, 1979) although many lung disorders may appear as pulmonary tuberculosis in a chest X-ray. Newer tomographic techniques to show sequential slices of cavities such as computerised axial or nuclear magnetic resonance tomography are particularly valuable in distinguishing between tuberculosis and, for example, cancerous lesions (Mitchison, 2005). Despite this, it can be difficult to distinguish between lesions of tuberculosis that are no longer active (indicating previous active infection that has become inactive) and active tuberculosis cavities. Any radiographic finding can therefore only be provisional.

One study found that, among the 17 study patients with culture-positive pulmonary tuberculosis and verified AIDS, only one patient had radiographic evidence typical of adult onset tuberculosis (Pitchenik and Rubinson, 1985). This suggests manifestation of tuberculosis differs between patients who are

HIV positive and those who do not, adding further complications to the X-ray image as a diagnostic tool.

The X-ray is commonly used for active case finding (Schluger and Rom, 1994), identifying those with undiagnosed suspected tuberculosis who can then be sent to give a sputum sample for a confirmatory smear or culture test. University College London Hospital has a mobile X-ray clinic mounted on the back of a lorry that is used to visit at-risk groups (such as those in homeless shelters and prisons) and diagnose TB[1]. In a current cluster-randomised trial in South Africa, evaluating the impact of isoniazid prevention therapy on TB incidence, the X-ray is used as the first tool for identifying active TB[2].

3.2.2 The Guinea-Pig Test

Guinea-pig inoculation was the established diagnostic tool when culture methods were inefficient (Marks, 1972). Tests involving animals are costly and undesirable and therefore, with improved culture methods, are no longer recommended on ethical grounds. The guinea-pig test is highly sensitive and there could be case for its use when an invasive procedure is used on a patient to obtain only scanty material in extra-pulmonary TB(Morris and Barton, 1983).

3.2.3 Tuberculin Skin Test

The tuberculin skin test (*TST*, also known as the *tuberculin sensitivity test* or *Purified Protein Derivative (PPD) Test*) is another diagnostic tool that is widely used in developed countries including the United States of America and the United Kingdom. It does not distinguish between active and inactive infection and therefore predominantly is used to diagnose latent disease and primary infection (Shingadia and Novelli, 2003). A positive TST result indicates that the subject has been exposed to TB at some point in their life, but has not necessarily ever had active disease. It is known to have low sensitivity in immunocompromised patients, requires patients to return 72 hours after the test to have the result read and it is thought that the results are affected by the BCG vaccine (Frieden et al., 2003).

---

[1]UCLH Mobile x-ray unit. http://www.uclh.nhs.uk/GPs+healthcare+professionals /Clinical+services/Infectious+diseases+(Hospital+for+ Tropical+Diseases)/Infectious+diseases+-+Mobile+x-ray+unit. Retrieved 23 Apr 2009.

[2]Thibela TB: CREATE. http://www.tbhiv-create.org/about/studies/thibela. Retrieved 23 Apr 2009. Personal communication from Katherine Fielding, Co-investigator.

3.2.4 Sputum Microscopy

The most widely used method of diagnosing tuberculosis in developing countries is by establishing the presence of acid-fast bacilli (AFB) in sputum using microscopic examination after staining. The sample is stained and examined under a microscope for the presence of M. *tuberculosis* bacilli. The number of visible bacilli can be counted and the *smear* graded on a scale from negative to heavily positive.

This is a comparatively quick way to diagnose tuberculosis, does not require expensive equipment and can therefore be carried out in most clinics across the world. The original procedure of direct field microscopy was replaced by fluorescence microscopy because it is five times more rapid with greater specificity (Mitchison, 2005) (and it is recommended today that the replacement of fluorescence bulbs with light-emitting diodes improves longevity at reduced cost (Khan et al., 2007)); but apart from this one improvement, the technology has remained largely unchanged since its introduction in the 1880s (Perkins et al., 2006). Smear sensitivity can depend greatly on the quality of the sputum and the skill of the laboratory technician carrying out the test. A randomised controlled trial in Pakistan showed women (but not men) were more likely to test smear positive for TB if they were given 'sputum-submission guidance' in addition to prevailing practice (Khan et al., 2007). While not particularly sensitive, direct smears have been found to be very specific for diagnosing tuberculosis (Mitchison et al., 1975). In studies, smear sensitivity for M. *tuberculosis* is commonly found to be around 50%, but specificity upwards of 95% (Schluger and Rom, 1994). A key problem with the smear test is that it does not distinguish between viable and non-viable bacilli (Youmans, 1979), although this is more of a problem for assessing disease progression than diagnosing TB. Smear examination cannot distinguish between the M. tuberculosis complex and other mycobacterium (Lee et al., 2003) and is also less sensitive in immunocompromised patients (Raviglione et al., 1992) due to their typically lower bacillary burden (Alisjahbana and van Crevel, 2007). Despite a low sensitivity, it is an attractive tool for public health programmes since it is sufficiently specific, gives a response in a short period of time and requires only one piece of equipment (Perkins et al., 2006).

3.2.4.1 Smear-Negative Pulmonary Tuberculosis

The low sensitivity of sputum microscopy leads to the phenomenon of *smear-negative pulmonary tuberculosis*. Patients who are smear-negative and culture-positive are less infectious (although studies have shown that smear-negative patients should not be considered non-infectious (Behr et al., 1999)), have a lower bacillary load and are recommended to be treated with the same regimens as patients who are smear-positive (World Health Organization, 2003). Smear negativity is more common in patients who are HIV positive (Long et al., 1991; Behr et al., 1999).

3.2.5 Culture Examination

Culture are more sensitive than smears and are of a comparable specificity, but provide results more slowly (Mitchison, 2005). If *M. tuberculosis* is present in the sputum, colonies will form on the medium. Only active, replicating bacilli will grow and methods exist to determine whether or not the growth is indeed *M. tuberculosis*. The culture result will tend to provide more information than the smear test as the actual number of colony forming units (CFUs) can be counted.

Culture growth on solid media is therefore the recognised diagnostic test for the presence of active tuberculosis. One positive culture for *M. tuberculosis* is required to determine a definite case of tuberculosis under WHO definitions of tuberculosis cases and treatment outcomes (World Health Organization, 2008) but, in countries where culture is not routinely available, one sputum smear positive for AFB is insufficient, two are required for a definite case.

Well equipped labs are required for culture growth and examination and, since the doubling time of the bacilli *M. tuberculosis* is long at approximately 20 hours (Youmans, 1979), results may not be available until 2 months after the sample was taken (Fortún et al., 2007).

It is common for patients to be recruited into a clinical trial on the basis of their smear test results due to the speed of the test, but their response to treatment would be assessed on the basis of sputum culture results (e.g., East African/British Medical Research Council, 1978a; Jindani et al., 2004).

3.2.5.1 Automated Liquid Culture Systems

First described in 1977 (Middlebrook et al., 1977), automated culture systems provide a less labour intensive and higher capacity diagnostic alternative. One of the first liquid culture systems was the BACTEC 460 (Becton Dickinson, Sparks, MD, USA). This was not used widely due to radioactive elements and other problems including the hazardous procedure of handling syringes and the high cost of the instrument (Hasegawa et al., 2002). In the past decade, the BACTEC MGIT 960 system (Becton Dickinson, Sparks, MD, USA) has been introduced. This system is non-invasive, non-radiometric, does not involve sharps and is fully automated. Both of these systems have been shown to have sensitivities comparable to solid culture methods while producing results much faster (Hanna et al., 1999).

The mycobacterial growth indicator tube (MGIT) is a recent diagnostic system for the rapid detection of *M. tuberculosis* in sputum. Processed sputum is inoculated into a liquid culture medium in the MGIT tube. A light at the bottom of the tube is activated if microbial growth is detected. The degree of positivity of a sample is measured in the time it takes the sensor to be activated, the *time to detection* (TTD). It has been shown that MGIT can detect the presence of *M. tuberculosis* in as short a time as two days (Epstein et al., 1998).

Some have suggested that solid culture methods are no longer necessary (Sharp et al., 2000), though higher contamination rates have been found in automated liquid culture systems than in solid culture methods (Hasegawa et al., 2002), up to 20% for the MGIT system compared to the generally accepted rate of 2%-5% on Lowenstein-Jensen solid medium (Diraa et al., 2003). While the MGIT system is cheaper than the earlier BACTEC 460, the cost is still estimated to be US $7 compared to US $0.10 per AFB stain (Diraa et al., 2003). Whilst the MGIT is highly accurate in the detection of mycobacteria, an individual TTD value depends on the size of the innoculum and the quality of the sputum in addition to the bacillary burden of the patient (Epstein et al., 1998).

Other automated liquid culture systems include the MB/Bact system (Organon Teknika, Boxtel, The Netherlands) and MB-Check (Nippon Roche Ltd., Tokyo, Japan).

3.2.6 Novel Diagnostic Markers

There are a large number of novel diagnostic tests that have been introduced in the last few decades. These included interferon-gamma assays such as the ELISPOT (Lawn et al., 2007), genetic markers (Cooke et al., 2008; Cliff et al., 2004), culture-based assays such as MODS (Moore et al., 2006) and novel diagnostic devices such as the Lung Flute®[3].

In 1993, backed by the WHO and with initial funding from the Bill and Melinda Gates Foundation, the independent non-profit Foundation for Innovative New Diagnostics (FIND) was established with the mission to 'development of rapid, accurate and affordable diagnostic tests for poverty-related diseases in the developing world'[4]. Their initial focus was on tuberculosis only and have since supported the search for new diagnostic methods for use in both the community health clinic and the technologically-equipped central laboratory.

3.3 Drug Action during Treatment

The main chemotherapy regimen for tuberculosis recommended by the WHO is a six month regimen consisting of an *intensive phase* of two months of isoniazid (H), pyrazinamide (Z), rifampicin (R) and ethambutol (E) taken daily followed by a *continuation phase* of four months of isoniazid and rifampicin also taken daily (denoted HRZE/HR) (World Health Organization, 2003). An eight month regimen starting with the same initial phase of HRZE but with a continuation phase of six months of ethambutol and isoniazid has been recommended as an alternative to the six month regimen and is still used in 31 countries worldwide (13 have plans to change to the six month regimen, World Health Organization (2008)) despite a recent randomised controlled trial having established that this eight month regimen is inferior to the standard six month regimen (Jindani et al., 2004). Treatment with a six or eight month regimen is called *short-course chemotherapy* in contrast to the earlier treatment plans of eighteen months or more (e.g., East African/British Medical Research Council, 1978a).

---

[3]Medical Acoustics introduces the Lung Flute®. http://www.medicalacoustics.com /Home/LungFlute. Retrieved 23 Apr 2009.

[4]FIND Diagnostics - Mission, Vision and Objectives. http://www.finddiagnostics.org /about/mission_vision.shtml. Retrieved 23 Apr 2009.

It is postulated that there are three phases of drug action (Mitchison, 1997) corresponding to as many as four different bacterial populations in man (Mitchison, 1985). The majority of tubercle bacilli are rapidly growing (population 1) and are killed during the first few days of treatment. This period of drug action is known as the *Early Bactericidal Activity* (EBA). Isoniazid has the most potent EBA (Jindani et al., 2003).

Remaining bacilli, termed *persisters*, are almost dormant, hardly metabolising and are therefore killed more slowly. It is hypothesised that these persisters fall into two populations: those in an intracellular acid environment killed most effectively by pyrazinamide (population 2) and those that have occasional spurts of active metabolism in an extracellular neutral or alkaline environment which are killed most effectively by rifampicin (population 3). It is bacterial factors that determine the speed of killing rather the drug action as it is only when a dormant bacilli begins to metabolise that it is killed by a drug (Jindani et al., 2003). After the first few days of EBA, this period of drug action until around the second month of treatment is called *bactericidal activity*. The speed of killing gets progressively slower during the six months of treatment and it is the ability of a drug to kill or sterilise the last few viable bacilli that is termed the *sterilizing activity*. Rifampicin and isoniazid are the most important drugs in the continuation phase beyond two months. There is a fourth population of bacilli that remain completely dormant and is unaffected by any drug.

Within the first two months, the cavities will close (Mitchison, 1996) and it is common for the majority of patients to become culture negative (see section 3.5.3).

3.3.1 Drug Resistance

Within a population of bacilli in an infected lesion, there may be mutant bacilli that are resistant to one or more anti-tuberculosis drugs. If such patients are treated with one of these drugs, their lesions will soon contain only resistant organisms and treatment will fail. It was demonstrated in 1948 that the emergence of resistance to either streptomycin or para-aminosalicyclic acid (PAS) was greatly decreased when both drugs were given in a combined treatment (Medical Research Council, 1950). The standard treatment regimen for tuberculosis consists of an intensive phase of four drugs given in the first two months. If a patient is initially resistant, or develops resistance to a particular

70

drug, then this phase ensures that there are still three others drugs working against the disease.

Despite the use of multi-drug combinations, an inadequate or poorly administered regimen will allow a drug-resistant strain to become dominant in the patient. Transmission of drug-resistant disease in a population is also a significant source of new drug-resistant cases (World Health Organization, 2006). Drug-resistance often arises in areas of poorly organised or poorly funded TB control programmes.

Mitchison and Nunn (1986) looked at the influence of initial drug resistance on treatment response, reviewing controlled trials conducted in East Africa and South-East Asia in collaboration with the British Medical Research Council (BMRC). The authors note that the response of patients with initial drug resistance dramatically improved with the introduction of the short-course regimens of six months including four drugs. They show that rifampicin was particularly responsible for the improvement in the response in initial isoniazid resistance and state that, in general, 'as the number of drugs in the regimen was increased, the failure rate fell' (Mitchison and Nunn, 1986). Giving only two drugs to a patient who was resistant to one of the drugs was effectively mono-therapy, but giving three or four drugs meant that there were still at least two to which the bacilli were susceptible to. The authors also highlight the high success rate of short-course regimens in the presence of initial resistance to isoniazid and streptomycin. Initial rifampicin resistance was rare, but resulted in a poor patient outcome and the authors prophetically comment: 'if rifampicin resistance became widespread, it would threaten the success of modern short-course treatment of tuberculosis' (Mitchison and Nunn, 1986).

Mitchison et al. (2007) notes that 'relapse cultures almost always have susceptibility patterns identical to those before treatment and respond to further treatment with the original regimen'. There is some evidence that HIV positive patients with low CD4 counts that do relapse, may be more likely have acquired rifampicin resistance in their relapse culture.

3.3.1.1 Multi-Drug Resistant Tuberculosis

Multi-drug resistant tuberculosis (MDR-TB) is defined as disease that is resistant to the two most important anti-tuberculosis drugs: isoniazid and rifampicin. The standard regimen for treating TB is then reduced to only pyraz-

inamide and ethambutol and therefore *second-line drugs* are required. Second-
line drugs for treating MDR-TB are less effective than first-line drugs and
therefore must be taken for longer (usually at least 18 months), have more
associated toxicities and are more expensive. Treatment for MDR-TB needs
to be tailored to the resistance patterns of individuals patients and very few
clinical trials have been conducted to evaluate regimens for the treatment of
MDR-TB (Caminero, 2006).

2005 saw the identification of *eXtensively-Drug Resistant TB (XDR-TB)* in
South Africa. XDR-TB is defined as MDR-TB with additional resistance to
one fluoroquinolone and one second-line injectable (capreomycin, kanamycin
or amikacin)[5]. There is very little treatment for XDR-TB and the prognosis is
very poor (Raviglione, 2006).

Since MDR-TB and XDR-TB require different treatment to drug-susceptible
TB and are considerably more life-threatening, it is common to consider them
almost as separate diseases.

3.4 Endpoint of Clinical Trials in Tuberculosis

The purpose of a surrogate marker is to substitute for the clinical endpoint
that is usually used in a clinical trial. It is important therefore to have an
objective and clearly defined clinical endpoint that is widely accepted to reflect
real patient benefit and treatment response. In this section, the endpoints
commonly used in clinical trials for anti-tuberculosis regimens are reviewed.

Among those who are culture negative at the end of treatment, some will
have recurrence of disease and with positive cultures some time after treat-
ment has ended. The World Health Organisation (WHO) defines a *relapse
case* as a 'patient previously declared cured but with a new episode of bac-
teriologically positive (sputum smear or culture) tuberculosis' (World Health
Organization, 2008). It is the continuation phase of short-course chemother-
apy that is vital to kill all or most of the persisting bacilli and therefore prevent
relapse.

Patients who do not respond to chemotherapy will remain culture positive
throughout the treatment period. In a clinical trial setting, *treatment failures*
at the end of treatment are rare, and therefore the primary endpoint used

---

[5]World Health Organisation Press Release (2006): WHO Global Task Force out-
lines measures to combat XDR-TB worldwide. http://www.who.int/mediacentre
/news/notes/2006/np29/en/index.html. Retrieved 23 Apr 2009.

in a clinical trial is sometimes relapse within 12-24 months after the end of treatment (O'Brien, 2002) and sometimes a combined endpoint of *poor outcome* defined as failure at the end of treatment or relapse during 12-24 months of follow-up (e.g. Jindani et al., 2004).

This period of extensive follow-up is costly and can mean that a TB trial could take five or more years to complete. Relapse rates under clinical trial conditions are often less than 5% and therefore large numbers of patients are required in a trial, even to show non-inferiority (Nunn et al., 2008). A properly validated surrogate marker for the sterilizing activity for anti-tuberculosis drugs would result in shorter trials with smaller sample sizes ultimately speeding the drug development process considerably (Global Alliance for TB Drug Development, 2001)

3.4.1 Reinfection and Relapse

In patients who appear to be cured at the end of treatment, a small proportion will have recurrent disease requiring retreatment. With the help of DNA fingerprinting of *M. tuberculosis*, recurrences can be separated into *exogenous reinfections* (the recurrent bacilli are of a different strain to the bacilli that the patient was first infected with) and *true relapses* or *endogenous reactivations* (the recurrent bacilli are of the same strain as the bacilli that the patient was first infected with) (Verver et al., 2005). Historically, many publications and medical textbooks stated that the majority of disease recurrence were relapses although there has been much debate more recently (van Rie et al., 1999). An editorial in the New England Journal of Medicine makes the following point:

> If exogenous reinfection, which is clinically indistinguishable from relapse, is common, then new regimens that effectively eliminate infection or treat disease will be unfairly judged in clinical trials

> Fine and Small (1999)

Estimates of the proportion of recurrences that are true relapses vary greatly depending on the population being studied. Table 3.1 shows the different terms that are used for recurrences or relapses and how these terms are used in this thesis. Older publications use *relapse* to refer to *recurrences* (before it was possible to distinguish between reinfections and reactivations), but newer

publications often use *relapse* to refer to reactivations in contrast to reinfections.

Verver et al. (2005) studied TB patients in the high-TB incidence area of Cape Town, South Africa to estimate the rate of recurrent TB attributable to reinfection after successful treatment. They followed 897 patients treated for TB, of whom 612 had a DNA fingerprint available at enrolment. Recurrent disease occurred in 108 (18%) of these of whom 68 had a DNA fingerprint in the second episode. The authors find that only 7 of the 31 (23%) recurrent patients who completed treatment were true relapses and 33 of the 37 (92%) recurrent patients who defaulted were true relapses.

Term	Definition
Recurrence	Any repeat episode of TB after a patient has been declared cured.
Relapse	A recurrence that has been identified as a reactivation.
(Exogenous) Reinfection	A recurrence of a different strain to that which the patient was originally infected with.
(Endogenous) Reactivation	A recurrence of the same strain as that which the patient was originally infected with.

Table 3.1: Terms used for recurrence of disease and their definitions as used in this thesis.

Glynn et al. (2004) fingerprinted cultures from patients who had been diagnosed with TB in Malawi from 1996 to 2001. The authors found 7 (58%) of 12 recurrences in HIV-positive patients were exogenous reinfections whereas all 8 recurrences in HIV-negative patients were true relapses.

Jasmer et al. (2004) looked at recurrences in two prospective clinical trials conducted in Canada and the US. They found that only 3 (4%) of the 75 recurrences with DNA genotyping results available did not match the pretreatment strain and were therefore thought to be exogenous reinfections. They conclude that 'recurrent tuberculosis in the United States and Canada, countries with low rates of tuberculosis, is rarely due to reinfection with a new strain of *M. tuberculosis*'.

Chiang and Riley (2005) performed a review of literature discussing exogenous reinfection. The reviewers found considerable experimental and epidemiological evidence in support of the qualitative role of exogenous reinfection as a cause of tuberculosis, but noted the lack of good quantitative estimates of its contribution.

A systematic review (Lambert et al., 2003) looked at thirteen studies using DNA fingerprinting to identify the occurrence of reinfection amongst recurrent disease. They found that among recurrences, reinfections did occur in both low-incidence and high-incidence countries among both HIV-positive and HIV-negative patients. Some studies found reinfection to be rare and some found reinfection to account for almost all of the recurrences. The authors suggest that a randomised controlled trial will still be able to compare the efficacy of different treatment regimens provided that the incidence of reinfection is likely to be similar in each arm of the trial. They also point out that true relapse rates will be overestimated in both arms leading to reduced power.

Before the introduction of DNA typing necessary for distinguishing between reinfection and reactivation, a recurrence was generally referred to as a relapse in any published literature. This is the case with all of the study reports for the MRC clinical trials published in the 1970s and 1980s. Because of this confusion, the term *recurrence* will be used in this thesis in place of the term *relapse*, unless DNA typing has been done to separate out exogenous reinfections from endogenous reactivations.

3.5   Prognostic and Surrogate Markers in TB

There has been only limited discussion of surrogate markers for response to treatment in TB; much of the discussion has instead focused around 'predictors', 'risk factors' or 'correlates'. There has been much written about the analysis and evaluation of surrogate markers in statistical journals (see Chapter 2) and this theory has filtered through to many diseases areas. A good example is HIV where several studies and meta-analyses have been published looking for surrogate markers for death and progression to AIDS (e.g., HIV Surrogate Marker Collaborative Group, 2000; Hughes et al., 1998). This has not been the case in the disease area of tuberculosis—there has been very little discussion of formal evaluation of surrogate markers, those authors choosing

to use the term 'surrogate' do so with a variety of different meanings. Using the definition from Prentice (1989), a surrogate marker should fully capture the treatment effect on the final endpoint and its purpose is as a substitute for that final endpoint in a clinical trial.

The papers considering predictors or prognostic markers do so with many different objectives: to understand the action of the drugs in question or to provide a cheap method to evaluate new therapies (Mitchison, 1993; O'Brien and Nunn, 2001), to identify patients who are more likely to fail (Barnes et al., 1988), to distinguish between active and latent TB (Jacobsen et al., 2007), to tailor treatment plans to individual patients (Wallis et al., 1998) or to predict MDR-TB (Salomon et al., 1995). There has been an acceptance recently for the need for a surrogate endpoint that can substitute for the final endpoint, but this call has really only been made in the last decade (Global Alliance for TB Drug Development, 2001; Burman, 2003). This divergence in the objective and function of the prognostic marker leads to a large variety of approaches to evaluating such markers. Some authors look for baseline factors as predictors of poor outcome, whilst some look only for markers measured after the start of treatment. Due to the distinction between failure at the end of treatment and long-term failure (defined as recurrence after the end of treatment), some authors look at predictors for short-term outcome and some look at predictors for long-term outcome.

Most trials use recurrence or recurrence combined with treatment failure as the endpoint in a clinical trial. The exact definition of what constitutes a bacteriological recurrence (it has been shown that isolated positive cultures are not necessarily indicative of recurrence, see section 3.5.3.2) or a treatment failure varies considerably between trials (e.g. Jindani et al., 2004; Benator et al., 2002). Other trials assess poor outcome using clinical symptoms (Epstein et al., 1998), respiratory failure or death (Barnes et al., 1988).

All this means that the literature on markers and predictors for response to treatment of tuberculosis is mixed and very varied. The following sections contain a review of the literature with the published research grouped by type of marker. A very large number of baseline risk factors for poor outcome have been proposed, and section 3.5.1 surveys those that are most important. Section 3.5.2 looks at some of the evidence for monthly sputum smears and 3.5.3 for monthly sputum culture results as predictors of recurrence. Section 3.5.4 looks at Early Bactericidal Activity, section 3.5.5 looks at Serial Sputum Colony Counts and section 3.5.6 looks at other miscellaneous markers.

Having read widely and discussed with experts in the field, it is clear which are the important papers providing evidence for different prognostic and surrogate markers for poor outcome. Except in the important sections on monthly culture results and serial sputum colony counts, the reviews are not systematic, but include the most important papers that have appeared in recent years. Section 3.5.2, in particular, follows an ongoing discussion in the International Journal of TB and Lung Disease (IJTLD) following recommendations by the International Union against TB and Lung Disease (IUATLD) regarding the two month sputum smear.

3.5.1 Baseline Risk Factors

Many studies have looked at baseline risk factors for treatment failure or recurrence. By its very nature, any measurement taken at baseline cannot be a surrogate marker as it is measured before the start of treatment. Baseline risk factors can be used to identify patients at risk of poor outcome and may be used for the purpose of prognosis.

The two most significant baseline risk factors for poor outcome to treatment are multi-drug resistance (Wallis et al., 1999) (although the presence of mono-drug resistance only results in slightly worse outcomes (Mitchison and Nunn, 1986)) and HIV co-infection (Nunn et al., 1991; Chang et al., 2004). A number of other factors have been found to be associated with treatment outcome in different settings.

Aber and Nunn (1978) was the first paper to look at factors or markers that could predict recurrence following treatment. Using data from two trials undertaken in East Africa and one in Hong Kong, the authors looked at the 'prognostic importance' of several pretreatment factors including age, sex, weight and viable colony count, radiographic severity of disease, radiographic extent of cavitation, baseline smear and culture results and baseline viable count of colony forming units. In the East African studies, the extent of cavitation was most correlated with recurrence and in the Hong Kong studies, patient age and viable CFU count at baseline were of prognostic significance.

A study in South India found that initial drug resistance, smoking and alcoholism were all associated with a higher risk of recurrence (Thomas et al., 2005). In one randomised controlled trial in the US and Canada enrolling 1004 participants, it was found that ethnicity, baseline weight, cavitation on chest radiograph and bilateral pulmonary involvement were risk factors for

failure/recurrence in a time-to-event analysis (Benator et al., 2002). Another trial in Poland found similar risk factors for recurrence: extensive disease, large cavities, heavy growth on pretreatment cultures and heavy use of alcohol (Zierski et al., 1980). A retrospective study using a community survey in South India found previous treatment, being male, alcoholism and being underweight were associated with default from treatment and patients with MDR-TB were more likely to fail treatment (Santha et al., 2002). In patients with HIV in Zambia and Malawi, Ciglenecki et al. (2007) found more advanced HIV disease was associated with increased TB-related mortality. In a systematic review, Slama et al. (2007) found a strong association between smoking and TB disease and limited association between smoking and TB retreatment or TB mortality. Baseline risk factors found by other studies include: cavitation on baseline radiograph (Chang et al., 2004), sex and age (Tam et al., 2002), and alcoholism (Barnes et al., 1988).

3.5.2 Monthly Smear Results during Treatment

Both the World Health Organisation (WHO) and the International Union against TB (IUATLD) recommend extending the intensive phase of two months of treatment by an additional month if a patient has a positive smear at two months. It is largely for this reason that smears are routinely taken at the end of the intensive phase and is based on scattered evidence.

To assess the importance of this recommendation a retrospective analysis of 726 new smear positive TB patients in one province in China was conducted. The authors show that those still smear positive at two months were more likely to fail treatment, although culture confirmation of treatment failure was not done and patients were not followed up for long-term treatment outcome (Zhao et al., 1997). Wilkinson (1998) reviews this article, incorporating results from an earlier publication by this author showing similar results (Wilkinson et al., 1998). The author notes that smears at 2 or 3 months have 'poor to moderate sensitivity (36-61%) and high specificity (88-90%)' and state that the positive predictive value is very low and therefore 'a positive smear at 2-3 months is *not* at all predictive of treatment failure' suggesting that it is no longer useful (Wilkinson, 1998).

Trébucq and Rieder (1998) express concern over this suggestion to throw 'overboard' this 'excellent management tool' of the two month smear. The authors note that the conclusions were based on retrospective cohort studies

rather than clinical trials, but agree that 'it is not a perfect tool for identifying those who will fail or relapse' but 'it is not such a bad tool either' and smears should still be taken at the end of the intensive phase as it provides 'very valuable information for evaluating the [national tuberculosis] programme as a whole'. Barker and Millard (1999) respond by observing that there is really no good evidence from clinical trials linking the two month smear and treatment outcome, the only data available is from epidemiological cohort studies.

In a retrospective case-control study in Malawi of 119 cases of TB patients smear positive after two months of treatment and 237 matched controls smear negative after two months, the cure rate was higher in the controls (77%) than the cases (71%) (Salaniponi et al., 1999). Following 297 smear positive patients in Madagascar (Ramarokoto et al., 2002), it was found that the majority of recurrences and failures were observed in the group that were smear-positive at two months (Of 152 patients smear-positive at two months, 16 (11%) were recurrences or failures and of 145 patients smear-negative at two months, there was only 1 (1%) failure).

It is clear that the smear result after two months of treatment is associated with long-term outcome to treatment. However, this association is weak and cannot reasonably be used as a prognostic marker as the two month smear is insufficiently sensitive for long-term outcome. It may be true that the two month smear is strongly associated with treatment failure as assessed with sputum smears (and it has been shown to 'strongly predict bacteriologic [smear] results beyond three months of treatment' (Rieder, 1996)), and fundamentally 'old habits die hard, but evidence is accruing that smears at 2-3 months do not serve their intended purpose' (Wilkinson, 1998).

3.5.3 Monthly Culture Results during Treatment

Aber and Nunn (1978) was the first paper to look at factors or markers that could predict recurrence following treatment. In the two East African studies, it was found that the culture results at 3 months and at 2 months showed the strongest association with recurrence. In the Hong Kong study, the extent of cavitation at the end of treatment was most highly prognostic of recurrence with the 2 month smear result also significantly associated with recurrence. The authors conclude by pointing out that, despite the differences between the East African and Hong Kong studies, the significant prognostic factors

79

were all ultimately related to 'the bacterial content of the lesions' and that the factors that were common to all three studies were those that 'measured the speed of sputum conversion during treatment'. These authors only looked at prognostic markers for recurrence and did not consider the question of surrogacy.

In a review article (Mitchison (1996), based on an earlier letter, Mitchison (1993)), the author looks at the correlation between proportions of patients still culture positive at each of 1, 2 and 3 months after the start of treatment with recurrence rates across eight published multi-armed trials. He concludes that there was 'good correlation between culture results and relapse rates', noting that the strongest correlation was at 2 months. It is not clear what statistical methods were used, and recurrence rates and the proportions positive at two months for each regimen only appear to have been available in six of the eight studies. However, across these six trials, the treatment regimen that had the lowest proportion culture positive at two months also generally had one of the lowest recurrence rates (suggesting the two month culture does capture some of the treatment effect on the final endpoint of recurrence). This therefore does indicate that the two month culture could be a useful surrogate, but clearly cannot constitute any formal evaluation.

These three papers (Aber and Nunn, 1978; Mitchison, 1993, 1996) are much cited and are the basis for the general acceptance that being culture positive at 2 months is a good predictor for relapse (Global Alliance for TB Drug Development, 2001), one author even going so far as to identify the 2 month culture bacteriology as a surrogate marker on the basis of this evidence (Sirgel et al., 2000). Fox (1981) identifies the proportion of patients who have achieved culture negative sputum at two months as a valuable clinical index of the sterilizing activity of a regimen. In an extensive review of biomarkers in TB, Perrin et al. (2007) state that the 2 month culture is the only potential biomarker that has been 'validated in clinical trials' without defining precisely what is meant by 'validated'. Spigelman (2007), writing as the President of the Global Alliance for TB Drug Development, identifies the two month culture as 'probably the best available surrogate marker for the relapse rate'. Burman et al. (2006b) state that 'two-month sputum culture conversion is an appropriate surrogate marker for the initial evaluation of a new drug regimen for tuberculosis treatment' basing this statement, in addition, on evidence from Study 22 conducted by the US TB Trials Consortium (TBTC).

In the report of Study 22 (Benator et al., 2002), the authors unknowingly

80

test the Prentice Criteria for the two month culture result combined with base-
line cavitation in a Cox proportional hazard model. They state that the hazard
ratio for the combined failure and relapse endpoint between the two experi-
mental treatments was 1.6 with 95% confidence interval (1.0,2.6), p=0.04. On
adjusting for the two month culture result and cavitation, the adjusted hazard
ratio becomes 1.34 with 95% confidence interval (0.83, 2.18), p=0.23. The treat-
ment effect is no longer statistically significant on adjusting for the two month
culture result and cavitation and it therefore satisfies Prentice's first criterion
(see section 2.5 on page 38). The two month culture and cavitation were both
strongly associated with the endpoint and therefore Prentice's second criterion
was satisfied (Prentice's third criterion was also satisfied). In this trial, the two
month culture result combined with cavitation satisfies Prentice's criteria for
surrogacy. This is not the case, however, in other published trials.

In the two regimens with the same continuation phase in the IUATLD
Study A (Jindani et al., 2004), the proportions positive after two months of
treatment (14% on 2EHRZ/6HE and 23% on 2(EHRZ)$_3$/6HE) did not ac-
curately reflect the proportions with an unfavourable long-term outcome to
treatment (5% on 2EHRZ/6HE and 5% on 2(EHRZ)$_3$/6HE) indicating that the
2 month culture did not fully capture the treatment differences between the
two regimens. In this same study, across all three regimens there was a strong
association between a positive culture at 2 months and an unfavourable out-
come (Mantel-Haenszel estimate of odds ratio stratified by regimen 5.61, 95%
CI (3.34, 9.43).

There is therefore good evidence that the two month culture result is
strongly associated with treatment outcome, and some evidence suggesting
that it might be a useful surrogate marker.

3.5.3.1 Capturing treatment effect before the end of treatment

A six month anti-tuberculosis regimen consists of the two month intensive
phase followed by the four month continuation phase. Two regimens that
are being compared in a clinical trial could have a different combination of
drugs given in the intensive phase or a different combination of drugs given
in the continuation phase or both. In a clinical trial comparing two treatment
regimens, the culture result after two months of treatment is clearly only af-
fected by the first two months of treatment. The two month culture result
cannot therefore *fully* capture the treatment effect when the two treatments

being compared have different drugs given in the continuation phase. (It is also true that the two month culture result cannot capture *any* treatment effect when the two treatment being compared have identical intensive phases.) The first phase of treatment is certainly more *intensive*; four drugs are given together compared to only two in the continuation phase and most of the bacilli are killed during this period. Drugs are given in the continuation phase to kill any remaining, more persistent, dormant bacilli. The most potent part of the anti-tuberculosis regimen is therefore the first two months and it is therefore reasonable to expect the two month culture result to capture a large part of the treatment effect and therefore be a useful surrogate. Nevertheless, it can never fully capture the treatment effect and be a *perfect* surrogate if the regimens being compared differ in their continuation phases.

3.5.3.2 Isolated Positive Cultures

In some of the early TB treatment clinical trials, a very small number of patients were found to have one or two isolated positive cultures, usually of low colony counts, during follow-up following a series of negative cultures. Several studies have demonstrated that these isolated positive cultures are not indicative of true recurrence, but are due to processing contamination or small lesions opening up in a patients' lung that do not cause recurrence of disease.

The first study was a bacteriological study involving three East African laboratories and one London laboratory (Aber et al., 1980). Sputum samples known to be either culture negative or probably culture positive were sent to three East African laboratories and one London laboratory. 45 (0.8%) of 5798 specimens known to be culture negative were found to be culture positive (with low colony counts). The authors concluded that this was due to transfer of bacilli from the positive to the negative samples and the occurrence of transfer was mainly due to the quality of the technician.

A second study involved the examination of 405 (1.1%) isolated positive cultures from 37,429 sputum specimens from three East African laboratories collected during four clinical trials (Mitchison et al., 1980). Since the incidence of these isolated positives decreased over time after the end of chemotherapy, the authors concluded that some isolated positives arose from the lesions of patients. Nevertheless, the authors also showed that some of the cultures were due to transfer in the laboratory due to varying rates across the three laboratories.

A third study applied DNA fingerprinting techniques to 266 isolates of *Mycobacteria tuberculosis* originating from 42 patients who experienced recurrence after the end of treatment and 42 patients who had isolated positive cultures after the end of treatment (Das et al., 1993). Specimens included those taken from these patients before the start of treatment. For only 5 (12%) of the 42 recurrence patients, the fingerprint of the recurrence isolate was different to that from the pretreatment isolate. For 36 (90%) of the 42 patients with isolated positives, the fingerprint of these isolated positives was different to that of the pretreatment isolate. These results showed that the majority of isolated positives were not true recurrences, but rather contamination from another source.

3.5.4 Early Bactericidal Activity

The first *Early Bactericidal Activity (EBA)* study was carried out in Nairobi, Kenya with the results being published in 1980 (Jindani et al., 1980). Counts of viable bacilli in overnight sputum were collected at two day intervals from pretreatment to 14 days after treatment. Patients were given one of 22 different combinations of anti-tuberculosis drugs (including some drugs given in mono-therapy at different doses). It was found that the fall in colony forming units (CFU, the number of colonies of *M. tuberculosis* growing on solid culture giving an indication of a patient's bacterial load) over the first two days differed between drugs and between doses of these drugs. The fall in CFU over the subsequent twelve days was reasonably similar across drug combinations and so the term Early Bactericidal Activity was used to describe the fall over the first forty-eight hours. The unit of this measure is $\log_{10}$ count per ml of sputum per day. The EBA was therefore seen as 'the best measure for discriminating between drugs' (Mitchison and Sturm, 1997); indeed EBA is deemed necessary by the Food and Drug Administration for licensing a new anti-tuberculosis drug (Jindani et al., 2003). EBA is not the only measure for assessing the performance of a drug. Pyrazinamide has a very poor EBA but is a vital drug in the treatment of TB (Mitchison, 2000). EBA studies are usually the one of the early studies for a new drug (e.g. Ciprofloxacin (Kennedy et al., 1993; Sirgel et al., 1997), Moxifloxacin (Pletz et al., 2004) and the new diarylquinoline discovered by Tibotec, TMC 207 (Rustomjee et al., 2008b)).

These data were re-analysed nearly thirty years after the original study (Jindani et al., 2003) to look at the activity of the drug regimens in the 12

days after EBA. The authors highlight the difference between the bactericidal (occurring during the first two days) and sterilizing (the ability to prevent relapse) properties of drugs are separate attributes. They suggest that sterilizing activity may be measured by the rate of decline in CFU after two days (the *Extended EBA*), but show that a study of duration eights days is insufficient to detect differences between treatment regimens and that significant differences were found when the study period was increased to twelve days. The authors suggest that a study of such length would be considered unethical if patients were on mono-therapy (due to potential development of drug resistance, see section 3.3.1) and therefore suggest a *Serial Sputum Colony Count (SSCC)* study where sputum colony counts are measured from the second day after starting treatment to the end of the first month to assess the sterilizing activity of a new drug in combination with other standard drugs.

3.5.5   Serial Sputum Colony Counts

A multi-centre study was initiated in 1993 at the suggestion of the WHO Steering Committee on Treatment of Mycobacterial Diseases to look at EBA and extended EBA up to five days after the start of treatment (Sirgel et al., 2000). The authors found that isoniazid had a highly potent EBA but was almost inactive beyond two days and stated that 'the most important implication of the study is the possibility of assessing the sterilizing activity of anti-tuberculosis drugs by extending the dosage period of EBA studies beyond the first 2 days'. Using data from another study in Nairobi (Brindle et al., 1993), Brindle et al. (2001) extended these ideas by looking at bactericidal activities of drugs during the 2-28 day period. In the original study, 122 patients with culture-positive untreated TB were recruited and sputum samples taken on days 0, 2, 7 ,14 and 28 after the start of treatment. One difficulty arises with extending the colony counts to 28 days is that some patients will become culture negative by this point. In this study, 7% of the samples were negative at 14 days and 20% were negative at 28 days (Brindle et al., 2001). Since the period of time was extended from an EBA study, this was known as *serial sputum colony counting (SSCC)*. Since the colony count is measured on the $\log_{10}$ scale, the authors fitted regression lines treating a negative culture as missing, the same as a culture that was not available, and therefore making the incorrect assumption that this censoring in non-informative. The authors show highly significant differences of the SSCC from days 2 to 28 between different treatment com-

84

binations, concluding that SSCC from days 2 to 28 (and perhaps extending SSCC from days 2 to 56) is 'a powerful method of demonstrating the long term bactericidal (sterilizing) activities of rifampicin and pyrazinamide in the SHRZ regimen' (Brindle et al., 2001). The authors do not have any relapse data from which to draw this conclusion, they are suggesting that the SSCC of these two drugs corresponds reasonably well with what is known about the abilities of rifampicin and pyrazinamide to prevent recurrence.

3.5.5.1 Analysis of Repeated Culture Counts

Data from an EBA or SSCC study consists of repeated CFUs taken at several specific timepoints from baseline to anything up to 56 days. There is considerable variation between serial measurements taken on the same patient as well as variation between patients (Sirgel et al., 2000). In the first EBA study (Jindani et al., 1980), CFUs were available at intervals of two days from baseline to day 14 meaning that each patient had up to eight separate counts. In the re-analysis of this data (Jindani et al., 2003), the authors calculate the rate of fall in $\log_{10}$ CFUs per ml of sputum per day over different treatment periods for each patient. These rates were calculated either based on the two $\log_{10}$ counts at the start and the end of the period or a the linear regression coefficient based on all $\log_{10}$ counts during the period. Multiple regression was used to model the effect of different drug combinations on these rates. They emphasise that there are two distinct phases suggesting that the two periods day 0 to day 2 and day 2 to day 14 were the most important capturing the bactericidal and sterilizing activities of the treatment respectively. They suggest a bi-exponential model with five parameters and two rate constants

$$\log_{10} \text{CFU} = C_1 e^{-k_1 t} + C_2 e^{-k_2 t} + S,$$

where C_1 and k_1 are parameters for the bactericidal activity, C_2 and k_2 are parameters for the sterilizing activity of the treatment, S is the asymptotic value for incomplete killing and t is the time from the start of treatment in days. They include this model merely as a suggestion noting that more results per patient over longer periods will be necessary to obtain good estimates of all five parameters.

Other authors (Brindle et al., 2001; Sirgel et al., 2000) use a similar method fitting straight lines over different treatment periods taking account of be-

85

tween patient variability by calculating rates of individual patients first or using weighted analyses.

A different approach of analysing SSCC data has been proposed by Gillespie et al. (2002), and applied by Gillespie and Charalambous (2003) and Gosling et al. (2003). The authors propose fitting a single exponential curve to the viable counts from each patient using an iterative process to remove the point which fits least well to the curve, refitting the model to this subset of points. This method has subsequently been widely criticised. The main criticism being that 'iterative discarding of data points because they do not fit an incorrect model is a source of bias rather than a help in discarding "discrepant" results' (Mitchison, 2003). Dore and Nunn (2003) also make this point noting that removing a data point from the analysis merely blurs the distinction between change in $\log_{10}$ counts over day 0 to 2 and days 2 to 14. Jindani et al. (2003) also advise against this model as it only contains a single rate parameter over the whole treatment period.

Davies et al. (2006a) use the earlier suggestion of a bi-exponential model to fit a hierarchical non-linear mixed effects (NLME) analysis of the SSCC data. They propose mono- and bi-exponential models as follows:

$$\log_{10} \text{CFU} = \log_{10} \left(e^{\theta_1 + e^{-\theta_2 t}} \right),$$
$$\log_{10} \text{CFU} = \log_{10} \left(e^{\theta_1 + e^{-\theta_2 t}} + e^{\theta_3 + e^{\theta_4 t}} \right),$$

where θ_2 and θ_4 are the parameters for the first and second rates of kill. The authors fit these models to the same data used by Brindle et al. (2001) and show that the preferred model is 'unequivocally bi-exponential', noting that this is consistent with the earlier work showing two different phases of killing. The advantage of a bi-exponential model in this situation is that the first exponential term will dominate for small t, corresponding to the steeper rate θ_2 representing the EBA, and the second term will dominate for large t, corresponding to the shallower rate θ_4 representing the slower sterilizing activity. The authors choose to use this particular dataset as the sputum was not decontaminated. In their discussion, the authors suggest that the finding of only a single population of bacilli in several studies was due to use of decontamination of sputum which may have been responsible for eliminating the first population of rapidly growing bacilli. The authors show that the second rate parameter, θ_3, explains the differences between treatments well

and could therefore be a good measure of the sterilizing activity of a treatment regimen. In this dataset, each patient only had five measurements over time so the authors acknowledge that a model with four parameters was the most complex that could be fitted. The modelling process does account for the between patient variability using mixed effects, but the same problem remains in that negative culture results are treated the same as a missing value (since the $\log_{10}$ of zero is undefined).

Davies et al. (2006b) look at optimal sampling strategies for measuring SSCC and estimating the four parameters in the bi-exponential model. They show that most published sampling schemes are relatively inefficient in estimating these parameters, suggesting instead a scheme of ten measurement taken over the first two months with more taken during the first two weeks.

3.5.5.2 Does SSCC predict sterilizing activity?

While the bi-exponential model proposed by Davies et al. (2006a) does seem to fit the data, allowing for the two different phases of killing and distinguishing between treatment groups in the second phase of killing, there has been no definitive study to explain how well this second rate parameter predicts sterilizing activity. Currently, the only sure measure of sterilizing activity is relapse after the end of treatment (Global Alliance for TB Drug Development, 2001), and there have been no published studies done where sputum samples have been taken more regularly that monthly over the first two months in addition to patients being followed up for long-term recurrence. This point is well made in a recent editorial in reference to extended EBA over days 2 to 7, but applies equally to SSCC measured up to day 56:

> Whether the extended bactericidal activity during days 2 to 7 does represent sterilizing activity of fluoroquinolones or any other drug remains speculative, and awaits the study of relapse rates following phase III clinical trials.

Donald (2006)

The REMoxTB Phase III trial that is currently enrolling, evaluating the addition of moxifloxacin in the standard treatment regimen, should provide some answers to this question. In one site, intensive colony counting is being done during the first two months of treatment in addition to patients being

followed up for relapse. It is expected that it will be several years until completion of follow-up and when this analysis will be possible (see Nunn et al., 2008, for more details).

3.5.6 Other Markers

3.5.6.1 Clinical Symptoms

3.5.6.1.1 Course of Fever during Treatment
Before the discovery of anti-tuberculosis drugs, it was accepted that persistent fever after several weeks of bed rest was indicative of progressive disease (Landis, 1920). Kiblawi et al. (1981) present the results of a retrospective study looking at the course of fever of patients who were hospitalised and treated for tuberculosis. They grouped patients who were febrile on admission into two categories: those whose fever persisted for more than two weeks and those whose fever cleared up within two weeks. They showed that the first group had significantly more patients with far advanced disease as assessed from a pretreatment X-ray. They include an extensive review of research done up to this point and conclude by saying that the course of fever after treatment is variable. They close by saying 'we could not demonstrate any differences in the course of fever that could be attributed to different modern anti-tuberculosis drug regimens' thereby using one of the Prentice criteria to dismiss the course of fever as a surrogate endpoint nine years before Prentice proposed them (Prentice, 1989).

Barnes et al. (1987) present the results of another study of 192 patients looking at the course of fever during treatment of pulmonary tuberculosis. They show that febrile patients are more likely to also have laboratory markers of advanced tuberculosis, and that duration of fever did not differ between different treatment regimens. A second paper (Barnes et al., 1988) on this data looked at prognostic markers for *short-term* outcome claiming to be the first systematic attempt to do so. The authors defined an unfavourable outcome as either respiratory failure requiring intubation or death from any cause during the patient's initial hospitalization. They developed a scoring system calculated from six of the most important prognostic variables that they identified that would identify those patients at greater risk of developing respiratory failure or death. The six variables were: total number of lymphocytes, percentage of neutrophils, age, smear-positive extra-pulmonary tuberculosis, alcoholism and cavitary disease. It is important to note that no cultures were available

88

in this study. They state that their findings were similar to those in Kiblawi et al. (1981), except that they did not find any significant association between extent of disease and prolonged fever.

3.5.6.1.2 Weight Gain Khan et al. (2006) evaluate weight gain during treatment as a marker for treatment failure using data from the randomised controlled trial Study 22 conducted by the Center for Disease Control and Prevention Tuberculosis Trials Consortium (Benator et al., 2002). In participants classified as underweight at diagnosis (defined as 10% or more below ideal body weight), the authors report a relative risk of 1.79 (95% CI 0.96, 3.32) comparing rates of recurrence among those with weight gain of 5% or less during the intensive phase of treatment and those with weight gain more than 5%. In a multiple covariate analysis adjusting for other risk factors, sex and age, this association was stronger (odds ratio 2.4, p = 0.03). Yew and Leung (2006) refer to an earlier study in Tanzania showing weight gain was an unreliable indicator of treatment response and comment that the prognostic importance of weight gain requires further study.

3.5.6.2 Automated Culture Systems on Liquid Media

A retrospective study of 26 patients receiving anti-tuberculosis treatment looked at the correlation of time to detection (TTD) from the MGIT automated culture system (see section 3.2.5.1 for details) with response to treatment (Epstein et al., 1998). The authors divided these patients into 'responders' and 'non-responders' on the basis of *clinical improvement* (defervescence, weight gain, decreased cough or hæmoptysis and increased appetite). They found that TTD fell sharply over time in the first group, but not in the second group. The authors found that the binary measure of TTD less than 20 days was a better prognostic marker than sputum AFB smear evaluation for treatment response although the measure of treatment response could be prone to subjectivity.

Wallis et al. (1998) conducted a study looking at 42 patients on six months of standard treatment for pulmonary tuberculosis. The authors monitored *Mycobacterium tuberculosis* antigen 85 complex and days-to-positivity (DTP) by BACTEC in addition to sputum culture and acid-fast smear. Patients were also followed up for six months after the end of treatment for relapse defined bacteriologically in conjunction with symptoms or radiographic findings consistent with active TB. Using the results of (Epstein et al., 1998), the authors

define *persistent disease* as subjects with at least one BACTEC culture at or after day 90 of treatment that became positive within 20 days (that is DTP $\leq$ 20). Four subjects had persistent disease and expression of antigen 85 complex increased significantly over the first two weeks of treatment only in these four subjects. The authors also found that its concentration on day 14 was a good predictor of persistence. Of the four who had persistent disease, two relapsed and there were no relapse in the rest of the cohort.

Looking at 177 TB patients from a cohort study in South Africa, Carroll et al. (2008) evaluated DTP measures taken during the first two weeks of treatment as predictors of subsequent smear and culture results. Defining a *response ratio*, *r*, as an algebraic combination of DTP values over the first two weeks of treatment, the authors found low sensitivities for *r* as a predictor of culture results at 2 (45%) and 3 (47%) months with high specificities (65% and 64% respectively).

Pheiffer et al. (2008) explored the relationship between DTP and treatment response in sputum samples from 125 patients with TB in one province in South Africa. The authors found that those with a negative smear at two months had had a longer DTP at diagnosis than those with a positive smear. The authors also looked at the relationship between CFU and DTP on the BACTEC 460 and the BACTEC MGIT 960 in sputum samples from 22 TB patients. They found only poor correlation between DTP and $\log_{10}CFU$, $R^2 = 0.22$ and $R^2 = 0.14$ for the BACTEC 460 and the BACTEC MGIT 960 respectively.

In a small sample of 39 isolates from patients with TB, Wallis et al. (1999) found a strong association between days to positivity (DTP) using the BacT/ALERT culture system (bio-Mérieux), $R^2 = 0.99$. Palaci et al. (2007) also found a good association between BACTEC DTP and CFU ($R^2 = 0.63$). These results are then 'confirming that DTP is a reliable surrogate of sputum bacillary load' (Palaci et al., 2007). This is not true as it has been shown in Chapter 2 that 'a correlate does not a surrogate make' (Fleming and DeMets, 1996); all that can be said is that DTP moderately correlates with CFU . There is much discussion about how DTP relates to CFU and more work is needed in this area. Automated liquid culture systems have clear benefits over the slower more labour-intensive solid culture methods (see section 3.2.5.1) and it is therefore important that this relationship is understood as more and more laboratories worldwide are using liquid culture systems.

3.5.6.3 Immunological Markers

In a study of 10 patients, Kennedy et al. (1994) compare *polymerase chain reaction* (PCR) with smear microscopy and culture results for assessing treatment response in patients undergoing chemotherapy for tuberculosis. They show that PCR detected effective treatment in most cases and suggest that it is promising but requires further study.

A study of 19 patients (Desjardin et al., 1999) found a rapid decline in *M. tuberculosis* messenger RNA (mRNA) transcribed from the 85B (alpha antigen) gene in sputum of tuberculosis patients during the first few days of effective treatment suggesting that this may indicate that this marker correlates well with microbial viability. The one patient in their survey who did relapse also had the highest levels of sputum 85B mRNA during the first fourteen days of treatment.

In a study of 22 patients, Thomsen et al. (1999) found that PCR remained positive much longer in patients suffering from extensive disease than in those with less extensive disease and suggest that therefore PCR may be applicable for monitoring the response to treatment for tuberculosis.

Chierakul et al. (2001) evaluate PCR as a predictor of outcome in 53 TB patients in a hospital in Bangkok. Patients tended to convert to negative later on PCR than on smear or culture, but the numbers were too few to determine whether this reflected progress of disease during treatment and the authors concluded that, on limited data, PCR was not useful in monitoring therapy in smear-positive PTB patients. Levée et al. (1994) show similar results in an even smaller group of 19 patients. Afghani et al. (1997) suggest, from a study of 94 specimens from 22 cases of smear-positive TB, that there is no great advantage in PCR over smear in predicting culture results.

3.5.7 Joint TDR/EC Expert Consultation

A meeting was initiated jointly by the Special Programme for Research and Training in Tropical Disease (TDR) and the European Commission in Geneva, Switzerland in June 2008 to evaluate the role of biomarkers and surrogate endpoints in the management of patients with tuberculosis. This meeting allowed for a comprehensive discussion on current biomarkers with the available evidence supporting these markers and sources of possible future biomarkers. Tables 3.2 on page 94 and 3.3 on page 95 show the current status of all avail-

able biomarkers for tuberculosis and are reproduced from the report of this meeting with permission (Zumla et al., 2008).

3.6 Discussion

As summarised in this chapter, a large number of candidate predictors, prognostic markers and surrogate markers for poor outcome to tuberculosis chemotherapy have been proposed in the literature. These can be grouped under three headings: *possible risk factors, recently proposed markers* and *markers with limited evidence suggesting their potential use as surrogates.*

Most of the proposed markers are *possible risk factors*. These are factors that may, at best, indicate that a particular patient is more or less likely to respond to treatment. Such risk factors include smoking, alcoholism, indicators of poor adherence, HIV infection, weight and radiographic assessments of disease and cavitation. These are independent of and measured prior to the initiation of treatment and therefore cannot be prognostic or surrogate markers. However, they may be useful in explaining some of the variation in treatment response between individuals (although this variation is usually removed with randomisation in a clinical trial) and the inclusion of important covariates might improve the fit of statistical models.

Some of the markers are *technologically advanced, recently proposed markers*. To measure these markers newer, advanced techniques often using expensive laboratory facilities are required. Such markers include those immunological and molecular markers and DTP from automated culture systems described in sections 3.2.6 and 3.5.6. To evaluate one of these markers as a surrogate endpoint, it must be assessed in patients during treatment in addition to patients being followed up for long-term treatment outcome. There is therefore often none, or very little such data available and analysis of these markers as prognostic and surrogate markers is therefore not yet possible. Typically, such markers are prohibitively expensive for use in laboratories with limited resources.

What remains of the list of possible markers are *SSCC* and *monthly culture results*. Monthly sputum smear results during treatment have too low sensitivity to be useful as prognostic markers and therefore will not be considered as surrogate markers. As described in section 3.5.5.2, whilst there are indications that summary parameters of SSCC may in fact prove to predict treatment

failure and be able to be used as a surrogate, insufficient data exists with participants followed up for relapse in addition to having SSCC data recorded. Extensive data do exist to evaluate monthly culture results as prognostic and surrogate markers. Several published articles have indicated that, in particular, the two month culture result has value as a prognostic or even surrogate marker, though actual analysis of the available data providing evidence of this is scant. Cultures are collected in some TB control programmes and the use of cultures as markers for treatment failure could therefore be of immediate use. Therefore, the primary objective of this research project is to evaluate monthly cultures as prognostic and surrogate markers for long-term response to treatment for tuberculosis and, following the recommendations from the literature (see section 3.5.3), the culture result at two months will provide the focus.

Candidate biomarker	Association	Study size and positive treatment outcome
Chest radiography		
Baseline chest X-ray	Recurrence	46/938
Baseline chest X-ray	Relapse	74/930
Baseline chest X-ray	Relapse	4/237
Baseline chest X-ray	Recurrence	24/175
Serial sputum microbiology		
M2 SCC	Recurrence	Many patients
Serial sputum CFU counts	Superior sterilizing activity	Few patients
Serial MGIT™ or BACTEC™ time to positivity	Anti-TB TrR, failure and relapse	Many patients
Early bactericidal activity	None	Many patients
TB-specific biomarkers		
Sputum antigen 85B RNA	Anti-TB TrR	1/18
Sputum antigen 85	TrR	2/40
Sputum antigen 85	Drug evaluation	40
Urine *Mtb* DNA	Anti-TB TrR	20
Urine lipoarabinomannan	Infection, active disease	Many patients
Anti-ESAT-6, 38 kDa protein, alanine dehydrogenase, malate synthetase	Extent of disease, TrR	168
IGRA	Anti-TB TrR	5/18
Breath biomarkers	Culture plates data Active disease	19/23
Nonspecific biomarkers of immune activation		
NKT cells at diagnosis	M2 SCC	8/21
Sputum IFN-γ	Anti-TB TrR	15
sIL-2R	Anti-TB TrR	44
sTNF-R, granzyme B at diagnosis	M2 SCC	18/36
Neopterin	Anti-TB TrR, relapse	11/39
		31
C-reactive protein	Anti-TB TrR, death	105
		100
		18
sICAM-1	TrR	30
suPAR	Death	101

Table 3.2: Current status of TB biomarkers. Continued in Table 3.3 on the next page.

Candidate biomarker	Association	Study size and positive treatment outcome
Functional studies of TB protection		
ELISPOT and QuantiFERON®	Vaccine effect Immune eradication of *Mtb* infection	Many patients
QFN	Prediction of disease in untreated contacts (using upper bound cut-off)	5/6
		6/6
Whole blood killing	TST effect BCG effect	12
		10
		50
Whole blood killing	AIDS effect	22
Whole blood killing	Combined ART effect	15
Whole blood killing	TNF monoclonal antibody effect	20
Whole blood killing	Vitamin D effect	192
Functional studies of anti-TB treatment		
IGRA	LTBI TrR	38
IGRA	Anti-TB TrR	5/18
Whole blood killing	Anti-TB TrR	Many patients
Whole blood killing	Correlation between serial CFU slope and M2 SCC	36
Whole blood killing	Anti-TB TrR	10
Highly multiplexed assays		
Transcriptomics	TB disease and infection	40
Proteomics	TB disease	60
Metabolomics	TB disease	NA

Table 3.3: Continued from Table 3.2 on the previous page. Current status of available biomarkers in TB from the 2008 joint TDR/EC expert consultation on biomarkers in tuberculosis. References for each study are available from the meeting report. This table is reproduced with permission from Zumla et al. (2008). ART, antiretroviral therapy; BCG, bacille Calmette-Guérin; CFU; colony forming unit; ESAT, early secretory antigenic target; IFN, interferon; IGRA, interferon-gamma release assay; IL, interleukin; LTBI, latent TB infection; MGIT™, Mycobacterium Growth Indicator Tube; M2 SCC; month-two sputum culture conversion; *Mtb, Mycobacterium tuberculosis*; NA, not available; NKT, natural killer T cells; sICAM-1, soluble intercellular adhesion molecule type 1; suPAR, soluble urokinase plasminogen activator receptor; TrR, Treatment Response; TNF, tumour necrosis factor; TST, tuberculin skin test.

Chapter 4

Overview of Data and Introduction to Analysis

4.1 Introduction

In the previous chapter, possible prognostic and surrogate markers for poor outcome to treatment were summarised. It was demonstrated that culture results during treatment are the most promising markers for which there is data available to evaluate as prognostic and surrogate markers.

In this chapter, the data used in this thesis to evaluate culture results as prognostic and surrogate markers will be introduced and described. In this section the justification for the data used will be discussed. Section 4.2 summarises the process of data entry and cleaning and section 4.3 describes the definition of the clinical endpoint. The data is described in section 4.4 with some tables and figures summarising the data presented in section 4.5.

4.1.1 Choice of Data

4.1.1.1 Trial vs. Observational Data

A surrogate marker is one which fully captures the effect of the treatment on the true endpoint. It is therefore particularly important that the observed treatment effect is indeed the true treatment effect and not contaminated by other factors. Data from prospective randomised controlled trials are therefore required for the evaluation of surrogate endpoints as this is the only known

way to control for unknown confounders (Lassere et al., 2007b). Retrospective observational studies contain serious potential biases that are removed when participants are prospectively enrolled into a trial and randomised to treatment (Pocock, 1983). To assess the treatment effect, it is also important that there is a control arm so that within-trial treatment comparisons are possible. In a clinical trial, participants enrolled will be followed-up after the end of treatment, but this is not always possible in a purely observational study. Follow-up is particularly important in the context of TB since relapse within twelve to eighteen months after the end of treatment is the main endpoint in clinical trials. For use in a clinical trial, a surrogate marker must first be evaluated in the context of previous clinical trials. Clinical trial data are therefore necessary for an evaluation of a surrogate endpoint.

4.1.1.1.1 Individual Patient Data

Daniels and Hughes (1997) and Gail et al. (2000) both emphasise the importance of having *individual patient data (IPD)* from clinical trials available for the evaluation of a surrogate marker, rather than just summary statistics. The authors suggest that it is less important to have IPD if the between-study variation is much greater than the within-study variation. In any TB clinical trial, there is often considerable variation in treatment outcomes and it therefore cannot be said that the within-study variation is minimal. Having IPD for the evaluation of surrogate markers in TB is therefore vital.

4.1.1.2 Datasets to be used

For the purpose of this research project, data will be used from selected tuberculosis clinical trials that were conducted by the British Medical Research Council (MRC) during the 1970s and 1980s to identify the most effective regimen for treating tuberculosis. A large amount of what we know today about the treatment of tuberculosis comes as a result of these trials. In 1999, an entire 49-page supplement to the International Journal of Tuberculosis and Lung Disease (IJTLD) was dedicated to these trials, (Fox et al., 1999). Other reviews also confirm the central importance of these MRC studies (D'Esopo, 1982; Christie and Tansey, 2005).

These trials were chosen for the following reasons:

- *Randomised controlled trials of high quality.* The streptomycin trial conducted in the UK by the MRC in the 1940s is generally regarded to be

the first published randomised controlled trial (Hill, 1990). Subsequent MRC TB trials were all randomised and conducted with an emphasis on the use and development of this scientific method. In 1979, when Archie Cochrane judged which disease area in medicine was the most evidence-based, he gave the 'gold medal' to the tuberculosis specialists (Cochrane, 1979). Clinical trials conducted to a high standard result in data quality of a high standard.

- *Large differences in recurrence rates between treatment arms.* Since the MRC trials were the means by which effective treatment regimens were evaluated, they naturally include less effective regimens leading to a high number of recurrences. Importantly, recurrence rates differed greatly between treatment arms within any one trial. A large treatment effect gives greater power for the evaluation of surrogate markers.

- *Six month regimens.* In common with the standard treatment recommended today, most of the treatment regimens evaluated on these trials were of duration six months. Trials arms with treatment duration other than six months are excluded from analyses. As discussed in Chapter 2, a surrogate can never be 'generally applicable' (Day and Duffy, 2000) and will only be valid in a trial involving the same drugs that were used for its evaluation. For the purpose of this research project, the data pertain to participants who were on six months of treatment using the same drugs that are largely still in therapeutic use today and will be used today in clinical trials (any new drug will be evaluated as part of a regimen involving three or four other drugs from the standard regimen). No new drugs for the treatment of TB have been developed for forty years (Spigelman, 2007). This means that data from historic clinical trials can still be used to evaluate surrogate markers for use in clinical trials being conducted today.

- *Multi-arm, large Trials.* Whilst authors disagree about the exact methods of evaluating surrogate markers (see Chapter 2), there is almost unanimous agreement that it must be done across a number of different trials. A surrogate marker that is shown to fully capture the treatment effect across a large number of treatment comparisons and a large number of individuals will be of greater use and be more convincing than one which is validated across only a handful of treatment comparisons.

These data consist of over 7000 individuals from 12 clinical trials includ-
ing 49 different treatment regimens.

- *Good Laboratory Data.* For each trial, samples were sent to a central lab-
oratory supported by MRC laboratories in Britain. Contamination was
kept to a minimum, non-tuberculosis mycobacteria were generally iden-
tified as such and resulting laboratory results were of a consistently high
standard.

- *Extensive Follow-up after the end of treatment.* In all trials, participants
were followed-up for a minimum of eighteen months after the end of
treatment and sputum samples taken on a monthly or in some trials
bimonthly or trimonthly basis during this follow-up period. In some
trials, participants were given vitamin tablets after the end of treatment
as an incentive to return for follow-up visits. On treatment cards from
all trials, there is clear documentation of home visits, written correspon-
dence and visits to family and friends in order to minimise participant
loss during follow-up.

- *Additional Covariates Recorded.* Additional variables recorded in all or
some of the trials include: age, sex, weight, drug resistance, baseline
radiographic measurements, smear results, race and centre. In a small
number of trials, radiographic measurements were also recorded at sev-
eral points during treatment. There is great variation in clinical outcome
and culture results during treatment between participants and some of
this variability may be removed by controlling for participant covariates
(and has been found in similar work on SSCC in Davies et al. (2006a)).

The following reasons relate to the homogeneity of the trial participants
included in these data. This homogeneity removes some complexity from
the analysis and interpretation of the results, but can have an impact on the
generalisability of conclusions drawn with the first point having the greatest
impact.

- *No HIV co-infection.* These trials were conducted in an era before the
existence of the human immunodeficiency virus (HIV) and so all partic-
ipants were uniformly HIV seronegative. It is known that the rifamycins
(in particular rifampicin) interact with some therapies given to treat
HIV and therefore participants with TB and HIV co-infection respond

to treatment differently (Blumberg et al., 2005) and prognostic and surrogate markers may be more or less valuable in patients with or without HIV.

- *Exclusively Pulmonary Tuberculosis.* Participants with extra-pulmonary TB require different treatment and have a different response to treatment (see section 3.1.2). Participants detected to have extra-pulmonary were excluded from the trials and only those with pulmonary tuberculosis were enrolled.

- *No MDR-TB.* Rifampicin use was not widespread in the communities in which these trials were conducted and therefore very few participants were rifampicin-resistant and therefore multi-drug resistant TB was not a problem.

These data therefore provide an unprecedented opportunity to determine the value of early culture results as prognostic and surrogate endpoints.

4.1.1.3 Trial Summary

Trials were identified using the aforementioned IJTLD article written by Wallace Fox in 1999 (Fox et al., 1999) and in communication with those who were involved with TB research during this period. Only trials for which individual patient data was available were considered. Trials were selected that met the following inclusion criteria (see above for details): six month regimens, regimens including standard drugs only, differing recurrence rates between regimens, at least eighteen months of follow-up and fairly large sample sizes (at least 50 in each treatment arm).

Regimens are described in a format such as: $2SHRZ/4S_2H_2$. The combination of letters before the forward slash (S, H, R and Z) correspond to the drugs given in the intensive phase and the combination of letters after the forward slash correspond to the drugs given in the continuation phase. The letter codes are as follows: E (ethambutol), H (isoniazid), R (rifampicin), S (streptomycin), T (thiacetazone), Z (pyrazinamide). The number at the beginning of the letter combination denotes the duration of the intensive phase in months (2 in the example). The number immediately following the forward slash denotes the duration of the continuation phase in months (4 in the example). The absence of a forward slash indicates that the intensive and the continuation phases included the same drugs at the same dosing frequency and are

therefore not shown separately. Any subscripts indicate the weekly dosing frequency of that particular drug (S_2H_2 indicates that streptomycin and isoniazid were given twice-weekly). An absence of subscript indicates that the drugs were given daily (S, H, R and Z were given daily in the intensive phase in the example). This closely follows the notation given in the original trial papers.

Twelve trials which satisfied the inclusion criteria were selected and are summarised below. Publications describing and resulting from the clinical trials, if not stated in the text below, are listed in Fox et al. (1999).

4.1.1.3.1 Trials Included

Study R (East Africa, 1970) This was the first large RCT to show that short-course chemotherapy of only six months was as effective as the standard eighteen months of treatment. The study demonstrated that a six month regimen containing the drug rifampicin (hence *Study R*) in addition to isoniazid and streptomycin (6SHR) had a superior recurrence rate to the standard eighteen month regimen of isoniazid, streptomycin and thiacetazone.(2SHT/16TH). Three other six month regimens were evaluated in this study (6SHT, 6SHZ and 6SH) and sites were located in Kenya, Tanzania, Uganda and Zambia. There were four publications resulting from this study: (East African/British Medical Research Council, 1972, 1973a, 1974b, 1977). Patients on the 18 month regimen are not included in analyses in this thesis.

Study T (East Africa, 1972) Following Study R, Study T showed the benefit of adding pyrazinamide in the first two months. This study compared four six months regimens: 6SHR, 6HR, 2SHRZ/4TH and $2SHRZ/S_2H_2Z_2$. There were two publications resulting from this study.

Study U, (East Africa, 1974) Study U expanded on Study T showing the sterilizing activity of SHRZ was better than SHR for the first two months, and that a one month intensive phase was too short. The trial compared four six month regimens (2SHRZ/4TH, 1SHRZ/5TH, $1SHRZ/5S_2R_2Z_2$ and 2SHR/4TH) with four of the same regimens with the continuation phase extended so that the total duration was 8 months.

This trial showed the eight month regimen of 2SHRZ/6TH was efficacious and therefore eight month regimens were widely adopted in many countries

(with thiacetazone later being replaced with ethambutol). The eight month regimen 2SHRZ/6EH was later found to be inferior to the six month regimen 2SHRZ/4HR in a recent IUATLD study (Jindani et al., 2004) and is subsequently no longer recommended. There were two publications resulting from this study. The eight month regimens are not included in analyses in this thesis.

Study X, (East Africa, 1976) Study X evaluated the effect of adding pyrazinamide and rifampicin to the continuation phase and reducing it to two months. Five four month regimens were compared in this study (2SHRZ/2HRZ, 2SHRZ/2HR, 2SHRZ/2HZ, 2SHRZ/2H and 2HRZ/2H). This study showed that four months of treatment was insufficient with these drugs (several trials are being conducted today to determine whether four month regimens incorporating fluoroquinolones or high dose rifamycins can be effective against TB) and was the first evidence that pyrazinamide gave no added benefit in the continuation phase. There were two publications resulting from this study.

High recurrence rates in all of the four month regimens were observed during the course of the trial and a decision was made to extend the continuation phase to four months for those in the study from that point forward. A number of patients originally enrolled in the study were therefore given six months of treatment and followed up for outcome to treatment. Data from these patients given six months of treatment is included in the analyses in this thesis, and those only given four months of treatment will be excluded. The five regimens included are therefore 2SHRZ/4HRZ, 2SHRZ/4HR, 2SHRZ/4HZ, 2SHRZ/4H and 2HRZ/4H.

Study Y, (East Africa, 1978) Study Y compared a 6 month continuation phase containing only isoniazid (2SHRZ/6H) with three 4 month continuation phases (2SHRZ/4H, 2SHRZ/4HR and 2SHRZ/4HZ). This study showed that recurrence rates decreased only after the addition of rifampicin, showing the importance of rifampicin throughout treatment. There were two publications resulting from this study. The eight month regimen is not included in the analyses in this thesis.

Tanzania Short-Course Chemotherapy Investigation (Study A, 1979) This study, conducted in Tanzania, compared 2SHRZ/4TH with 2SHRZ/4H and

demonstrated that the addition of thiacetazone in the continuation phase did reduce recurrence rates. There was one publication resulting from this study.

First Hong Kong Short-Course Study (1972) Alongside, and slightly later than the studies in East Africa, a number of trials were conducted in Hong Kong and Singapore. These studies 'concentrated on intermittent regimens well suited to their urban communities, and were less restricted by drug costs' (Fox et al., 1999), particularly the high cost of rifampicin which was prohibitively expensive in an East-African setting.

None of the regimens compared in the first study in Hong Kong contained rifampicin, in contrast with regimens in subsequent studies which all contained rifampicin at least in the intensive phase. This study would now be described as a 2x3 factorial study and compared the regimen given SHZ given either daily, twice weekly or thrice weekly and given for either six months or nine months. The number of doses given during a week did not greatly affect recurrence rates, but the recurrence rates in the six month regimens were too small with acceptable recurrence rates only in the nine month regimens. This study showed that a six month regimen without rifampicin resulting in unacceptable recurrence rates, indicating that rifampicin was necessary in a six month regimen. There were two publications resulting from this study. The nine month regimens are not included in the analyses in this thesis.

Second Hong Kong Short-Course Study (1974) The second Hong Kong study compared the control regimen of 6SHR against a selection of intermittent regimens (2SHRZ/$S_2H_2Z_2$, 2SHRE/$S_2H_2E_2$ and 4$S_3H_3R_3Z_3$/$S_2H_2Z_2$) either as six month regimens or with the continuation regimens extended to make eight month regimens. Recurrence rates in the regimens not containing pyrazinamide were poor and was the first evidence showing the poor sterilizing activity of ethambutol and the importance of pyrazinamide in the intensive phase. There were two publications resulting from this study. The eight month regimens are not included in the analyses in this thesis.

Third Hong Kong Short-Course Study (1977) The third Hong Kong study compared a daily control regimen of 6EHRZ with four thrice-weekly four-drug regimens: 6$E_3H_3R_3Z_3S_3$, 6$S_3H_3R_3Z_3$, 6$E_3H_3R_3S_3$ and 6$E_3H_3R_3Z_3$. All regimens, with the exception of the regimen not containing pyrazinamide,

103

had very low recurrence rates. There were three publications resulting from this study.

Fourth Hong Kong Short-Course Study (1979) The fourth Hong Kong study explored the duration of pyrazinamide in a six month regimen. $6H_3R_3Z_3$ was compared with $6S_3H_3R_3Z_3$ where the pyrazinamide was given for two, four or six months. Recurrence rates were low in all four regimens showing that pyrazinamide gave little added benefit after two months. Unusually, 2% of those on the regimen without streptomycin failed treatment showing the importance of four drugs in an intermittent six month regimen. During the course of this study some patients were given the same regimens except with the HRZ given as a combined formulation. There was no reported differences in this study between those who took the combined tablet and those who took separate tablets and so this distinction is not considered in the analyses in this thesis and the fourth Hong Kong study only provides four regimens. There was one publication resulting from this study.

First Singapore Short-Course Study (1973) The first Singapore study further explored the benefit of pyrazinamide beyond two months comparing 2SHRZ/2HR and 2SHRZ/4HR with the same regimens given pyrazinamide in the continuation phase. Recurrence rates were very low in both six month regimens, but only moderately low in the four month regimens. The two four month regimens are not included in the analyses in this thesis. There were three publications resulting from this study.

Third Singapore Short-Course Study (1983) The third Singapore study explored the use of a combined tablet of isoniazid, rifampicin and pyrazinamide. Six regimens were compared: $2SHRZ/4H_3R_3$, $1SHRZ/5H_3R_3$ and $2HRZ/4H_3R_3$ with HRZ given in the intensive phase as a combined preparation or as separate tablets. recurrence rates in the regimens with the combined formulation were slightly higher in those with separate tablets. There was one publication resulting from this study.

4.1.1.3.2 Trials Excluded

Kenya Levamisole Study (1981) The purpose of this study was to determine whether the addition of the immunostimulant Levamisole affected the

104

activity of the 2SHRZ/TH regimen. While the individual patient data was available, this drug is not used today for the treatment of TB and so the data from this study are not used in this thesis (Kenyan/Zambian/British Medical Research Council, 1989).

Second Tanzanian Chemotherapy Investigation (1982) This study was conducted to compare two seven month treatment regimens. No six month regimens were included in this trial and so the data are not used in this thesis (Tanzania/British Medical Research Council, 1996).

Second Singapore Short-Course Study (1978) Three different six month regimens were compared in this study. All three regimens were highly effective and recurrence rates within two years from the start of therapy were only 1% in each arm. The treatment effect size is too small to be of any use in evaluating surrogate endpoints. These data are not used in this thesis (Singapore Tuberculosis Service/British Medical Research Council, 1985, 1988).

The Hong Kong Silico-tuberculosis Study (1980) This study compared a six month regimen with an eight month regimen for the treatment of silico-tuberculosis. Silico-tuberculosis is a particular occurrence of tuberculosis of the lung that is inflamed with the presence of foreign bodies (often inhaled particles of silica) and is found primarily among workers at mines and quarries. Patients with silico-tuberculosis respond differently to those with non-silicotic TB and data from this trial are therefore not used in this thesis (Hong Kong Chest Service/British Medical Research Council, 1991b).

4.2   Data Entry and Validation

4.2.1   Introduction

The twelve trials described above were all completed by the mid-1980s and the results published by the beginning of the 1990s. The analyses were mostly, if not entirely, conducted by hand and no electronic copy of the data exists from this time. Therefore, while individual patient data is available from each of these trials, it exists in the form of patient cards that have been kept in storage. The first, lengthy step in the analysis of these data is converting them into an electronic form (*data entry*) and checking the quality of these data (*validation*).

4.2.2 Methods

4.2.2.1 The First Four East African Trials

The first four East African trials (Studies R, U, T and X) had been entered onto the computer over a period of around five years prior to the start of this research project. This was done with the purpose of performing a re-analysis of data from the key trials using more modern methods of statistical analysis (such as adopting the intention-to-treat principle, survival analysis and imputation techniques). This re-analysis is still ongoing and the results are yet to be published.

The data had been entered using the process of *double-data-entry* (*DDE*). For each of the four trials, the data were entered in full on two separate occasions by two independent persons. Two databases therefore exist for each trial. While some studies have shown that double data entry is neither necessary nor sufficient for ensuring good quality data (Day et al., 1998; Gibson et al., 1994), most of the human error in transcribing the data from case record forms (CRFs) to computer can be identified and eliminated with DDE. This task of DDE validation is also a time-consuming task as individual CRFs will have to be re-checked for every discrepancy that is found between the two databases.

Prior to the start of this research project, these data had been entered, but no formal comparison of these two databases or any analysis had been carried out and the databases largely remained in their raw, uncleaned state. Using Stata 9.2 (StataCorp. 2005. *Stata Statistical Software: Release 9*. College Station, TX: StataCorp LP.), the two independent copies of data from each of the four trials were compared record-by-record to find discrepancies. Discrepancies were then checked against the original CRFs and the records updated on a final database.

In addition to the process of DDE validation, treatment cards were manually checked to determine the endpoint for each participant. The different classifications of treatment failure and treatment cure are not necessarily exactly the same as those used in the original trial analyses (discussed below in section 4.3) and so it was important, even after the double data entry validation had been performed, to check each endpoint. Any doubtful recurrences were studied carefully and classified after discussion with other TB experts. These discussions and resulting classifications were made blind to the allocated treatment.

Treatment cards were then separated into groups of treatment regimens

and manually checked against that stored in the database. In addition to endpoint classification, treatment allocation is an important data element and so the regimen marked on each treatment card was checked.

4.2.2.2 The Remaining Eight Trials

No electronic copy of the data from the remaining eight trials existed prior to this research project. A data entry environment was set up using the software EpiInfo 6.04d (CDC, USA; WHO) and data entered during the months of July and August 2006. Due to time constraints and the limited additional benefits of DDE, these data were only single-data entered.

Each treatment card contains a large variety of information including participant characteristics, medicines given, drug sensitivity results, laboratory results, details of postmortem (if performed), some clinical symptoms and any additional comments by trial clinicians or investigators. Not all of this information is of use for the purpose of this research project, so only a subset of these data elements were selected for data entry. Table 4.1 on the following page summarises the list of variables selected for data entry. Participant characteristics were chosen as those most likely to be risk factors. In some of the trials, x-rays were taken during treatment. Radiographic extent of disease and cavitation has been recorded where they were assessed.

As with the first four trials, each treatment card was manually checked against the database to confirm the endpoint classification. Any undecided classifications were confirmed on consultation with TB experts. Again, the allocated treatment arm was manually checked against the database for every treatment card to ensure that the regimens recorded on the database were correct.

A random selection of 5% of treatment cards from these eight trials were selected and each data element on the treatment card checked against the database to assess the overall quality of the data. The results are shown in Table 4.2 on page 109.

The percentage error was calculated assuming 25 pieces of information on each of the treatment cards sampled. The overall error rate of 1.2% with 95% confidence limits of 0.9% and 1.5% indicates a high quality of data entry. It was decided that this was an acceptable error rate and there was therefore no need to check the data further or employ double data entry.

Variable	Type	Comments
Trial-specific Data Elements		
Trial[†]	string	Name of trial
Region[†]	string	East Africa, Hong Kong or Singapore
Trial Number[†]	string	Unique participant identifier
Treatment Arm[†]	string	Allocated treatment regimen
Control Arm[†]	boolean	Flag to identify the control arm in a trial (see section 6.1.1)
Date of Randomisation	date	
Patient Characteristics		
Age	integer	
Sex	boolean	Male, Female
Weight	float	
Race	string	Chinese, Indian, Malay or Eurasian (recorded in Singapore studies only)
Baseline Resistance[†]	string	Susceptibility to Isoniazid and/or Streptomycin
Baseline Smear	string	Nil, Slight, Medium or Heavy
Baseline Culture	string	0, 1-19, 20+, Innumerable Colonies (IC) or Confluent Growth (CG)
Baseline Disease[‡]	string	Radiographic extent of disease
Baseline Cavitation[‡]	string	Radiographic extent of cavitation
Assessments During Treatment		
Smear Results	string	Measured at months 1, 2, 3, 4, 5 and 6
Culture Results	string	Measured at months 1, 2, 3, 4, 5 and 6
Extent of Disease[‡]	string	Measured at months 4 and/or 6
Extent of Cavitation[‡]	string	Measured at months 4 and/or 6
Endpoint Classification		
Unfavourable	boolean	Unfavourable at end of treatment
Last Culture	integer	Month of last non-missing culture
Recurrence[†]	boolean	
Month of recurrence	integer	
Died[†]	boolean	
Date of Death	date	
Month of Death	integer	
Cause of Death	string	Free text including conclusions of clinical examination and post-mortem (if conducted)
TB Death	boolean	Cause of death TB or other respiratory disease
Endpoint	string	Final classification of endpoint (see section 4.3)
Time to endpoint	integer	Time from randomisation to failure or censoring

[†]Denotes variables that cannot be missing.
[‡]On chest x-ray.

Table 4.1: Variables selected for data entry.

Study	Number of Participants		Errors	Percentage Error [c]
	Entered [a]	Sampled [b]		
Study Y	552	28	13	1.9%
Tanzania	318	16	6	1.5%
Hong Kong 1	246	12	6	2.0%
Hong Kong 2	392	20	6	1.2%
Hong Kong 3	1207	60	12	0.8%
Hong Kong 4	1605	80	19	1.0%
Singapore 1	201	10	4	1.6%
Singapore 3	310	16	4	1.0%
Overall (exact binomial 95% confidence intervals)				1.2% (0.9%, 1.5%)

[a]The total entered is usually greater than the total included in the analysis. See section 4.4.2.

[b]Size of sample was taken as the nearest integer to 5% of the total entered.

[c]The number of data pieces entered from each card varied across trials and, for the purpose of this calculation, was taken to be 25. Percentage error was calculated as (Number of Errors/(25 · Size of Sample))%.

Table 4.2: Error checking report

4.3 Endpoint Definition

Culture results were available monthly during the six months of treatment and almost as regular after the end of treatment for a time. In most trials, follow-up was for twenty-four months after the end of treatment, but in a small number of trials (see table 4.5 on page 116 for details) follow-up was only for eighteen months. Culture results were taken monthly for most of the follow-up period, and every two or three months in the remainder of the follow-up period.

A combined endpoint, *poor outcome*, defined as failure at the end of treatment or recurrence following apparently successful treatment, is the clinical endpoint for the purposes of this research project . In some trials, participants were followed up beyond the scheduled end of follow-up, but any culture results or deaths occurring more than three months after the end of follow-up have been ignored.

There are three possible outcomes for participants enrolled in these trials. Either an individual will be classified as having a *poor outcome*, they respond to treatment without recurrence and have a *fair outcome*, or they have a *missing outcome*. Individuals with a fair outcome are classified as censored for time-to-event analyses and will simply be considered as having a fair outcome when considering a binary treatment response. Similarly, individuals with a

109

missing event will be excluded from all analyses (assuming the missingness mechanism is non-informative). Table 4.3 gives the classification algorithm which is described in more detail below.

Classification	Outcome	Event or Censoring Time
FAIR OUTCOME	Non-TB death	Month of Death
	Loss to follow-up	Month of Last Culture
	Long-term cure	Month 27 or 33*
POOR OUTCOME	Failure during Treatment	Month 6
	Failure after successful treatment	Month of first indication of recurrence
	TB death	Month of Death
MISSING OUTCOME	Default before the end of treatment	N/A
	Death during Treatment	N/A

*Three months after the scheduled end of follow-up to allow for late failures. In the second, third and fourth Hong Kong trials, follow-up ended twenty-four months after randomisation. In the remaining trials, follow-up ended thirty months after randomisation.

Table 4.3: Endpoint Classification

- *Default or death during the first four months of treatment*—MISSING OUTCOME. If an individual defaults or dies before the end of treatment with a last non-missing culture at month 4 or earlier, the end of treatment status for this participant will be unobtainable and the individual will be classified as having a missing status. It was felt that individuals who default or die within the first four months of treatment have had insufficient time to demonstrate a response to treatment, either favourable or otherwise, and so cannot be assigned an outcome. If a culture result is available at months 5 or 6, the end of treatment status will be determined on the basis of these results.

- *Failure during Treatment*—POOR OUTCOME. Failure during treatment, corresponding to treatment non-response, can only be assessed at the end of treatment, during the fifth and sixth months after randomisation. If a participant has heavily positive cultures[1] at months five or six, or cultures with any number of colonies at months five and six, then the participant will be said to have an unfavourable status at the end of treatment. Such individuals are classified as failures at 6 months. If there is not

---

[1] A heavily positive culture is defined as a growth of 20 or more colonies.

110

such strong evidence for failure during treatment, the individual will not be classified as a failure during treatment. Such strict criteria may miss some of those failing at the end of treatment, but these cases will be subsequently be picked up as recurrences after the end of treatment.

- *Failure after Successful Treatment*—POOR OUTCOME. A participant who responds to treatment, (not classified as failure during treatment), but subsequently shows conclusive evidence of tuberculosis disease in follow-up is defined as failure after successful treatment or *recurrence*. The time of failure will be defined as the month of the earliest indication of recurrence. One of the following is required to show evidence of recurrence

  1. Two heavily positive cultures (20+ colonies) within three consecutive months.

  2. Three positive cultures within four consecutive months.

  3. Evidence of retreatment after a clinician's assessment of recurrence of disease.

  4. Participant defaults and is lost to follow-up immediately following a positive culture. Participants showing even scant evidence of recurrence (such as only a single positive culture with a low colony count) before being lost to follow-up are classified as recurrences since there is no more information to confirm or discount.

- *Death after Successful Treatment*—POOR OUTCOME / FAIR OUTCOME. A cause of death from clinical inspection and, in some cases postmortem, is usually available for individuals who die after the end of successful treatment and, in most cases, it is clear whether the death was related to tuberculosis or not. If there is clear evidence that the death was unrelated to tuberculosis or any other respiratory disorder, then it can be classified as a censored observation. If, on the other hand, the cause of death is related to any respiratory disease then this is a recurrence whether or not it has been confirmed bacteriologically and is classified as a treatment failure. The time of the failure or censoring is taken to be the month of death.

- *Long-term Cure / Loss to follow-up*—FAIR OUTCOME / MISSING OUTCOME. Participants who respond to treatment and who have not died or shown evidence of recurrence up to three months beyond the end of follow-up

111

(twenty-four or eighteen months after the end of treatment, depending on the trial) are classified as having long-term cure.

Some participants are lost to follow-up before the end of follow-up. If such participants do not show evidence of recurrence and have completed at least twelve months of follow-up, then they too are classified as having a fair outcome. When considering the endpoint as binary, participants who are lost to follow-up less than twelve months after the end of treatment will be classified as having a *missing outcome*. In a time-to-event analysis, such censoring can be accounted for, but if the endpoint is considered as binary, such participants must be classified as having a missing event. A participant must have at least twelve months of follow-up after the end of treatment to be classified in the analysis of the binary outcome as having a favourable outcome.

There are a handful of participants who did not clearly fit into one or other of these categories. Such participants were classified on a case-by-case basis in consultation with more experienced TB specialists who were not told the treatment allocation.

4.4 Description of the Data

4.4.1 Introduction

Data from 7307 trial participants were initially included in the database. The participants were from 12 trials conducted over two decades in 6 countries. These data include 50 treatment arms, removing the 19 regimens from these 12 trials that were excluded as described previously (see section 4.1.1.3).

4.4.2 Exclusion of participants

Tuberculosis can also infect parts of the body other than the lungs (see section 3.1.2). All twelve of these MRC trials were assessing treatment for pulmonary TB and, for this reason, a common exclusion criteria was presence of extra-pulmonary TB or any other serious disease. Participants found to have such disease were usually not enrolled into these trials. Despite this, a handful of participants have been identified as having been enrolled with

extra-pulmonary TB or other serious non-TB disease and these nine have been excluded.

Enrolment into a TB trial was on the basis of between one and three positive sputum smears. Treatment was started immediately, and this diagnosis was later confirmed using the more sensitive sputum culture. A small number of individuals who were diagnosed with TB on the basis of the sputum smear results were then found to be culture negative and disease-free. It was common in these trials for such individuals to have their treatment discontinued and be withdrawn from the study. An inclusion criteria for these trials was therefore a positive smear and a positive culture. Due to the delay in receiving culture results, individuals could have had several months of treatment, some even completing their six month course before being withdrawn. Since such individuals were most likely disease free and were intentionally withdrawn by the investigators, those with negative culture results at baseline were excluded from the final analysis. 324 such trial participants were excluded.

For some unknown reason, the number of participants allocated to one arm of the 2nd Hong Kong Study was only one fifth of the number allocated to each of the other three arms (22 trial participants compared to 118, 123 and 129 participants in the other three arms). This arm was therefore excluded from the analysis.

One patient enrolled into Study X was found to have already been enrolled in Study U (these two trials evidently overlapped in some sites) and was therefore excluded from Study X.

This resulted in a total of 6974 trial participants included in the analysis.

4.4.3 Patterns of Missing Data

Each of the trials used for this analysis were conducted by the British MRC over a period of around twenty years. Many of the same researchers, clinicians, lab technicians, microbiologists and statisticians were involved and the trials had very similar laboratory and clinical protocols and were therefore conducted in very much the same way. Most of the data elements recorded for each individual were the same across trials, but some were measured in only a subset of trials:

- *Weight*. The dosage of some drugs is dependent on a patient's weight. A heavier patient will be given a slightly larger dose than a lighter patient.

For this reason, weight is measured at randomisation and a trial partic-
ipant's treatment regimen tailored to this. Weight is also indicative of
the general health of a trial participant, underweight participants have
been found to have a greater chance of treatment failure (e.g. Benator
et al. (2002), see section 3.5.1). Weight has therefore been included as a
variable entered onto the database and was available on treatment cards
from all but one of the trials. It is clear from the published report of
the results of the trial where weights were not available (the third Hong
Kong study, Hong Kong Chest Service/British Medical Research Coun-
cil (1981)) that dosages were adjusted for trial participants' weight, but
an individual's weight does not appear to have been recorded on their
treatment card. The weight must have been recorded elsewhere and is
no longer available.

Trial	Month 0	Month 4	Month 6
1. Study R	✓		
2. Study T	✓		
3. Study U	✓		
4. Study X	✓	✓	
5. Study Y	✓		✓[†]
6. Tanzania	✓[†]		
7. Hong Kong 1	✓		✓[*]
8. Hong Kong 2	✓		
9. Hong Kong 3	✓		
10. Hong Kong 4	✓[†]		
11. Singapore 1	✓	✓[†]	✓[†]
12. Singapore 3			

[†]Unknown categorical scale used.
[*]Only extent of cavitation and not extent of disease was recorded at month 6 in this trial.

Table 4.4: Timings of assessments of radiographic extent of disease and cavi-
tation.

- *Radiographic extent of cavitation and disease.* Cavitation and extent of dis-
  ease were assessed from chest x-rays at various timepoints during the
  treatment period across the twelve trials. These measures were mostly
  assessed on the same standard categorical scale: nil, trivial, slight, lim-
  ited, moderate, extensive, gross. In some trials, this scale was not used
  and a different system of coding (apparently a 26-point scale from A to
  Z) was used. There is no documentation remaining of how this second

114

scale corresponds to the first. Contact has been made with a number of those involved in these trials, but there remains no explanation of how to interpret this system of grading. Until an explanation is found, extent of disease and cavitation will be considered as missing for these trials, see Table 4.4 on the preceding page. In addition to this, no radiographic measurements were recorded for the third Singapore trial at baseline or at any other point during treatment.

- *Race.* Studies have suggested that TB patients from East Asia respond slightly differently to some regimens compared to those from Africa (see Fox et al. (1999)). It is not clear whether it is geographical or genetic differences that contribute to these differences. Race (Chinese, Indian, Malay or Eurasian) was only recorded in the two Singapore trials and so will be of limited use in this analysis, but has been entered and is available on the database.

- *Isoniazid and Streptomycin Resistance.* Resistance to isoniazid or streptomycin based on cultures done before the start of treatment was marked on a treatment card with a green **H** or a green **S**, or a green **HS** for resistance to both. Susceptibility was not marked and so individuals with missing baseline sensitivity tests were indistinguishable from those who were found to be fully drug susceptible. All individuals not marked as resistant to isoniazid or streptomycin are assumed to be fully susceptible and therefore these two variables are non-missing.

- *Rifampicin Resistance.* One individual on the Tanzanian study and one on the third Singapore study were marked as having Rifampicin resistance. Rifampicin was developed later that the other anti-TB drugs and therefore did not see widespread use in any of these regions during the trial periods. Drug resistance develops as a result of poor treatment compliance and therefore cannot occur until that drug has been introduced into clinical practice. For this reason, in most of these trials, there was no procedure written into the trial protocol for identifying rifampicin resistance. It was therefore decided that these two individuals were the result of laboratory errors and the entire population was taken to be rifampicin-susceptible.

4.4.4 Baseline Summary Tables

After the exclusions detailed in the previous section, 6974 individuals allocated to 49 treatment arms across 12 trials were included in the data to be used for the analysis. Table 4.5 gives the breakdown by trial. Duration of follow-up is given in months after the end of treatment. East African studies had centres in Kenya, Tanzania, Uganda and Zambia.

	Location of Centres	Year of Start	Duration of Follow-up	Regimens Included	Patients Randomised and Included
1.	East Africa	1970	24 months	4	761
2.	East Africa	1972	24 months	4	902
3.	East Africa	1974	24 months	4	421
4.	East Africa	1976	24 months	5	310
5.	East Africa	1978	24 months	3	533
6.	Tanzania	1979	24 months	2	296
7.	Hong Kong	1972	24 months	3	246
8.	Hong Kong	1974	18 months	3	369
9.	Hong Kong	1977	18 months	5	1142
10.	Hong Kong	1979	18 months	8	1489
11.	Singapore	1973	24 months	2	198
12.	Singapore	1983	24 months	6	307
			Total	49	6974

Table 4.5: Summary of data for final analysis.

Trial	N	Weight* /kg	Missing§	Age* /years	Missing§	Sex†	Missing§	Resistance† Isoniazid	Streptomycin
1. Study R	761	49 (10)	0.0%	30 (17)	0.0%	520 (68%)	0.0%	65 (9%)	16 (2%)
2. Study T	902	48 (9)	0.9%	32 (19)	1.2%	566 (63%)	0.9%	55 (6%)	36 (4%)
3. Study U	421	47 (10)	0.0%	30 (15)	0.0%	268 (64%)	0.0%	28 (7%)	17 (4%)
4. Study X	310	48 (8)	0.0%	34 (19)	0.0%	197 (64%)	0.0%	25 (8%)	4 (1%)
5. Study Y	533	47 (10)	0.8%	30 (17)	0.9%	339 (64%)	0.4%	28 (5%)	9 (2%)
6. Tanzania	296	49 (12)	0.7%	34 (20)	1.4%	208 (71%)	0.3%	24 (8%)	11 (4%)
7. Hong Kong 1	246	46 (9)	0.4%	37 (27)	0.0%	175 (71%)	0.4%	27 (11%)	33 (13%)
8. Hong Kong 2	369	47 (9)	0.3%	36 (34)	0.0%	288 (78%)	0.0%	46 (12%)	48 (13%)
9. Hong Kong 3	1142	N/A		31 (30)	0.0%	824 (72%)	0.0%	89 (8%)	104 (9%)
10. Hong Kong 4	1489	48 (10)	0.5%	32 (27)	0.1%	1061 (71%)	0.1%	134 (9%)	147 (10%)
11. Singapore 1	198	45 (10)	0.0%	43 (25)	0.0%	129 (65%)	0.0%	4 (2%)	9 (5%)
12. Singapore 3	397	49 (13)	0.0%	38 (26)	0.0%	199 (65%)	0.0%	6 (2%)	9 (3%)
Total	6974	48 (10)	0.4%	33 (24)	0.3%	4774 (69%)	0.2%	531 (8%)	443 (6%)

*Median (Inter-quartile range) rounded to nearest integer.
†Number of male participants (percentage of total non-missing).
‡Number with resistance (percentage of total).
§Percentage of total with missing values.

Table 4.6: Summary of baseline characteristics.

		Lab Results				Radiographic Readings			
		Culture*		Smear†		Cavitation‡		Disease‡	
Trial	N		Missing§		Missing§		Missing§		Missing§
1. Study R	761	733 (98%)	1.5%	279 (37%)	0.7%	373 (59%)	16.8%	324 (51%)	16.0%
2. Study T	902	854 (96%)	1.7%	217 (24%)	1.0%	467 (60%)	13.3%	451 (52%)	4.0%
3. Study U	421	406 (96%)	0.0%	157 (37%)	0.0%	242 (59%)	3.1%	200 (49%)	3.1%
4. Study X	310	299 (97%)	0.7%	96 (31%)	0.3%	48 (17%)	10.0%	99 (35%)	9.0%
5. Study Y	533	509 (97%)	1.8%	241 (46%)	1.3%	56 (12%)	9.2%	220 (45%)	9.0%
6. Tanzania	296	264 (91%)	2.0%	47 (16%)	1.7%	N/A			
7. Hong Kong 1	246	220 (89%)	0.0%	129 (52%)	0.0%	28 (11%)	0.0%	51 (21%)	0.0%
8. Hong Kong 2	369	343 (93%)	0.0%	125 (34%)	0.0%	48 (13%)	0.5%	67 (18%)	0.3%
9. Hong Kong 3	1142	1097 (96%)	0.1%	298 (26%)	0.0%	18 (2%)	0.2%	126 (11%)	0.1%
10. Hong Kong 4	1489	1428 (97%)	1.1%	503 (34%)	0.4%	N/A			
11. Singapore 1	198	192 (97%)	0.5%	123 (62%)	0.0%	4 (2%)	0.0%	49 (25%)	0.0%
12. Singapore 3	307	302 (98%)	0.0%	161 (52%)	0.0%	N/A			
Total	6974	6647 (96%)	0.9%	2376 (34%)	0.5%	1284 (28%)	7.1%	1587 (34%)	5.1%

*Number heavily positive, more than 20 colonies (percentage of total non-missing).
†Number graded as *Heavy* (percentage of total non-missing).
‡Number graded as *Extensive* or *Gross* (percentage of total non-missing).
§Percentage of total with missing values.

Table 4.7: Assessments of microbiology and disease severity on chest x-ray at baseline.

118
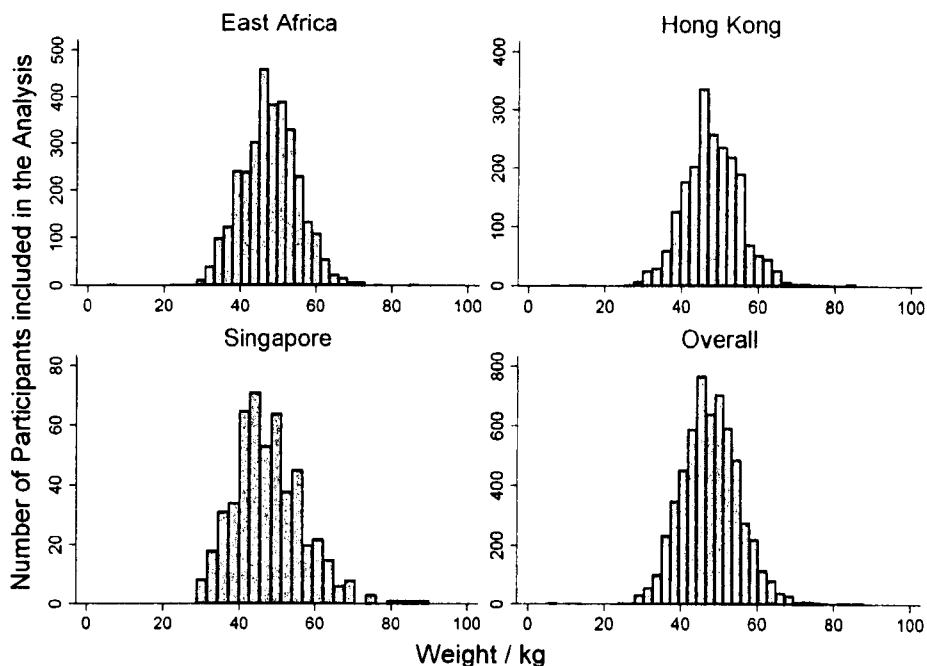
Figure 4.1: Histograms of weight by geographical region.

Table 4.6 on page 117 summarises the baseline characteristics of the individuals included in the analysis data. All were assessed and recorded at enrolment prior to initiation of treatment. The distribution of weights does not differ substantially between trials and is characterised across geographical regions by the Gaussian bell-shaped curve as shown in Figure 4.1. These bell-shaped curves are representative of the histograms of weight within individual trials in that geographical region.

A larger proportion of streptomycin resistance was seen in the Hong Kong trials with very little drug resistance seen in the Singapore trials (see 4.2 on the following page). There is no evidence of increased resistance over time as might be expected with the increased availability of these drugs over time.

There were a larger number of males enrolled and this is seen across all trials with the proportion of males slightly greater in the Hong Kong trials. The age distribution differs between geographical regions as shown in the population pyramids in Figure 4.3 on page 121 (each typical of the population pyramids of the individual trials from those regions). A larger number of males over fifty and fewer individuals of both sexes in their twenties and
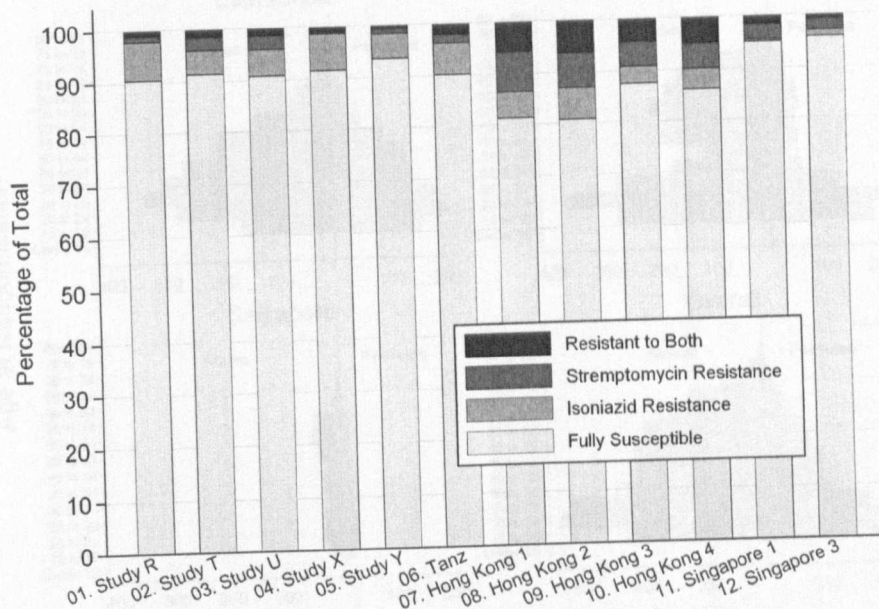
Figure 4.2: Drug resistance patterns by trial.

thirties were enrolled in Hong Kong trials as compared to the East African trials. It is likely that this is a feature of the demographics in each of these populations from which TB patients are drawn rather than a difference in the actual process of enrolment. There are comparatively too few individuals from the Singapore trials to draw any conclusions from this region.

Table 4.7 on page 118 (and figure 4.4) summarise the radiographic extent of cavitation and disease assessed before the start of treatment, and show the severity of cavitation and disease varies between trials. It appears that cavitation and disease was more serious in the East African trials compared to the trials conducted in East Asia. Since the trials were conducted under the same sponsor of the MRC and used very similar trial protocols, it seems likely that these differences are due to real difference in severity of disease of participants enrolled in the studies rather than differences in the grading. It is also clear that the differences are not limited to geographical region, but there are differences in the severity of disease and cavitation between trials in the same region. Figure 4.5 on page 122 shows the distributions of sputum smear and culture gradings at enrolment.
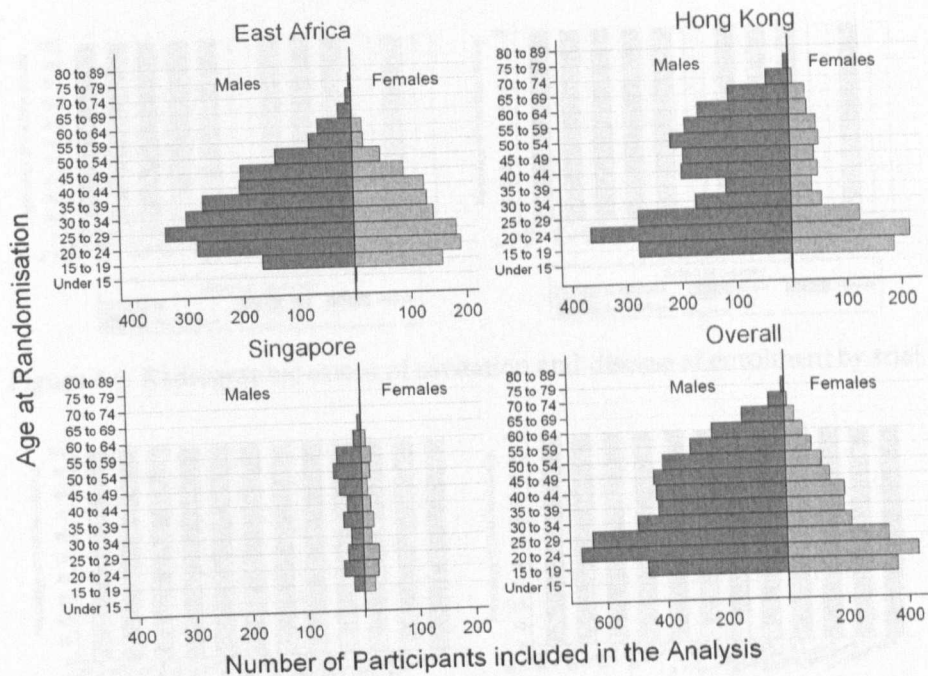
Figure 4.3: Distribution of age and sex by geographical region.

4.4.5 Longitudinal Data During Treatment

Table A.1 on page 248 shows the total numbers culture positive and culture negative at each month during treatment in each geographical region. Figure 4.6 on page 123 shows the proportions culture positive (among those with a non-missing result) at each month during treatment by geographical region. Overall, only 57% of participants with a non-missing culture result at month 1 had a positive culture and only 21% still had a positive culture at month 2. Patients in East Africa tended to culture convert later than those in Hong Kong or Singapore. 70% of participants still had a positive culture at month 1 in East Africa compared to 46% and 51% in Hong Kong and Singapore respectively. 29% of participants still had a positive culture at month 2 in East Africa compared to 15% and 7% in Hong Kong and Singapore respectively. This difference in culture conversion rates between East Africa and East Asia is likely to have an impact on the benefit of cultures during treatment as prognostic and surrogate markers and so possible differences between geographical regions will be explored in analyses in later chapters.

121
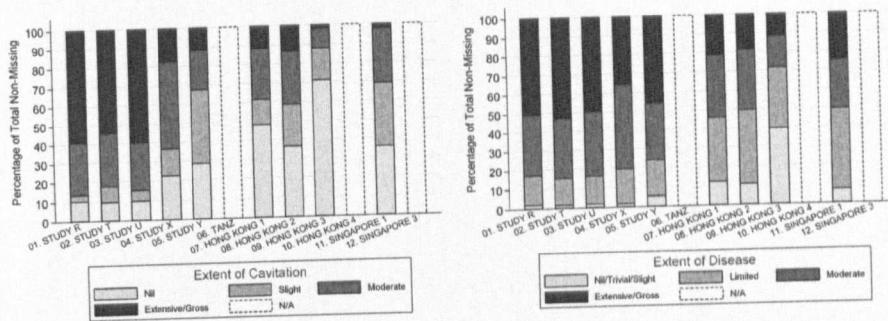
Figure 4.4: Radiographic extent of cavitation and disease at enrolment by trial.
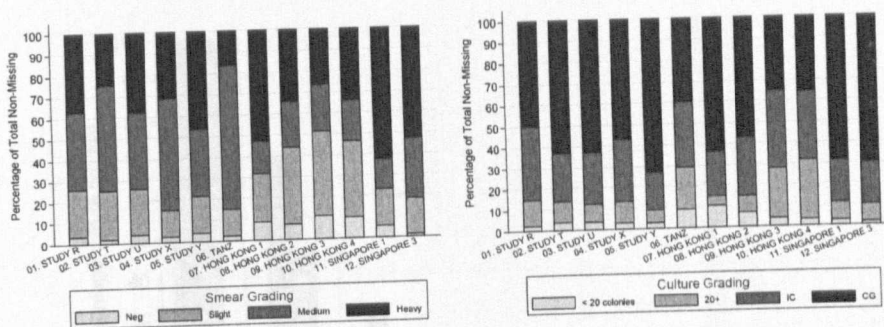
Figure 4.5: Sputum smear and culture grading at enrolment.

4.5 Summary Data Analysis

4.5.1 Summary of Outcomes

Of the total number of participants included in the analysis, just under 10% had a poor outcome. Proportions of poor outcomes ranged from 3% to 26% across the twelve trials. Table 4.8 shows the breakdown of events by trial. Overall, only 7% of participants had a missing outcome.
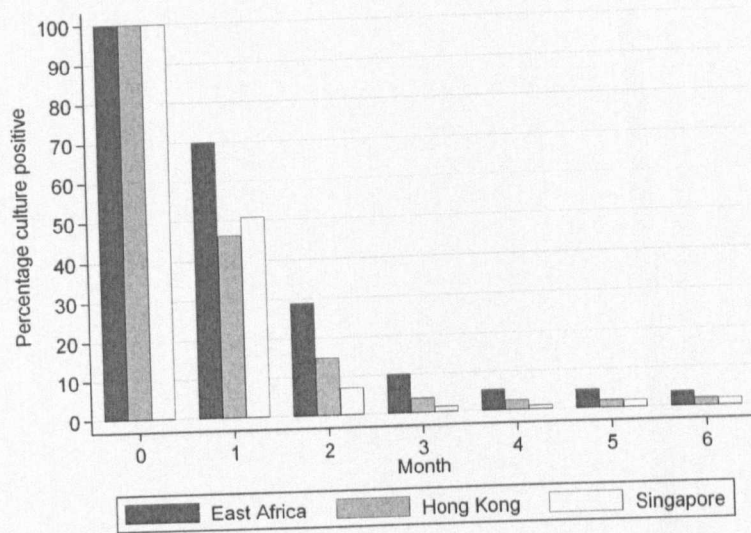
Figure 4.6: Proportions culture positive at each month during treatment by geographical region.

Trial	Total	FAIR OUTCOME			POOR OUTCOME				MISSING EVENT		
		Non-TB Death N	Long-term Cure N	Total N (%)	Treatment Failure	Recurrence	TB Death	Total N (%)	Default	Early Death	Total N (%)
1. Study R	761	6	564	570 (75%)	17	100	4	121 (16%)	42	28	70 (9%)
2. Study T	902	8	746	754 (84%)	11	45	12	68 (8%)	55	25	80 (9%)
3. Study U	421	3	325	328 (78%)	3	67	4	74 (18%)	11	8	19 (5%)
4. Study X	310	2	252	254 (82%)	5	17	2	24 (8%)	26	6	32 (10%)
5. Study Y	533	2	416	418 (78%)	6	33	5	44 (8%)	58	13	71 (13%)
6. Tanz.	296	2	232	234 (79%)	5	26	2	33 (11%)	14	15	29 (10%)
7. HK 1	246	1	165	166 (67%)	10	52	1	63 (26%)	15	2	17 (7%)
8. HK 2	369	0	299	299 (81%)	1	54	1	56 (15%)	11	3	14 (4%)
9. HK 3	1142	12	999	1011 (89%)	8	41	12	61 (5%)	51	19	70 (6%)
10. HK 4	1489	3	1281	1284 (86%)	11	113	4	128 (9%)	69	8	77 (5%)
11. Sing. 1	198	3	182	185 (93%)	1	3	1	5 (3%)	4	4	8 (4%)
12. Sing. 3	307	3	277	280 (91%)	3	12	4	19 (6%)	7	1	8 (3%)
Total	6974	45	5738	5783 (83%)	81	563	52	696 (10%)	363	132	495 (7%)

Table 4.8: Summary of participant endpoints by trial.

4.5.2 Analysis of Baseline Risk Factors

4.5.2.1 Time-to-event vs. Binary Outcome

As described in Table 4.3 on page 110, individuals can be classified as either having a fair or a poor outcome (those with a missing outcome will be excluded from subsequent analyses) and the number of months to outcome is also recorded in each case. One approach to analysing these data is to consider the time to the event and perform a survival analysis.

In this dataset, a large number of individuals (5307, 76% of the total) completed follow-up with no recurrence of disease and were classified as having a fair outcome. Even in the first Hong Kong study where no treatment arm included rifampicin, all regimens were highly effective and the number of these with a poor outcome was very few. This means that the time to the event was unobserved for 76% of the individuals and some standard survival analysis techniques (such as estimating the median survival time without extrapolation) are not possible (since the survivor function never crosses the 0.5 mark). It appears therefore that the analysis population contains a large proportion of *immunes*, or *long-term survivors*—those individuals who will never have a poor outcome. Maller and Zhou (1996) give a number of different methods for analysing survival data with long-term survivors. These methods involve using mixture distributions modelling the hazard of survival and the probability of being a long-term survivor separately. This would add additional complexity to any time-to-event analysis.

In a TB trial, the endpoint of interest is a poor outcome of treatment—either during or some time after completion of therapy. This is in contrast to an oncology or an HIV trial where the endpoint is often death and a time-to-event analysis is relevant. Under the WHO definitions of treatment outcomes (World Health Organization, 2008), a TB patient classified as having a fair outcome at the end of follow-up would be classified as *cured*. A TB treatment is deemed to be beneficial if it reduces the number of poor outcomes rather than merely delaying the onset of recurrence.

Therefore, while a time-to-event analysis might be the most natural approach, analyses considering poor outcome as a binary endpoint are clearly of great value. A surrogate endpoint that can be a substitute for the binary endpoint of poor outcome will still be extremely useful even if it proves to be a poor surrogate for the time to that outcome. The analysis considering the endpoint as a binary variable will be much less complex than that considering

the endpoint as a time to failure. This means that, while a surrogate found for poor outcome will be less comprehensive than that found for the time to poor outcome, such a surrogate may be easier to find due to fewer restrictions and the analysis will lead to a clearer conclusion that will be easier to interpret. The analyses in subsequent chapters will therefore be focused on considering poor outcome as a binary variable and analysed as such. These models will be simpler and more powerful.

Covariate		Odds Ratio	95% Confidence Interval
Continuous Covariates			
Weight (per 10 kg)		0.86	(0.77, 0.97)
Age (per 10 years)		1.29	(1.22, 1.36)
Categorical Covariates			
Sex	Female	1.00	
	Male	1.78	(1.47, 2.15)
Isoniazid Resistance	Susceptible	1.00	
	Resistant	3.70	(2.96, 4.62)
Streptomycin Resistance	Susceptible	1.00	
	Resistant	2.45	(1.88, 3.18)
Sputum Smear	Negative	1.00	
	Slight	0.89	(0.56, 1.41)
	Medium	1.74	(1.11, 2.72)
	Heavy	2.90	(1.88, 4.48)
Sputum Culture	1-19 Colonies	1.00	
	20+ Colonies	0.89	(0.49, 1.60)
	Innumerable Colonies	1.53	(0.89, 2.63)
	Confluent Growth	2.73	(1.61, 4.61)
Extent of Cavitation	Nil	1.00	
	Slight	1.77	(1.21, 2.59)
	Moderate	3.04	(2.18, 4.23)
	Extensive/Gross	6.44	(4.54, 9.15)
Extent of Disease	Nil/Trivial/Slight	1.00	
	Limited	3.37	(1.84, 6.19)
	Moderate	5.13	(2.81, 9.35)
	Extensive/Gross	9.17	(5.04, 16.70)

Table 4.9: Baseline characteristics as risk factors for a poor outcome. Odds ratios are adjusted for differences between trial.

4.5.2.2 Summary Tables

Table 4.9 on the previous page shows the odds ratios of treatment failure and associated 95% confidence intervals for each of the baseline covariates. All odds ratios are estimated from a mixed effects models including the trial effect as a random effect. Participants who are heavier, younger or female are all more likely to have a fair outcome. As expected, baseline resistance to isoniazid or streptomycin were risk factors for treatment failure with a comparatively higher odds ratio for resistance to isoniazid. More serious grading of radiographic extent of disease and cavitation were strong predictors of treatment failure, more so than sputum lab results. Participants with extensive or gross extent of disease had odds of treatment failure nearly nine times those of participants with nil, trivial or slight disease. Participants with extensive or gross cavitation had odds of treatment failure nearly six times those of participants with nil cavitation.
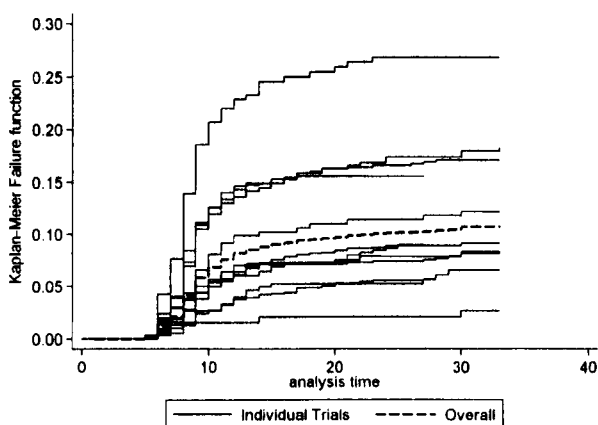
Figure 4.7: Estimate of Kaplan-Meier failure function by trial.

4.5.2.3 Analysis of Survival Time

Figure 4.7 shows the Kaplan-Meier estimate of the failure function and figure 4.8 on the next page the hazard function by trial. The failure function is $1 - S(t)$ where $S(t)$ is the Kaplan-Meier estimate of the survivor function. Figure 4.8 suggests that the assumption of proportional hazards across trials is reasonable. The hazard is not monotonic (and therefore does not follow a

127
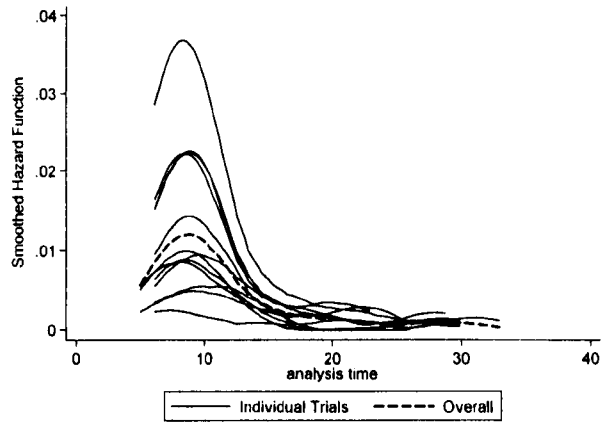
Figure 4.8: Estimate of hazard function by trial.

distribution from the Weibul family), but is unimodal. The hazard increased steeply to a peak around month 9 before decreasing to a low constant from around month 18 onwards. This behaviour is consistent with the model of treatment failure being a mixture of true relapse (almost all occurring within twelve months of follow-up) and exogenous reinfection being observed as a background constant instantaneous hazard (see section 3.4.1).

Chapter 5

Culture Positivity as a Prognostic Marker for Poor Outcome

5.1 Introduction

Having described and performed some exploratory analyses of the data in Chapter 4, the first step in evaluating a marker as a surrogate is to establish its value as a prognostic marker. The second of the Prentice Criteria is that the surrogate markers should have some 'prognostic implication' for the true endpoint (see section 2.5 on page 38). A prognostic marker is one that can be used for predicting outcome and therefore requires less strict criteria than for a surrogate marker, which must capture the treatment effect on the true endpoint. In the context of this thesis, the prognostic importance of culture results as predictors of poor outcome must first be established using the trial data before they can be evaluated as surrogates.

Validation of a prognostic marker involves looking at the association with and prediction of the binary endpoint of poor outcome across treatment arms controlling for the trial effect. In this chapter, culture positivity during treatment will be explored as a prognostic marker pooling individuals across treatment arms before validating this marker as a surrogate in Chapter 6.

As described in the previous chapter, sputum culture results were recorded

on a semi-continuous scale ranging from negative to CG (confluent growth). Historically in TB prognostic studies (e.g., Aber and Nunn, 1978; Mitchison, 1996), the culture result was only considered as a binary variable—positive or negative. This approach leads to a loss of information, but is easier to use as a marker in clinical practice and simplifies the evaluation process resulting in a clearer interpretation. A positive culture indicates presence of mycobacteria in the lungs while a negative culture indicates absence of detectable mycobacteria in the lungs. The number of colonies is thought to give an indication of the bacillary load in the patient's lungs and is therefore more commonly given the unit of *colony forming units* or *CFUs*. In this chapter, culture results will be considered both as ordered categorical and as binary exploring whether the loss of information in the latter is large.

One way to incorporate the number of CFUs into a binary variable is to vary the *point of dichotomy* of the binary variable. A point of dichotomy of 11, for example, means classifying a CFU count of less than 11 a 'negative' culture and classifying a CFU count of 11 or more as a 'positive' culture with this binary variable being assessed as a marker. It might be suggested that a low CFU count is no more predictive of poor outcome to treatment than a negative culture (indeed, isolated scantily-positive cultures are not indicative of failure; see section 3.5.3.2) and that they should be grouped together for the prognostic marker. This may reduce the *False Positive Fraction (FPF)* of the marker without substantially reducing the *True Positive Fraction (TPF)*, see section 2.4.2 for definitions. The effect of varying the point of dichotomy for the binary marker is also explored in this chapter.

As described in Chapter 4, the data used in this thesis are from twelve TB clinical trials conducted across East Africa and East Asia in the 1970s and 1980s. A total number of 6974 trial participant allocated to 49 treatment regimens across the twelve trials were included.

5.2 Exploratory Methods

5.2.1 Introduction

In section 5.3, the *Receiver Operating Characteristic curve (ROC curve)* is used to evaluate cultures at each of months 1, 2, 3 and 4 from the start of treatment as prognostic markers for poor outcome. In this section, simpler, exploratory methods are applied to the data to explore the value of culture results during

treatment as prognostic markers for a poor outcome to treatment.

This section is divided into two parts. In the first (section 5.2.2), culture results are assumed to be categorical with 7 levels. In the second (section 5.2.3), culture results are considered as binary, exploring the variation in the point of dichotomy which leads naturally into section 5.3 bringing these results together in the ROC curve.

(a) Month 1   (b) Month 2

(c) Month 3   (d) Month 4
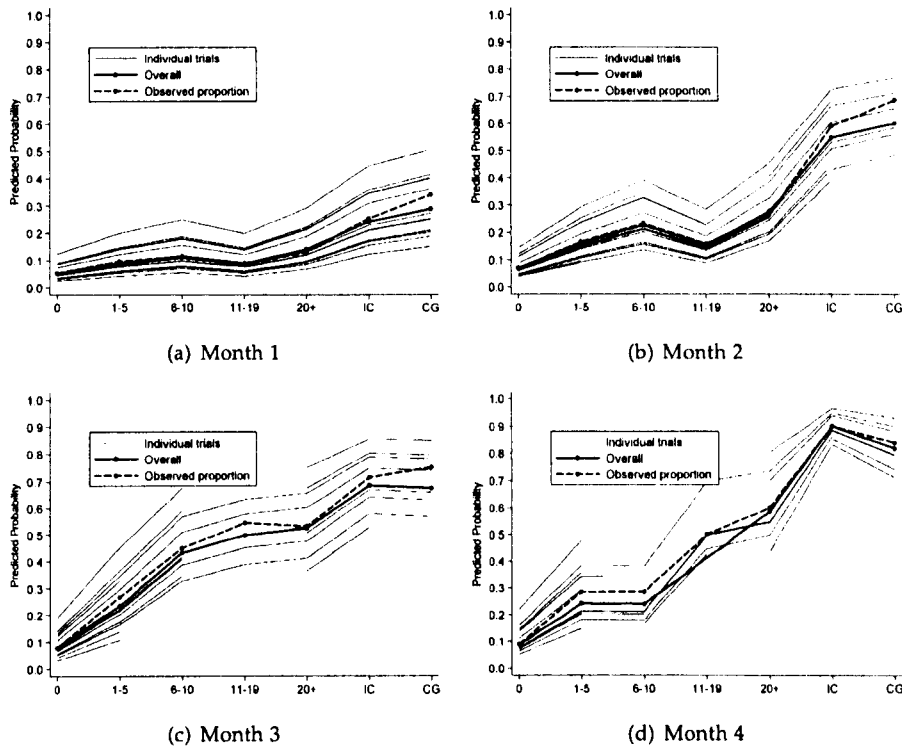
Figure 5.1: Predicted probabilities of failure for culture results at months 1 to 4.

5.2.2 Cultures as Categorical

5.2.2.1 Methods

In this section, culture results are considered as categorical with 7 levels. These levels are: Negative, 1-5, 6-10, 11-19, 20+ (20-100 colony forming units), IC (Innumerable Colonies) and CG (Confluent Growth). It is expected that the probability of a poor outcome will vary across trials and therefore the prob-

131

ability of a poor outcome was modelled conditional on the culture results using logistic regression, with clustering within trial being accounted for by including a random intercept varying across trials. The model is as follows:

$$\text{logit}\, P(T_{ik} = 1|S_{ik}) = \alpha + \sum_{l=1}^{l=7} \beta_l I_l(S_{ik}).$$ (5.1)

for individual k in trial i, where T_{ik} is the is the binary outcome variable, S_{ik} is the culture result taking values $s_1, s_2, \ldots, s_7$ and $I_l(S_{ik})$ is the indicator function where:

$$I_l(S_{ik}) = \begin{cases} 1 & \text{if } S_{ik} = s_l \\ 0 & \text{if } S_{ik} \neq s_l \end{cases} , \quad l = 1, \ldots, 7.$$ (5.2)

The clustering within trials was included in the model by including a random intercept, u_i, letting $\alpha = a + u_i$ where $u_i \sim N(0, \sigma^2)$. A random slope (adding random effects to each of the β_l) has not been included since this makes the model unnecessarily complex leading to instability and problems with convergence.

5.2.2.2 Results

An ideal prognostic marker would be one with a very low (zero for a perfect marker) predicted probability of poor outcome for negative cultures and a very high (unity for a perfect marker) predicted probability for positive cultures. A prognostic marker that could be used as a binary marker is one showing a sudden increase in predicted probability, where this sudden increase is at the point of dichotomy.

Figure 5.1 on the previous page shows the predicted probabilities of poor outcome for each of the different categories of culture results at months 1 to 4. The thin solid lines are the predicted probabilities for the twelve individual trials, the thick solid line is the predicted probability across trials and the dashed line shows the overall proportion of participants with a poor outcome observed for each culture result.

The thin solid lines show that there is great variability between trials, but that the average across trials is a meaningful representation of the data. It is clear that a culture result reflecting a higher bacillary load at a later month is more predictive of poor outcome than a culture result an earlier month. This suggests that individuals with heavily positive cultures at months 3 or 4 are

highly likely to have a poor outcome to treatment.

A strong positive trend is apparent in Figures 5.1(b), 5.1(c) and 5.1(d). This means that, at months 2, 3 and 4, the predicted probability of poor outcome increases with increasing levels of CFUs measured. Therefore, not only is the presence of *Mycobacteria tuberculosis* in an individual's sputum during treatment prognostic of a poor outcome, but a higher number of CFUs corresponds to a greater prognostic value. This is seen across all of months 2 to 4 to a lesser or greater extent.

The slope of the line of the overall predicted probability in Figure 5.1(a) (month 1) is shallow showing that a positive culture is not a much better predictor than a negative culture. An increase is seen after 11-19 suggesting that a CFU of less than 20 at month 1 is not prognostic of poor outcome, but a CFU of 20+, IC or CG is more prognostic of poor outcome. However, the predicted probabilities are small for results of 20+, IC or CG at month 1.

For the month 2 culture, the line of the average across trials is steeper from 20+ colonies suggesting that there is a difference in the prognostic value between heavily positive (more than 100 colonies) and more scantily positive cultures. This suggests that the binary variable dichotomised at 100 colonies could be a useful prognostic marker, but the predicted probability is not insignificant for cultures of less than 100 colonies and therefore grouping these with negative cultures is questionable.

The lines for months 3 and 4 are steeper from the start with the same small predicted probability of poor outcome for a negative culture. The predicted probability of a poor outcome differs greatly between a negative culture and the various levels of a positive culture and this suggests that cultures at months 3 and 4 are good prognostic markers. It was seen in Chapter 4 that there are few individuals with positive cultures at months 3 and 4 (7% and 4% overall respectively, see table A.1 on page 248), and fewer with heavily positive cultures, making the use of these as prognostic markers impractical. This aspect, the actual numbers of participants with a particular culture result, is not captured by the predicted probability and it is therefore important to look at other measures of prognostic value. Other measures are considered below.

It is interesting to note that the predicted probabilities for between 1 and 5 colonies across all four of these graphs is greater than that for a negative culture. This means that individuals with scanty colonies at any month are more likely to fail treatment than those with no positive cultures, even if this difference is small (particularly at month 1). Whilst it has been shown in

previous studies that isolated scantily-positive cultures during follow-up are not indicative of a poor outcome (see section 3.5.3.2), these data suggest that patients with scantily-positive cultures during treatment might be more likely to have a poor outcome to treatment than those with a negative culture.

5.2.3 Cultures as Binary

5.2.3.1 Methods

In this section, the culture result is considered as a binary variable exploring the effect of varying the point of dichotomy. Where the point of dichotomy is given as 11, for example, this means that all individuals with 11 or more colonies are counted as being 'culture positive' and those with 10 colonies or less as being 'culture negative'.

As described in section 2.4, measures to evaluate binary prognostic markers can be calculated empirically from the data or estimated from a parametric model. The latter is preferable as it allows for adjusting for covariates or clustering using random effects.

In this section, four different sets of measures will be used to evaluate culture results as prognostic markers and the different results compared.

1. *Odds Ratios.* The standard measure of effect summarising the association of explanatory variables with outcome in a logistic regression model is the odds ratio. The odds ratio express the ratio of odds of a poor outcome given a positive culture against the odds given a negative culture. While popular, the odds ratio is not an appropriate measure to evaluate a prognostic marker (Pepe, 2003), but is useful as it reflects the association between the markers and a poor outcome.

2. *Pseudo-R^2 Statistics.* In linear regression, the R^2 statistic measures the proportion of variation in the response variable that is captured by the explanatory variables. This measure does not naturally follow in logistic regression, but a number of different *pseudo-R^2* measures have been proposed (see section 2.4.4) and the four most useful are used to evaluate culture results as prognostic markers.

3. *True and False Positive Fractions (TPF and FPF).* The TPF (often called the *sensitivity*) and the FPF (equal to the *specificity* subtracted from one) are the most common measures for evaluating prognostic markers. The TPF

134

is the proportion of those with a poor outcome that have a positive culture result and the FPF is the proportion of those with a fair outcome that have a positive culture result. A good marker is one with a high TPF and low FPF. Increasing the point of dichotomy leads to classifying less individuals as culture positive and results in a decreased FPF at the expense of a decreased TPF.

4. *Positive and Negative Predictive Values (PPV and NPV).* The PPV and the NPV are also common in prognostic studies. The PPV is the proportion of those with a positive culture result that have a poor outcome and the NPV is the proportion of those with a negative culture result that have a fair outcome to treatment. Unlike the TPF and the FPF, the PPV and the NPV are affected by the prevalence of the outcome that the markers are intended to predict. It is therefore important that both the TPF and FPF *and* the PPV and NPV are examined in a prognostic study.

To calculate the odds ratio, pseudo-R^2 statistics and positive and negative predictive values for poor outcome to treatment, the following model is used:

$$\text{logit}\, P(T = 1|S) = \alpha_1 + \beta_1 S. \tag{5.3}$$

Here, the probability of a poor outcome ($T = 1$) is modelled conditional on the binary culture result, S. To calculate the true and false positive fractions, the following model is used:

$$\text{logit}\, P(S = 1|T) = \alpha_2 + \beta_2 T. \tag{5.4}$$

Here, the probability of a positive culture result ($S = 1$) is modelled conditional on the treatment outcome, T. Functions of the parameters, α and β, are used to estimate these measures in each case using the methods that are described in more detail in section 2.4. In these data, patients are clustered within trials. Mixed effects logistic regression models are used to adjust for this clustering. Random effects are added to both models as follows:

$$\alpha_1 = a_1 + u_{1i}, \quad \alpha_2 = a_2 + u_{2i}, \tag{5.5}$$

$$\beta_1 = b_1 + v_{1i}, \quad \beta_2 = b_2 + v_{2i}, \tag{5.6}$$

where $u_{1i} \sim N(0, \sigma_{\beta_1}^2)$, $u_{2i} \sim N(0, \sigma_{\beta_2}^2)$, $v_{1i} \sim N(0, \sigma_{\alpha_1}^2)$ and $v_{1i} \sim N(0, \sigma_{\alpha_2}^2)$

135

for trial i. The random effects are assumed to be independent (using the *independent covariance structure*).
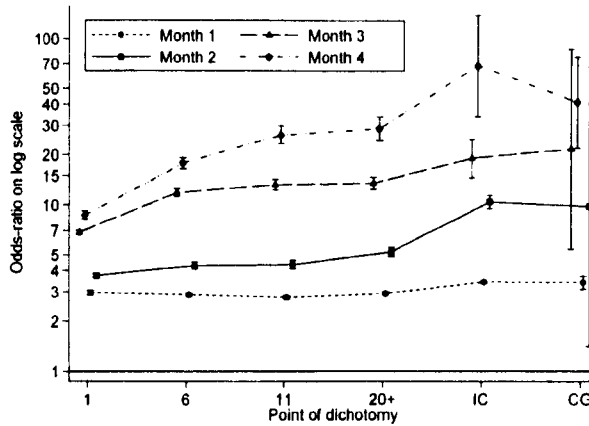
Figure 5.2: Odds ratios of poor outcome for culture results at different points of dichotomy for months 1 to 4 plotted with 95% confidence limits. The point of dichotomy is given on the horizontal axis. Plotted points have been perturbed slightly in the horizontal for clarity of presentation.

5.2.3.2 Results

5.2.3.2.1 Odds Ratios The odds ratio of poor outcome on the log scale for different points of dichotomy is shown in Figure 5.2, plotted with 95% Wald confidence limits.

From the graph it can be noted that the estimates of the odds ratios are very large and the widths of the confidence intervals are generally very small. The large amount of data (from nearly 7000 trial participants) results in the narrow confidence intervals giving very precise estimates of the odds ratios (except when the point of dichotomy is large). As the point of dichotomy increases, the proportion of patients with a positive culture decreases and therefore the width of the confidence interval grows. The width of the confidence interval is also larger for later months for a similar reason. Very few participants have heavily positive cultures beyond month 1 and therefore the confidence intervals for a point of dichotomy of CG are particularly large for months 2, 3 and 4.

The point estimates of the log-odds ratios are greater for later months

136

and also increase with point of dichotomy, particularly for months 2, 3 and 4. Considering a point of dichotomy of 1, the odds of a poor outcome for a patient with a positive culture at month 2 is 3.8 times that for a patient with a negative culture, and this ratio increases to 8.2 for cultures at month 4 and for a point of dichotomy at CG at month 2, the odds ratio is 10.3.

Considering the traditional binary variable (a point of dichotomy at 1), the confidence intervals for each of the cultures are very narrow giving good evidence for the odds ratio of a poor outcome for the culture result at month 4 being greater than that at month 3 which in turn is greater than that at month 2 which is greater than that at month 1. No longitudinal analysis was conducted to formally compare the odds ratios at different months, these conclusions are based on estimates from separate models and so can only be approximate. Nevertheless, the differences are striking. The point estimates do separate a little as the point of dichotomy increases, but the width of the confidence intervals also increases.

Despite these wide confidence intervals, all of the 95% confidence intervals exclude an odds ratio of 1 giving evidence of varying strength that all of these markers (at each month with each point of dichotomy) are strongly associated with a poor outcome and fulfil the requirement stated at the beginning of this chapter that the marker should have 'some prognostic implication for the true endpoint'.

5.2.3.2.2 Explained Variability: Pseudo-R^2 Statistics

Figure 5.3 shows each of the four pseudo-R^2 statistics introduced in section 2.4 for the different points of dichotomy for months 1, 2, 3, and 4.

Overall, the plotted lines are largely flat or decreasing as the point of dichotomy is increased. There is some suggestion that a point of dichotomy of 6 gives marginally greater values of R^2 than at 1, particularly in Figures 5.3(b) and 5.3(d). In all four graphs, each of the four the lines representing culture results at each of months 1 to 4 were close to each other, and there is no consensus across the different pseudo-R^2 statistics as to which monthly culture result is superior with a different marker being suggested to be superior by each R^2 statistic.

Apart from the Count R^2, all of the plotted points are below the line $R^2 = 0.2$. If these pseudo-R^2 statistics are in any way equivalent to the OLS R^2 then this suggests that less than 20% of the variation is explained by the culture result despite the very high odds ratios presented in the previous sec-

(a) Modified McFadden's R^2 (R^{2l}_{McF})

(b) Efron's R^2

(c) McKelvey and Zavoina's R^2

(d) Count R^2
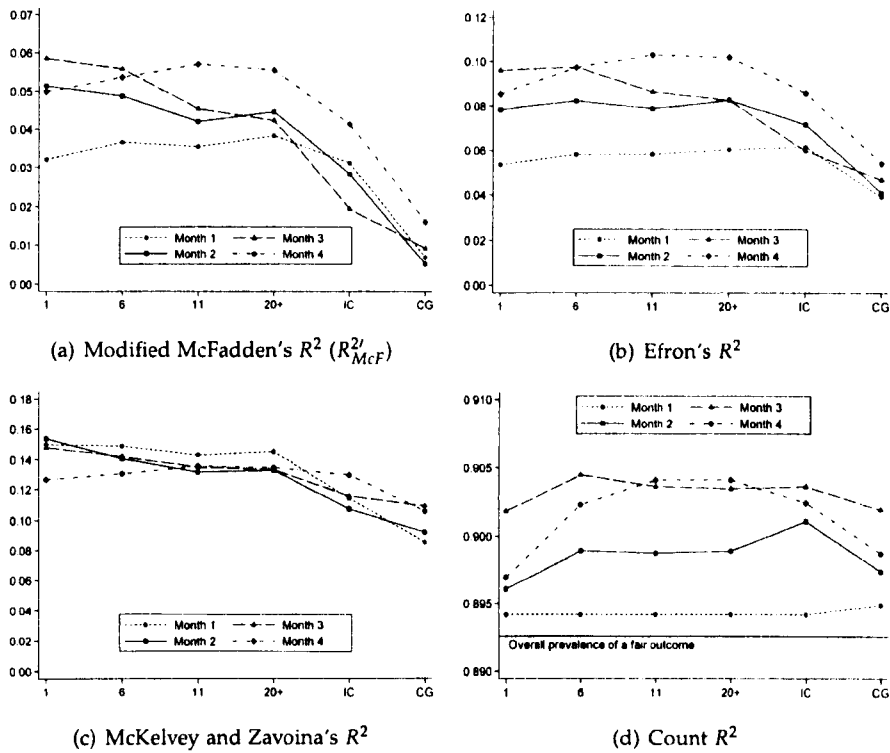
Figure 5.3: Various pseudo-R^2 assessing the explained variability of models for culture results at months 1 to 4, for different points of dichotomy.

tion. This demonstrates how important it is to consider different measures when evaluating culture results as prognostic markers. The Count R^2 values are very close to the overall prevalence of a fair outcome (89.3%), excluding those with a missing outcome. If we consider a hypothetical marker that classifies all participants as having a fair outcome, each of these markers correctly identify the outcome in only a fraction more participants than this completely uninformative hypothetical marker.

5.2.3.2.3 True and False Positive Fractions Figure 5.4 on the following page shows the *true positive fraction (TPF)* and *false positive fraction (FPF)* for different points of dichotomy at different months. All these are estimated from the parametric model (equation 5.4) which models the probability of a positive culture conditional on the treatment outcome. The number of colonies on the horizontal axis is the point of dichotomy. The thin solid lines show the

138

(a) Month 1   (b) Month 2

(c) Month 3   (d) Month 4
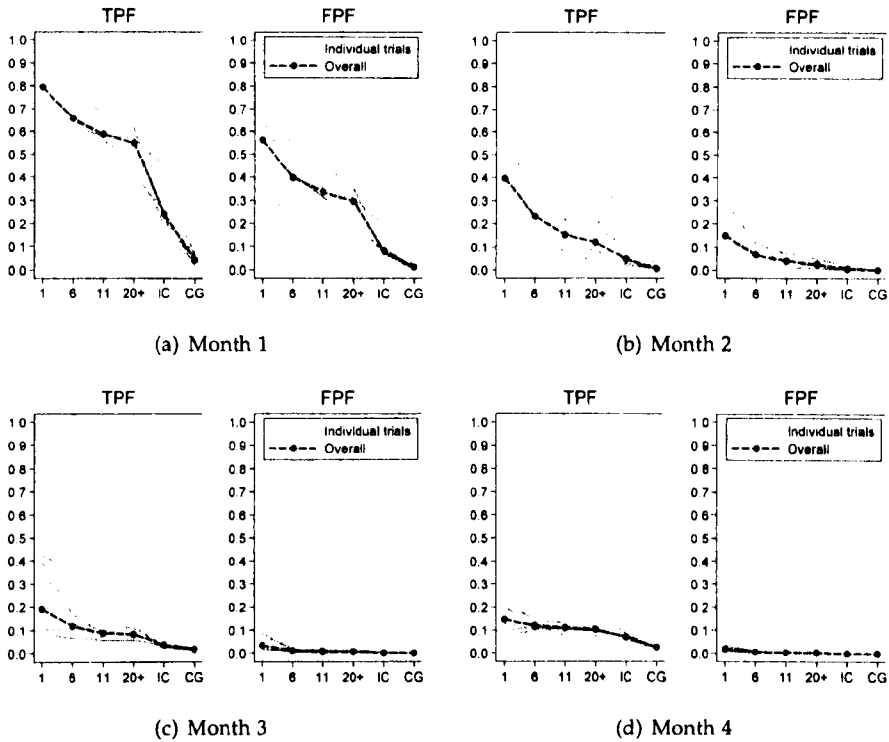
Figure 5.4: True Positive Fraction (TPF) and False Positive Fraction (FPF) for culture results for different points of dichotomy for months 1 to 4. The point of dichotomy is given on the horizontal axis.

estimates of TPF and FPF for each trial and the thick dashed line shows the overall estimates of TPF and FPF across all twelve trials.

The lines for the individual trials show that the overall curve is representative of curves for individual trials. There is greater variability across trials in TPF than in FPF, due to the considerably smaller numbers of participants with a poor outcome than with a fair outcome.

It is clear from Figures 5.4(c) and 5.4(d) that the FPF at months 3 and 4 is very low, however the TPF is unacceptably low. Considering a point of dichotomy of one, an individual with a fair outcome will almost certainly have a negative culture at month 3 or month 4 (and therefore the FPF is very low), but too few participants with a poor outcome have a positive culture. At a point of dichotomy of one, only 19% of individuals who have a poor outcome to treatment have a positive culture at month 3 and only 15% have a

139

positive culture at month 4 (see Table 5.1).

	Point of Dichotomy					
	1		20+		IC	
	TPF	FPF	TPF	FPF	TPF	FPF
Month 1	80%	56%	55%	30%	24%	9%
Month 2	40%	15%	13%	3%	5%	< 0.5%
Month 3	19%	4%	8%	1%	4%	< 0.5%
Month 4	15%	2%	10%	1%	7%	< 0.5%

Table 5.1: True Positive Fraction (TPF) and False Positive Fraction (FPF) at different months for three different points of dichotomy.

At a point of dichotomy of one, the TPF at month 2 is improved compared to that at months 3 or 4, but the FPF is also slightly increased. Of those individuals who go on to have a poor outcome, 40% have a positive culture at month 2. However, it still means that 60% of those individuals with a poor outcome will not be identified by the month 2 culture. At month 1, of those individuals that fail treatment, 80% have a positive culture at a point of dichotomy of one, but 56% that have a fair outcome also have a positive culture.

Table 5.1 also shows the TPF and FPF for a point of dichotomy at 20+ and a point of dichotomy at IC. At month 1, a point of dichotomy of 20+ yields a TPF of 55% with a moderate FPF of 30% which is more acceptable than that at a point of dichotomy of one.

5.2.3.2.4 Positive and Negative Predictive Values The *positive predictive value (PPV)* and the *negative predictive value (NPV)* are estimated from the parametric model (equation 5.3) modelling the probability of a poor outcome conditional on the culture result.

Figure 5.5 on the following page shows the PPV and NPV of culture as a binary variable for different points of dichotomy for months 1 to 4. The thin solid lines are the values of the PPV and NPV for the twelve individual trials and the thicker dashed line is the overall PPV and NPV estimated across all trials.

One additional point is plotted for each graph, the PPV and NPV for the logistic models containing the trial effect only (labelled *Null model* on the horizontal axis). From this model, the PPV is effectively an estimate of the proportion of the whole population that go on to have a poor outcome to treatment
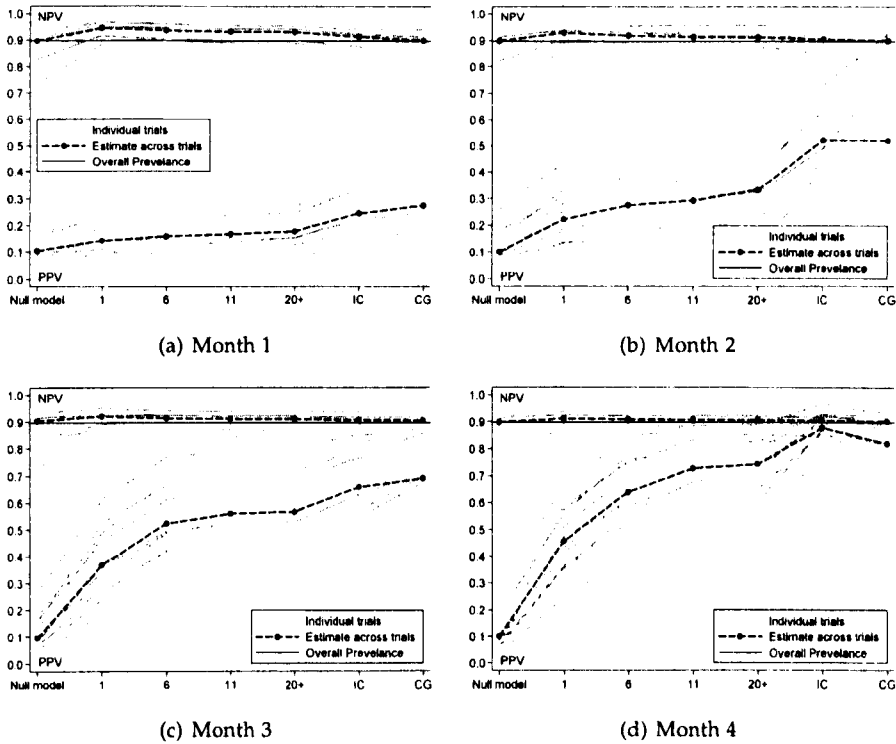
Figure 5.5: Negative Predictive Value (NPV) and Positive Predictive Value (PPV) for different points of dichotomy for months 1 to 4. The point of dichotomy is given on the horizontal axis.

and the NPV of the proportion of the whole population that go on to have a fair outcome to treatment. The PPV for this point can be interpreted as the crude diagnostic test of classifying all individuals as having a poor outcome and the NPV as the crude diagnostic test of classifying all individuals as having a fair outcome. They can therefore be considered as reference values for the PPV and the NPV. Since the PPV and NPV are affected by the prevalence of a poor outcome and must be interpreted with reference to this. The overall prevalence of a fair outcome 89.3% is also plotted as a reference line

Looking across all months and the different points of dichotomy, the NPV estimated across trials is only marginally greater than the overall proportion of fair outcomes with the highest NPV being with the point of dichotomy at 1 (also see Table 5.2 on the next page). With such a high proportion of fair outcomes, it is hard to interpret the NPV.

141

	Point of Dichotomy							
	Null Model		1		20+		IC	
	PPV	NPV	PPV	NPV	PPV	NPV	PPV	NPV
Month 1	10%	90%	14%	95%	18%	93%	24%	91%
Month 2	10%	90%	22%	93%	33%	91%	52%	90%
Month 3	10%	90%	37%	92%	57%	91%	66%	91%
Month 4	10%	90%	46%	91%	74%	91%	88%	91%

Table 5.2: Positive Predictive Value (PPV) and Negative Predictive Value (NPV) at different months for three different points of dichotomy.

The shape of the plot of the PPV is not dissimilar to the plots in Figure 5.1 which shows the predicted probability of a poor outcome conditional on the *culture lying in the specified range*. Here, a point on the graph shows the predicted probability of a poor outcome conditional on a *culture greater or equal to the corresponding point of dichotomy*.

The slope of the PPV is greater for later months. At month 1, the PPV is fairly flat showing that the culture result at month 1 is not highly predictive of poor outcome. At month 2, the PPV for a point of dichotomy of one is 22%—only 22% of those with a positive culture at month 2 go on to have a poor outcome to treatment. This compares with 37% and 46% at months 3 and 4 respectively. It is clear from Figures 5.5(c) and 5.5(d) that a highly positive culture at months 3 or 4 is highly predictive of poor outcome. Since the NPV hardly varies, on the basis of predictive values along, one would select the month 4 culture with a point of dichotomy at IC yielding at PPV of 88% and a NPV of 91%.

5.3 Receiver Operating Characteristic Curve

Figure 5.6 on the following page shows the receiver operating (ROC) curves for sputum culture results at each of months 1 to 4 during treatment. The TPF and FPF are estimated for every possible point of dichotomy and plotted separately for each month. The TPF and FPF are estimated as in section 5.2.3.2.3 (equation 5.4), modelling the probability of a positive culture given the true outcome including a random effects intercept term in the model to allow for clustering within trials. A binormal curve is fitted to the data points as described in section 2.4.3.1, and this is shown on each graph.

It is immediately clear from the graphs is the ROC curves are very short.

(a) Month 1

(b) Month 2

(c) Month 3

(d) Month 4
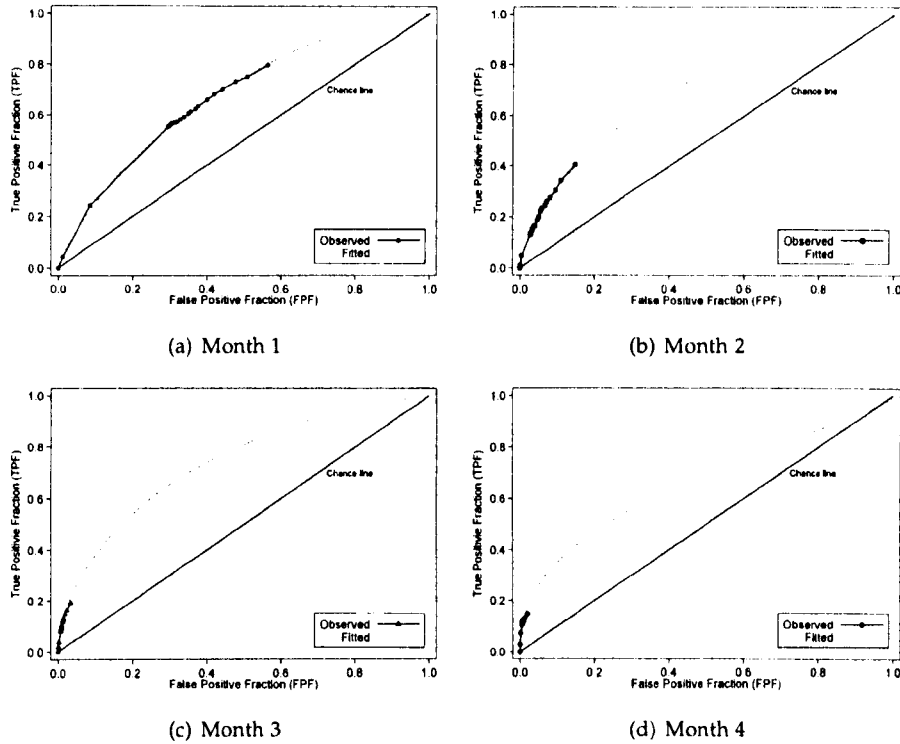
Figure 5.6: Receiver Operating Characteristic (ROC) Curve for culture results as prognostic markers at months 1 to 4.

At month 1, there is no point of dichotomy giving a TPF of greater than 0.8 and this limit drops to 0.4 at month 2 and below 0.2 for months 3 and 4. For each of these graphs, the point with the largest TPF corresponds to a point of dichotomy of 1. This indicates that a not insubstantial proportion of individuals with a poor outcome have a negative culture at month 1 with this proportion increasing steadily over months 2, 3 and 4.

The fit of the binormal curve is good for months 1, 2 and 3 ($R^2 = 0.99$ in each case). The fit is less good at month 4 ($R^2 = 0.83$) and this is suggested by the non-symmetric fitted curve actually crossing the chance line around FPF $= 0.96$ in Figure 5.6(d). All of the fitted curves at months 2 to 4 require considerable extrapolation and rely heavily on the assumption that the binormal curve is the most correct model, and therefore too much weight should not be placed on the AUC.

Table 5.3 on the following page shows the area under the fitted curve

143

Month	AUC	95% confidence interval
1	0.68	(0.66,0.70)
2	0.70	(0.64,0.75)
3	0.74	(0.60,0.82)
4	0.68	(0.54,0.82)

Table 5.3: AUC for fitted binormal curve

(AUC) with 95% bootstrap confidence intervals. The confidence intervals were bias-corrected using 1000 bootstrap repetitions (Carpenter and Bithell, 2000). Only six different points of dichotomy (1, 5, 10, 20, IC and CG) were used to calculate the confidence intervals as the bootstrap procedure for all 22 points of dichotomy was too computationally intensive. Since an uninformative marker has ROC curve equal to the chance line with AUC 0.5, the AUC lies in the range [0.5,1.0]. All of these AUCs are small with the curve lying close to the chance line. This shows that none of these markers are particularly effective prognostic markers. The AUC is the largest for month 3, but they are all very similar across the four months with wide, overlapping confidence intervals.

5.3.0.3 Summarising Culture Results across months 2 to 4

The proportion of individuals with a fair outcome to treatment that have a positive culture at month 3 or 4 (the FPF) is small, but the proportion of individuals with a poor outcome that have a positive culture (the TPF) is also small. Compared to the culture result at 2 months, the summary marker of *the heaviest culture in month 2, 3, or 4*, is likely to have a higher TPF without a considerable increase in FPF since the FPF at months 3 and 4 is so low. Figure 5.7 on the next page shows a selection of the graphs from previous sections plotted for the summary marker of the heaviest culture in months 2, 3 or 4.

Comparing the odds ratios in Figure 5.7(a) with those for the 2 month culture Figure 5.2 on page 136, the estimates are larger with the 95% confidence interval for the point of dichotomy at CG much narrower. The PPV for the heaviest culture at months 2, 3 or 4 is marginally greater than that for the 2 month culture at each point of dichotomy and the pseudo-R^2 statistics are still very low. Comparing Figure 5.7(c) on the next page with Figure 5.4(b)

(a) Odds Ratios, PPV and NPV.

(b) Pseudo-R^2 Statistics.

(c) TPF and FPF.

(d) ROC Curve with fitted binormal curve.
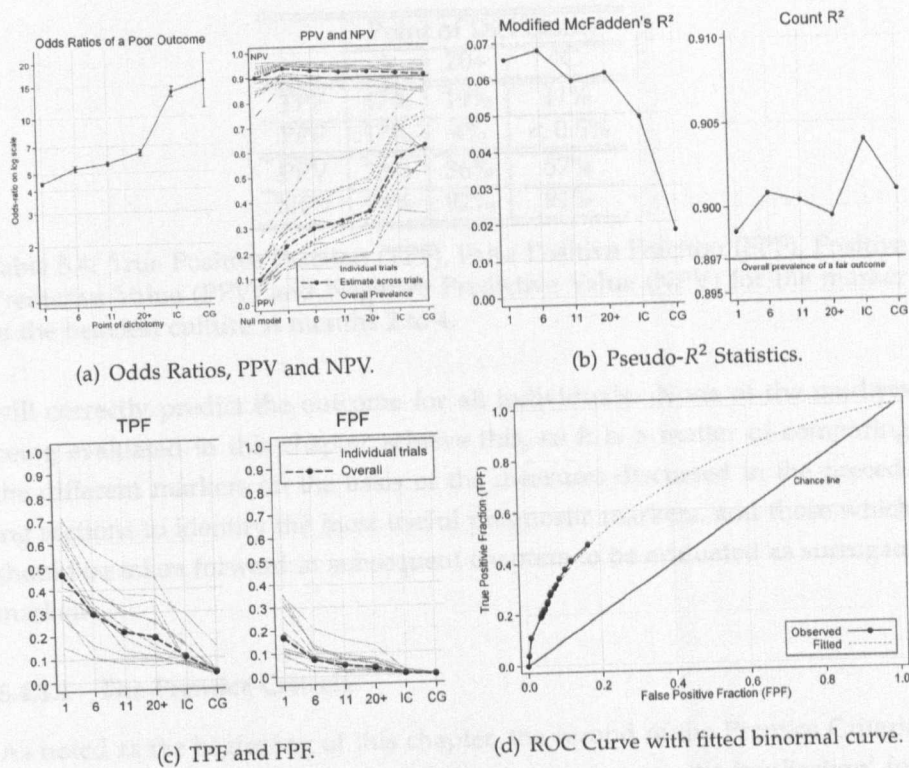
Figure 5.7: Odds ratios, PPV, NPV, TPF, FPF and pseudo-R^2 statistics for the marker of the heaviest culture at 2, 3 or 4 months.

on page 139, the shapes are very similar with higher overall TPF (47% compared to 40% at a point of dichotomy of 1) with only a slight increase in FPF (17% compared to 15% at a point of dichotomy of 1, see Table 5.4). The ROC curve is still close to the chance line and the area under the fitted binormal curve (AUC) is only 0.71 (95% confidence interval (0.67, 0.75)), larger than the 2 month culture result but not the 3 month culture result.

5.4 Discussion

5.4.1 Introduction

The results and graphs in this chapter give a variety of perspectives on the same research question and need to be combined to properly evaluate culture results during treatment as prognostic markers. A perfect prognostic marker

145

	Point of Dichotomy		
	1	20+	IC
TPF	47%	19%	11%
FPF	17%	4%	< 0.5%
PPV	23%	36%	57%
NPV	94%	92%	91%

Table 5.4: True Positive Fraction (TPF), False Positive Fraction (FPF), Positive Predictive Value (PPV) and Negative Predictive Value (NPV) for the marker of the heaviest culture at months 2 to 4.

will correctly predict the outcome for all individuals. None of the markers being evaluated in this chapter achieve this, so it is a matter of comparing the different markers on the basis of the measures discussed in the preceding sections to identify the most useful prognostic markers, and those which should be taken forward in subsequent chapters to be evaluated as surrogate markers.

5.4.1.1 The Prentice Criteria

As noted at the beginning of this chapter, the second of the Prentice Criteria is that the surrogate markers should have some 'prognostic implication' for the true endpoint. The author expressed this criterion (Prentice, 1989) for a time-to-event true endpoint, T, and a continuous surrogate endpoint, S, as: $\lambda_T(t; S) \neq \lambda_T(t)$, that is that the surrogate should have some effect on the hazard of failure. This can be expressed similarly for a binary true endpoint, T, and a binary or continuous surrogate, S, as: $P[T = 1|S] \neq P[T = 1]$. The null hypothesis, $H_0 : P[T = 1|S] = P[T = 1]$, is equivalent to the null hypothesis, $H_0 : b_1 = 0$, where b_1 is the log odds ratio of a poor outcome comparing levels of the binary surrogate from Equation 5.3 on page 135 (see also Equation 5.6). The odds ratios comparing the levels of the binary culture results at months 1 to 4 for different points of dichotomy are shown in Figure 5.2 on page 136. For each point of dichotomy at each month, the estimate of the odds ratio and the widths of the 95% confidence intervals vary widely, but all of the confidence intervals exclude 1. There is strong evidence against the null hypotheses, $H_0 : \exp(b_1) = 1$, can be rejected in each case. On this basis, each culture result at each point of dichotomy formally meets the first Prentice criterion and can therefore be taken forward to be evaluated as surrogate markers. The odds ratios do not tell the whole story and therefore only a

146

selection of the best of these markers will be taken forward to be evaluated as surrogates.

5.4.1.2 False Positives and False Negatives

The term *false positive* refers to those who are incorrectly identified as having a poor outcome and *false negative* refers to those who are incorrectly identified as having a fair outcome. Each of the markers discussed in this chapter lead to a varying number of false positives and false negatives. The ideal marker will have no false negatives or false positives; but if this is not possible, what is the relative *'cost'* of a false positive as compared to a false negative? In clinical trials in modern settings, recurrence rates are commonly no more than 10% (Nunn et al. (2008) identify this as a conservative estimate). The purpose of a prognostic marker will be to identify these 10% of individuals as early as possible without the need of a long period of follow-up. If an individual is identified early as being likely to ultimately have a poor outcome, their combination of drugs can be changed or the duration of treatment extended.

For an individual who will relapse but is not picked up (*a false negative*), retreatment is possible (the WHO recommended retreatment regimen costs between US $15 and US $30 compared to the first-line regimen costing between US $10 and US $20, World Health Organisation (2003)) but only after the individual has manifested symptoms of recurrence and sought medical help. This could be at considerable loss of quality of life of the individual and they could die before retreatment was started.

For an individual who will not have recurrence but has been identified by the marker as having a poor outcome (*a false positive*), treatment will be extended or changed. The burden of treatment for the individual will be increased (unnecessarily so, as the individual will not have recurrence) as will any possible side-effects.

While both are undesirable, a case could be made that a false negative will be more *'costly'* to both the patient and the national tuberculosis programme than a false positive. It is more important to make every effort to identify those who will ultimately have a poor outcome at the expense of wrongly classifying a few who will have a fair outcome. Using this argument the *false positive fraction (TPF)* holds more weight that the *false negative fraction (FPF)* in comparing prognostic markers and this will guide the discussion in this section.

5.4.2 Comparing Markers

Odds ratios were high with 95% confidence intervals excluding 1 for all markers showing that all markers were strongly associated with poor outcome, but not necessarily that any were good prognostic markers. Pseudo-R^2 statistics were all low. As discussed in section 2.4, each of these statistics were developed for a mixed logistic model to mirror the proportion of variation explained, R^2, from linear regression. However, none are completely equivalent to the R^2 from linear regression and it is therefore not straightforward to interpret these. Nevertheless, these pseudo-R^2 statistics do suggest that none of the markers capture a high proportion of the variation in treatment outcome. Prognostic markers are more usefully compared using the TPF, FPF, PPV, NPV and the ROC curve.

5.4.2.1 Month 2 Culture

From other studies looking at predictors of long-term outcome for treatment for tuberculosis summarised in section 3.5, the month 2 culture result has been thought to be the most useful marker. The odds of a poor outcome to treatment for a patient with a positive culture at 2 months are nearly four times those for a patient with a negative culture at 2 months (see Figure 5.2 on page 136). Despite this, in these data, the 2 month culture has a TPF of only 40% at a point of dichotomy of 1 reducing to 1% as the point of dichotomy increases. At a point of dichotomy of 1, the PPV is 22%. At the individual level, this means that only 40% of those with a poor outcome to treatment have a positive culture at 2 months and only 22% of those with a positive culture at 2 months go on to have a poor outcome to treatment.

The 2 month culture does not discriminate well between long-term poor and fair outcomes of treatment and this is shown by the short ROC curve and a low AUC for the fitted binormal curve. Nevertheless, it appears that it may be better than either earlier or later months (see below) and due to its prevalence in the literature, *the two month culture at a point of dichotomy of 1 will be taken forward to be evaluated as a surrogate marker*.

5.4.2.2 Months 3 and 4 Cultures

Only 7% of the 6974 individuals across the twelve trials had a positive culture at 3 months and only 4% at 4 months (see Table A.1 on page 248). It is there-

fore unlikely that cultures at either of these months will be useful prognostic markers.

The TPF for the culture result at month 3 at a point of dichotomy of 1 is only 19% and at month 4 is 15%, though the FPF are also low at 4% and 2%. At the same point of dichotomy of 1, the PPV at month 3 is 37% and at month 4 is 46%.

The slope of PPV at months 3 and 4 is steep (Figure 5.5 on page 141; the slope is similarly steep for the predicted probabilities considering the culture result as continuous, Figure 5.1 on page 131), rising in each case to an approximate plateau around a point of dichotomy of 11. This indicates that a culture result at month 3 or 4 with a higher point of dichotomy may be a better prognostic marker, but the TPF drops off sharply (Table 5.1 on page 140) to 8% for month 3 and 10% for month 4 at a point of dichotomy at 20+, making the use of either marker unfeasible.

The AUC for the fitted binormal ROC curve at month 3 is the greatest (even than month 2), though the 95% confidence interval is wide (since the ROC curve is very short).

There are more positive cultures at month 3 than at month 4 and therefore, *the three month culture at a point of dichotomy of 1 will be taken forward to be evaluated as a surrogate marker.* The four month culture will not be included in the analyses of surrogate markers in subsequent chapters.

5.4.2.3   Month 1 Culture

At a point of dichotomy, the month 1 culture was found to have a too high FPF of 56%. The plot of the PPV across points of dichotomy was flat but rising with point of dichotomy. At a point of dichotomy of 20+ the TPF is 55%, higher than that for the 2 month culture at any point of dichotomy, and the FPF is a reasonable 30%. While a positive culture at a point of dichotomy of 1 at month 1 is not strongly prognostic of a poor outcome, a heavily positive culture (20+ or greater) at month 1 does discriminate better between fair and poor outcomes and could be a passable prognostic marker. Therefore, *the one month culture at a point of dichotomy of 20+ will be taken forward to be evaluated as a surrogate marker.*

5.4.2.4 Summary Marker: Heaviest Culture in Months 2-4

Compared to the month 2 culture, the summary marker of the heaviest culture at month 2, 3 or 4 was shown to have superior TPFs for the different points of dichotomy at only a marginal increase in the corresponding FPFs and also a superior AUC for the fitted binormal ROC curve. This marker appears to be a better prognostic marker, although it may not be as useful in practice as the 2 month culture. The earlier availability of the 2 month culture result may offset the slim benefit in discrimination of the heaviest culture from months 2 to 4. *This marker will therefore not be taken forward to be evaluated as a surrogate marker.*

5.4.2.5 Receiver Operating Characteristic Curve

As shown in figure 5.6 on page 143, the ROC curves for cultures at months 2, 3 or 4 were very short showing the TPF and FPF even for a point of dichotomy of 1 was still low. These results suggest that the culture result after month 1 using solid media does not capture enough of the disease action to be of great use, but that perhaps a culture method that is more sensitive in identifying low colony counts, perhaps a liquid culture method (although there is disagreement about whether liquid culture is more sensitive or is measuring a different population, see section 3.2.5.1), could yield a higher TPF and be a better marker.

5.4.3 Conclusions

All of the markers evaluated in this chapter at each point of dichotomy satisfied the second Prentice criteria that there was some association between the marker and the true endpoint. Nevertheless, not all were useful prognostic markers and only three will be taken forward in subsequent chapters to be evaluated as surrogate markers:

1. the two month culture at a point of dichotomy of 1,

2. the three month culture at a point of dichotomy of 1, and

3. the one month culture at a point of dichotomy of 20+.

Chapter 6

Culture Positivity as a Surrogate Marker for Poor Outcome

6.1 Introduction

In the previous chapter, culture results at months 1, 2, 3 and 4 were evaluated as prognostic markers for poor outcome. Results indicated that culture results at months 2 and 3 were strongly associated with the endpoint of a poor outcome, but were only fair prognostic markers on the basis of the ROC curve. The point of dichotomy of 1 was favoured for both these cultures. The culture result at month 3 was found to be marginally superior as a prognostic marker than the 2 month culture. The point of dichotomy of 20+ for the month 1 culture (classifying all cultures 20+ or higher as positive and 19 colonies or lower as negative) was also found to be a passable prognostic marker. Therefore, these three markers: *culture positive at month 2, culture positive at month 3* and *heavily culture positive at month 1*, will be taken forward in this chapter to be evaluated as surrogate markers. In this chapter, a positive culture at month 2 or at month 3 will be for a point of dichotomy of 1 and a positive culture at month 1 will be for a point of dichotomy at 20+ unless otherwise explicitly stated.

6.1.1 Treatment Ordering

Not all of the trials that yielded data for the analyses in this thesis included an obvious control regimen against which the other regimens were compared. The trials formed part of a larger research programme evaluating the most effective tools for the treatment of TB and were often linked, with each trial building on final, and sometimes interim, results of previous trials. A surrogate endpoint is one which captures the difference on the final endpoint between two treatments in a treatment comparison and it is therefore necessary to identify one regimen from each trial as a nominal 'control' so that treatment comparisons from that trial can be defined in relation to this regimen. In each trial, the regimen that has the highest proportion of poor outcomes (the regimen of least efficacy) is identified as the control. This is so that the difference in risk of poor outcome between the experimental and control regimens is greatest and therefore the treatment ordering is such that the most amount of information is available for evaluating culture results as surrogate endpoints. The only exception to this is in the fourth East African study (Study X, see section 4.1.1.3.1 on page 101) where the 2HRZ/4H regimen is identified as the control regimen. This is so that the intensive phase (the first two months) of treatment of the nominal control regimen is different to that of each of the experimental regimens.

6.1.2 Treatment comparisons included

These data are from 12 trials, comprising 49 treatment regimens, giving a total of 37 treatment comparisons.

The initial 2 month intensive phase in each regimen in three of these trials (East African study Y, the Tanzanian study and the first Singapore study) is the same (SHRZ in each case) and therefore it is not possible for the culture result at month 1 or at month 2 to capture any of the treatment effect in the risk of poor outcome as the treatment difference only occurs in the continuation phase after the end of the second month. Therefore, these four comparisons are excluded from analyses evaluating month 1 and month 2 culture positivity in this chapter. In addition, the two regimens in one comparison in the third Singapore study had the same combination in the first month (S with the combined formulation of HRZ) and was therefore excluded from analyses evaluating month 1 culture positivity in this chapter.

No patients allocated to the control regimen (2SHRZ/4HR) in the first Singapore study, or to the $2HRZ/4H_3R_3$ regimen in the third Singapore study had a positive culture at three months and therefore the log odds ratio of a positive culture at three months could not be estimated for these two treatment comparisons. Analyses evaluating month 3 positivity in this chapter therefore did not include these two treatment comparisons.

In summary, 32 treatment comparisons are used to evaluate the month 1 culture result as a surrogate marker, 33 to evaluate the month 2 culture result and 35 the month 3 culture result.

6.1.3   Chapter Outline

In this chapter, these three binary markers will be evaluated as surrogate markers. In section 6.2 the Prentice criteria (see section 2.5 for a definition) will be evaluated for each of these three markers across the included treatment comparisons and the results from this will be discussed. In section 6.3 a selection of the single trial summary measures introduced in section 2.6 that attempt to quantify the surrogate value of a putative marker will be explored and the results from these measures discussed. In section 6.4, meta-analytic methods introduced in sections 2.8 and 2.9 are applied to these three markers allowing for a comprehensive analysis. In section 6.5 data from two recent tuberculosis trials are added to the database and the most promising models from the previous sections applied to these new data to assess the value of these results in trials today. In section 6.6 the results from each of these analyses are discussed and conclusions drawn.

6.2   Evaluating the Prentice Criteria

Prentice (1989) provided the first clear statistical criteria for defining and evaluating a surrogate marker; and the Prentice Criteria are still accepted today as the starting point for any discussion of surrogate markers (see section 2.5).

6.2.1   Methods

In his original paper, Prentice, to suit his application, stated his criteria with the true endpoint as a time-to-event variable and the surrogate as being time dependent. In this chapter since both the true and the surrogate endpoint are

binary variables, these three criteria can be restated for a binary true endpoint, T, and a binary surrogate endpoint, S:

$$P(T = 1|S = s, Z = z) \equiv P(T = 1|S = s), \qquad (6.1)$$

$$P(T = 1|S = s) \neq P(T = 1), \qquad (6.2)$$

$$E\left[P(T = t \cup S = s|Z = z)\right] \neq E\left[P(T = t \cup S = s)\right] \qquad (6.3)$$

where Z is an indicator variable corresponding to the active or the control treatment.

The second criterion (equation 6.2) states that the surrogate must have some association with the final endpoint. This has been shown to be the case in Chapter 5 for each of the three markers in question. The first criterion (equation 6.1) states that the probability of a poor outcome conditional on the surrogate and the treatment effect is exactly equivalent to the probability of a poor outcome conditional only on the surrogate. This is the key criterion and it will be evaluated in this section. The third criterion (equation 6.3) is necessary to avoid certain artificial situations where the first two criteria are met but the marker is not technically a surrogate. See section 2.5 for discussion. It is evident that this third criterion is met for each marker.

Prentice's original paper combined the tasks of evaluating and using a surrogate in a single trial confusing the issue slightly. In this chapter, culture results are being evaluated as surrogate markers so that they may be used in a future trial. It is therefore necessary to include an additional criterion:

$$P(T = 1|Z = z) \neq P(T = 1). \qquad (6.4)$$

That is, there should be a statistically significant treatment effect on the true endpoint, unadjusted for the surrogate. If this condition does not hold, then there is no treatment effect on the true endpoint and there is therefore nothing for the surrogate to 'capture'.

If these four criteria hold, adjusting for the surrogate renders the treatment effect non-significant and can therefore be said to *(fully) capture the treatment effect.*

To test the Prentice criteria, three models need to be fitted and null hypotheses tested on the parameters from each of these models. These models

are:

1. the model assessing the effect of treatment on the true endpoint:

$$\text{logit} \, P(T_{ik} = 1) = \nu_i + \beta_{ij} Z_{ijk}, \tag{6.5}$$

2. the model assessing the relationship between the surrogate endpoint and the true endpoint:

$$\text{logit} \, P(T_{ik} = 1) = \mu_i + \gamma_i S_{ik}, \quad \text{and} \tag{6.6}$$

3. the model assessing the effect of treatment and the surrogate endpoint on the true endpoint:

$$\text{logit} \, P(T_{ik} = 1) = \nu_i^* + \beta_{ij}^* Z_{ijk} + \gamma_i^* S_{ik}, \tag{6.7}$$

where Z_{ijk} is the indicator variable denoting the treatment comparison:

$$Z_{ijk} = \begin{cases} 1 & \text{if patient } k \text{ was allocated treatment } j \text{ in trial } i \\ 0 & \text{if patient } k \text{ was not allocated treatment } j \text{ in trial } i \end{cases} . \tag{6.8}$$

6.2.1.1 Missing Data

A missing surrogate endpoint, the culture result, is caused by a contaminated result, a lost sputum sample or the patient failing to produce sputum. In most of these trials, patients were kept in hospital for at least the first two months (and in some cases all six months) of treatment. It can therefore be assumed that the culture results that have been observed are representative of the entire population and the data are missing completely at random (MCAR).

A missing true endpoint, poor outcome of treatment, is caused by a patient that is lost during follow-up less than twelve months after the end of treatment. There is evidence on the treatment cards that continued efforts were made to find patients who had failed to come for a follow-up visit to determine whether the patient had had a poor outcome. These efforts included writing letters, visiting the patient's home and talking to friends and relatives. Loss to follow-up was therefore kept to a minimum and was mostly caused by patients moving away or by deaths not related to tuberculosis. It is unlikely that the missingness was related to the unobserved outcome. It is

assumed that the outcomes that have been observed are representative of the entire population and the data are also MCAR.

For each of these models, individuals with either the treatment outcome (the true endpoint T) or the culture results (the surrogate endpoint S) missing are not included in the models. This complete-case analysis is therefore unbiased on the assumption of MCAR. Evaluating the Prentice criteria for each of the three candidate surrogates is therefore on slightly different subsets of trial individuals, as well as a slightly different subset of treatment comparisons (see section 6.1.2).

6.2.2 Results

6.2.2.1 Evaluating for each treatment comparison

Treatment effect on poor outcome (unadjusted)	Effect of 2 month culture on poor outcome	Treatment effect on poor outcome (adjusted for 2 month culture)
21 (64%) Non-significant	→ 6 (29%) Non-significant	
	→ 15 (71%) Significant	
12 (36%) Significant	→ 2 (17%) Non-significant	→ 0 (0%) Non-significant → 2 (100%) Significant
	→ 10 (83%) Significant	→ 3 (30%) Non-significant → 7 (70%) Significant

Table 6.1: Results of testing Prentice criteria for the 2 month culture for individual treatment comparisons

Table 6.1 shows the breakdown of the 33 treatment comparisons by statistical significance for testing the three null hypotheses (here defining *statistically significant* as a two-sided p-value of less that 0.05) for the 2 month culture result. The three null hypotheses are:

1. $H_0 : P(T = 1|Z = 1) = P(T = 1|Z = 0)$; $\beta_{ij} = 0$, testing the unadjusted treatment effect on a poor outcome.

156

2. $H_0 : P(T = 1|S = 1) = P(T = 1|S = 0)$; $\gamma_i = 0$, testing the effect of the two month culture result (the first candidate surrogate) on poor outcome.

3. $H_0 : P(T = 1|S, Z = 1) = P(T = 1|S, Z = 0)$; $\beta_{ij}^* = 0$, testing the treatment effect on poor outcome adjusted for the two month culture result.

Of the 33 treatment comparisons, the ratio of odds of a poor outcome is shown to be different from one in only 12 (36%) comparisons (criterion 4 shown in equation 6.4). Among these 12 treatment comparisons, there was a significant association between the two month culture and a poor outcome in 10 (83%) (criterion 1 shown in equation 6.2) and among these 10 treatment comparisons, the ratio of odds of poor outcome became nonsignificant on adjusting for the surrogate in 3 (30%) (criterion 2 shown in equation 6.1). Therefore, of the 12 treatment comparisons showing a significant effect on a poor outcome, the two month culture satisfied the Prentice criteria in only 3 (25%).

Treatment effect on poor outcome (unadjusted)	Effect of 2 month culture on poor outcome	Treatment effect on poor outcome (adjusted for 2 month culture)
24 (65%) Non-significant	→ 9† (38%) Non-significant	
	→ 13 (54%) Significant	
13 (35%) Significant	→ 5(38%) Non-significant	→ 1 (20%) Non-significant → 4 (80%) Significant
	→ 8 (62%) Significant	→ 0 (0%) Non-significant → 8 (100%) Significant

†In two treatment comparisons, culture positivity at 3 months predicted poor outcome perfectly and is therefore not including in this column.

Table 6.2: Results of testing Prentice criteria for the 3 month culture for individual treatment comparisons

Treatment effect on poor outcome (unadjusted)	Effect of 2 month culture on poor outcome	Treatment effect on poor outcome (adjusted for 2 month culture)
20 (62%) Non-significant	→ 6† (30%) Non-significant	
	→ 13 (65%) Significant	
12 (38%) Significant	→ 4 (33%) Non-significant	→ 0 (0%) Non-significant
		→ 4 (100%) Significant
	→ 8 (67%) Significant	→ 1 (13%) Non-significant
		→ 7 (88%) Significant

†In one treatment comparison, culture positivity at 1 months predicted poor outcome perfectly and is therefore not including in this or the next column.

Table 6.3: Results of testing Prentice criteria for the 1 month culture dichotomised at 20+ for individual treatment comparisons

Table 6.2 shows the results of testing the Prentice criteria for the 3 month culture result as a surrogate for each of the 37 treatment comparisons. Of the 13 treatment comparisons showing a significant effect on poor outcome, the three month culture did not satisfy the Prentice criteria in any treatment comparison.

Table 6.3 shows the results of testing the Prentice criteria for the 1 month culture result taking the point of dichotomy at 20+. Of the 12 regimens with a significant effect on a poor outcome, the 1 month culture result at a point of dichotomy of 20+ satisfied the Prentice criteria in only 1 (8%) treatment comparison.

These results suggest that the month 2 culture is marginally superior to the other two markers as the Prentice criteria were met in more comparisons, but all three are poor.

6.2.2.2 Evaluating across treatment comparisons

When combining the data and testing each of the three null hypotheses across all treatment comparisons, the picture is not much improved. Figure 6.1 shows

an estimate of the unadjusted treatment effect and the treatment effect adjusted for each of the candidate surrogate with 95% confidence intervals, using a mixed logistic regression model adjusting for clustering within trials.

In each case, the surrogate had a statistically significant effect on a poor outcome satisfying criterion 2. For month 2 the log odds ratio was -0.74 with 95% confidence interval (-1.01, -0.47), for month 3, the log odds ratio was -0.77 with 95% confidence interval (-1.01, -0.54), and for month 1 the log odds ratio was -0.75 with 95% confidence interval (-1.03, -0.48). It is clear to see that the origin is not contained in the 95% confidence interval of the unadjusted treatment effect satisfying criterion 4, but the same is also true for the treatment effect adjusted for the surrogate and therefore criterion 1 is not satisfied. For month 3 and month 2 cultures, the point estimate and the confidence interval for the adjusted treatment effect are shifted towards the origin, with a greater shift occurring with the month 2 culture, but this shift is very slight. For the month 1 culture, the point estimate is shifted slightly towards the origin, but the width of the confidence interval is increased.
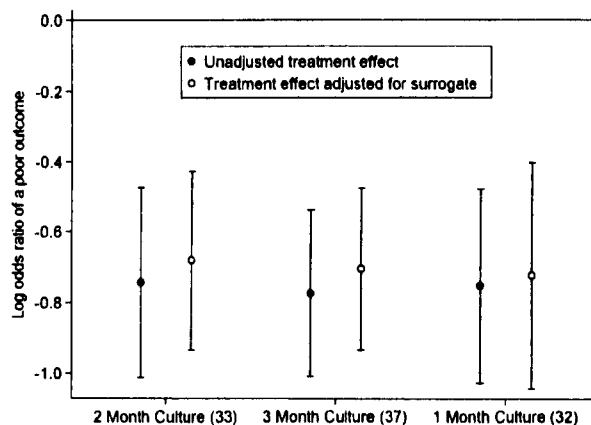
Figure 6.1: Unadjusted treatment effect on a poor outcome and treatment effect adjusted for the surrogate marker for each of three candidate markers. 95% confidence interval is shown and number of treatment comparisons included is shown in parentheses on the horizontal axis labels.

It should be noted that this evaluation across treatment comparisons assume a common log odds ratio of poor outcome across all treatment comparisons. This is not a realistic assumption given that each treatment comparison

is of different treatment regimens, but is useful as an indicator of the behaviour of the marker.

6.2.3 Conclusions

Each of the three candidate surrogate markers being evaluated in this chapter, culture positivity at months 2 and 3 and heavy culture positivity at month 1, have failed to satisfy the Prentice criteria in all treatment comparisons.

Testing the criteria for individual treatment comparisons showed that in only a small proportion of comparisons were the Prentice criteria satisfied, with the 2 month culture having marginally the largest proportion. This means that in the majority of comparisons, the Prentice criteria demonstrated that the markers were *not* surrogates.

When evaluating the criteria across all treatment comparisons, adjusting for the surrogate marker did appear to capture a very small fraction of the treatment effect, but certainly did not *fully* capture the treatment effect and therefore the Prentice criteria were not satisfied. The 2 month culture result produced the greatest movement in the point estimate and confidence interval suggesting this captured more of the treatment effect.

The Prentice criteria has been criticised as only being a means of excluding failed surrogates (see section 2.5.1), and not useful for evaluating the value of surrogate markers that may be less than perfect but still useful. A non-significant result for the hypothesis test of criteria 2 does not necessarily imply surrogacy, it only means that surrogacy cannot be excluded. This criticism is irrelevant in this case as all three markers fail to satisfy the Prentice criteria overall and therefore should be classed as failed surrogates on this basis. None of these candidate surrogate markers are perfect surrogates and in fact appear to be very poor. Despite these disappointing results, subsequent sections use additional methods to explore the use of these three candidate markers as surrogates and to attempt to quantify the value of each.

6.3 Single Trial Summary Measures

6.3.1 Introduction

Following the route of the development of methods for evaluating surrogate markers (see Chapter 2) and having now tested the Prentice criteria for the

160

three candidate markers, summary measures for quantifying the proportion of treatment effect captured by the surrogate are now calculated. Even though the Prentice criteria have not been satisfied, indicating these none of the three candidates fully capture the treatment effect on the true endpoint, single trial summary measures and subsequent methods in the rest of this chapter will be used to quantify the degree to which these candidate markers can be used as surrogates.

A variety of different measures for the proportion of treatment effect explained by the marker were reviewed in sections 2.6 and 2.8, and four have been selected for assessment using these data: (i) the *Proportion of Treatment Effect* (Freedman et al., 1992), (ii) the *Risk Reduction* (Li et al., 2001), (iii) *F* and *F'* (Wang and Taylor, 2002), and (iv) the *Relative Effect* and the *Adjusted Association (AA)* (Buyse and Molenberghs, 1998).

6.3.2 Methods

The formulae for each of these measures contains a quotient in some form of other involving estimates of model parameters in both the numerator and the denominator. Standard errors on these individual parameters can be estimated in the usual way, but calculating a standard error for the quotients is not straightforward. For calculating confidence intervals, most authors have recommended using Fieller's method (Fieller, 1940) or the δ-method (Freedman, 2001), both being approximate methods. With the continual improvements in computing power, Monte Carlo methods are now more accessible than when these measures were first proposed and confidence intervals are calculated in this section using bootstrap methods. 1000 replications were used in a nonparametric bootstrap to calculate the bias corrected and accelerated (BCa) 95% confidence intervals Carpenter and Bithell (2000).

Point estimates and 95% confidence intervals for each of these four measures are calculated for each of the treatment comparisons, evaluating each of the three candidate markers.

6.3.2.1 The Proportion of Treatment Effect (PTE)

The idea of quantifying the proportion of treatment effect that is captured by a surrogate marker rather than merely testing whether it is fully captured or not was first suggested by Freedman et al. (1992) and the PTE was the

measure that they proposed. As the first measure proposed and the measure that ushered in the change in focus from testing to quantifying, this measure is included in this analysis, despite widespread criticism.

Two separate models are fit for each treatment comparison j in trial i:

$$\text{logit}\, P(T_{ik} = 1) = \nu_{ij} + \beta_{ij} Z_{ijk}, \tag{6.9}$$

$$\text{logit}\, P(T_{ik} = 1) = \nu_{ij}^* + \beta_{ij}^* Z_{ijk} + \gamma_{ij}^* S_{ik}, \tag{6.10}$$

where Z_{ijk} is the indicator variable denoting the treatment comparison:

$$Z_{ijk} = \begin{cases} 1 & \text{if patient } k \text{ was allocated treatment } j \text{ in trial } i \\ 0 & \text{if patient } k \text{ was not allocated treatment } j \text{ in trial } i \end{cases}, \tag{6.11}$$

T_{ik} and S_{ik} are the true and surrogate endpoints for participant k in trial i.

The PTE is then:

$$\text{PTE}_{ij} = \frac{\hat{\beta}_{ij} - \hat{\beta}_{ij}^*}{\hat{\beta}_{ij}}, \tag{6.12}$$

where $\hat{\beta}_{ij}^*$ is the estimate of the treatment effect adjusted for the surrogate, β_{ij}^*, and $\hat{\beta}_{ij}$ is the estimate of the unadjusted treatment effect, β_{ij}. This yields one value of the PTE for each treatment comparison.

6.3.2.2 The Risk Reduction

Li et al. (2001) propose this measure and consider the risk of a poor outcome to a patient before the start of treatment and the reduction in risk as a result of the treatment. The Risk Reduction (RR) is the proportion of this reduction in risk that is captured by the surrogate.

One model is fit for each treatment comparison j in trial i:

$$\log P(T_{ik} = 1) = \nu_{ij}^* + \beta_{ij}^* Z_{ijk} + \gamma_{ij}^* S_{ik}. \tag{6.13}$$

This link function for this generalised linear model is the log function, rather than the logit function yielding estimates of risks rather than odds. The RR is:

$$RR_{ij} = \frac{1 - \exp\left(\hat{\gamma}_{ij}^* \Delta_S\right)}{1 - \exp\left(\hat{\beta}_{ij}^* + \hat{\gamma}_{ij}^* \Delta_S\right)} \tag{6.14}$$

where Δ_S is the difference in the proportion of participants with $S_{ik} = 1$ in the control regimen subtracted from the proportion of participants with $S_{ik} = 1$ in the experimental regimen for treatment comparison j.

6.3.2.3 F and F′

A similar measure to the RR, F, is the reduction in risk due to the change in the surrogate induced by the treatment divided by the total reduction in risk between treatment groups, where the risks are estimated differently (Wang and Taylor, 2002). F is estimated based on the distribution of the true endpoint given the treatment and the surrogate endpoint and the distribution of the surrogate endpoint given the treatment.

For binary true and surrogate endpoints, S and T, the measure is:

$$F_{ij} = \Big[P(T_{ik} = 1 | Z_{ijk} = 0, S_{ik} = 1) - P(T_{ik} = 0 | Z_{ijk} = 0, S_{ik} = 0) \Big]$$
$$\cdot \frac{P(S_{ik} = 1 | Z_{ijk} = 0) - P(S_{ik} = 1 | Z_{ijk} = 1)}{P(T_{ik} = 1 | Z_{ijk} = 0) - P(T_{ik} = 1 | Z_{ijk} = 1)}, \tag{6.15}$$

and a complementary form:

$$F'_{ij} = \Big[P(T_{ik} = 1 | Z_{ijk} = 1, S_{ik} = 1) - P(T_{ik} = 0 | Z_{ijk} = 1, S_{ik} = 0) \Big]$$
$$\cdot \frac{P(S_{ik} = 1 | Z_{ijk} = 0) - P(S_{ik} = 1 | Z_{ijk} = 1)}{P(T_{ik} = 1 | Z_{ijk} = 0) - P(T_{ik} = 1 | Z_{ijk} = 1)}, \tag{6.16}$$

where these are calculated for each treatment comparison j in trial i. These probabilities are estimated empirically from the data. For example, $P(T_{ik} = 1 | Z_{ijk} = 0)$ is estimated as the proportion of those trial participants in the control group for treatment comparison j ($Z_{ijk} = 0$) that have a poor outcome ($T_{ik} = 1$), and $P(T_{ik} = 1 | Z_{ijk} = 1, S_{ik} = 1)$ is estimated as the proportion of those trial participants with a positive culture result ($S_{ik} = 1$ for the candidate marker being evaluated) in the experimental treatment arm for comparison j ($Z_{ijk} = 1$) that have a poor outcome ($T_{ik} = 1$). Since F_{ij} and F'_{ij} are complementary, only the results of F_{ij} will be presented.

6.3.2.4 The Relative Effect and the Adjusted Association

Two measures were proposed to replace the concept of a single measure attempting to quantify the proportion of treatment effect captured by the marker

(Buyse and Molenberghs, 1998). The relative effect (RE) is ratio of the treatment effect on the true endpoint and the treatment effect on the surrogate endpoint and the adjusted association (AA) is the association between the true and surrogate endpoints adjusted for the treatment.

Three separate models are fit for each treatment comparison j in trial i:

$$\text{logit}\, P(S_{ik} = 1) = \mu_{ij} + \alpha_{ij} Z_{ijk}, \tag{6.17}$$

$$\text{logit}\, P(T_{ik} = 1) = \nu_{ij} + \beta_{ij} Z_{ijk}, \tag{6.18}$$

$$\text{logit}\, P(T_{ik} = 1) = \nu_{ij}^* + \beta_{ij}^* Z_{ijk} + \gamma_{ij}^* S_{ik}. \tag{6.19}$$

The RE is:

$$RE_{ij} = \frac{\beta_{ij}}{\alpha_{ij}}, \tag{6.20}$$

and the AA is:

$$AA_{ij} = \gamma_{ij}^*. \tag{6.21}$$

6.3.3 Results

Figures A.1, A.2 and A.3 in Appendix A show the point estimates and 95% bootstrap confidence intervals for each of the five measures for each of the three candidate markers across all 37 treatment comparisons. Confidence intervals have been censored in the figures at +3 and -2 with an arrow indicating if the bounds of the confidence interval lie outside the interval $[-2, +3]$. For some treatment comparisons, the point estimate and the 95% confidence interval lay completely outside of the interval $[-2, +3]$ and are therefore not shown on the graphs. In addition, some point estimates are not shown due to problems with convergence. Table 6.4 shows summaries of the point estimates of the five measures across all treatment comparisons.

There is great variation in the point estimates and wide 95% confidence intervals for each of the single trial measures. No more than 80% of point estimates lie in the interval [0,1] and can therefore validly be called 'proportions'.

The medians for each of the PTE, the RR and F are very low and all lie between 0.01 and 0.11. The medians for the RE and the AA are greater showing a clear contrast.

164

Measure	Month	$N^\dagger$	$n^\ddagger$	Min	Max	Median	IQR	Proportion $\in [0,1]$
PTE	Month 1	31	0	-2.14×10^{14}	1.82	0.01	0.11	45%
	Month 2	22	0	-0.63	3.18	0.11	0.26	77%
	Month 3	14	0	-0.12	0.81	0.03	0.09	79%
F	Month 1	36	0	-1.04	1.75	0.03	0.10	67%
	Month 2	37	0	-10.71	2.32	0.11	0.36	59%
	Month 3	36	0	-0.04	0.80	0.05	0.15	75%
RR	Month 1	37	0	-2.04	1.62	0.03	0.13	54%
	Month 2	37	0	-1.93	1.90	0.07	0.25	57%
	Month 3	37	0	-3.93	0.54	0.02	0.06	73%
RE	Month 1	25	2	-8.55	25.44	2.11	4.08	
	Month 2	23	1	-17.77	21.49	0.82	3.27	
	Month 3	35	1	-1.00×10^{15}	112.07	1.37	2.43	
AA	Month 1	25	0	-0.02	0.11	0.04	0.04	
	Month 2	22	0	0.18	1.98	1.13	0.29	
	Month 3	34	0	0.08	3.28	1.65	0.66	

$\dagger$The number of treatment comparisons for which the point estimate was calculated.

$\ddagger$The number of treatment comparisons for which the entire 95% confidence interval and point estimate lie outside the interval $[-2, +3]$. The total number of treatment comparisons plotted on the graphs in Appendix A is therefore N – n.

Table 6.4: Summary results of the singe trial measure across the 37 treatment comparisons.

6.3.4 Conclusions

Some authors have proposed a weighted average of a single trial measure across trials to yield an overall measure of the proportion of treatment effect that is captured by the surrogate (see section 2.9). With such great variation and wide confidence intervals in these results, a weighted average would not be an appropriate summary as it would conceal the fact that so many of the point estimates and confidence intervals are outside of the interval [0,1]. The medians for each of the PTE, the RR and F all lie between 0.01 and 0.11 either suggesting poor surrogacy or, more likely, demonstrating uninformative measures.

Freedman et al. (1992) suggest that a surrogate could be deemed important if the lower limit of the confidence interval of the PTE is greater than a critical value such as 0.5 or 0.75 (see section 2.6). On this basis, these markers could only be called 'important' in very few of these treatment comparisons.

Li et al. (2001) do give an explanation for a RR of greater than 1, noting that this corresponds to treatment effects on the true and surrogate endpoint that are in opposite directions. This could be an explanation of those values of RR greater than 1, but not of those that are less than 0.

Neither the RE, nor the AA are purported to be 'proportions' and therefore have a clear interpretation if the estimate lies outside of [0,1]. Nevertheless, the RE is simply the ratio of the treatment effect on the surrogate endpoint and the treatment effect on the true endpoint and is of little value when calculated in a single trial. The AA is useful as a treatment-comparison-specific estimate of the association between the true and surrogate endpoints adjusted for treatment. The results suggest that this association increases from month 1 to month 2 to month 3, but the spread (and therefore the standard error of any overall estimate) also similarly increases.

6.4 The Meta-Analytic Approach

6.4.1 Introduction

In section 6.2, the Prentice criteria were tested and each of the three markers were shown to satisfy only two of the three criteria. In section 6.3, four different approaches were used in an attempt to quantify the proportion of treatment effect captured by the surrogate. These measures were shown to have a number of deficiencies, not least that they are all single trial measures and there is no obvious way to extend these measures across trials. Any marker must be a valid surrogate across a number of trials and therefore it is meta-analytic methods that are most appropriate for evaluating the three candidate markers as surrogates.

In Chapter 2, two different meta-analytic paradigms to surrogate marker evaluation were summarised. Section 2.8 discussed a series of methods based on the work of several Belgian authors (including Burzykowski et al. (2005), the first and only textbook on surrogate marker evaluation) hereafter denoted as the *Belgian Paradigm*. In section 2.9, a two-stage graphical approach was summarised that was developed and applied largely in the disease area of HIV (see HIV Surrogate Marker Collaborative Group (2000) and Chapter 17 of Burzykowski et al. (2005)) hereafter denoted as the *HIV Paradigm*. In both of these approaches, it is necessary first to estimate the effect of the treatment on the surrogate and the true endpoint before modelling the relationship between

the two.

In the *Belgian paradigm*, these two stages are effectively combined into a one-stage mixed effects model, including the effect of trial as a random effect. The key is that the true endpoint and the surrogate endpoint are modelled jointly yielding two statistics, R^2_{indiv} and R^2_{trial}, that assess the quality of the surrogate at the individual and the trial level respectively. For Gaussian true and surrogate endpoints, the joint distribution is multivariate Gaussian which can be modelled using standard mixed model theory and R^2_{indiv} and R^2_{trial} easily estimated. The joint distribution is more complicated if one or both endpoints are not normally distributed. Various authors have suggested using copulas to specify the relationship between two random variables given standard marginals. Some work has been done in this area, but the fitting of the model when both endpoints are not Gaussian is not straightforward and does not lend itself to ready application, particularly as copulas are not widely used in medical statistics, being more common in financial statistics (Genest and MacKay, 1986).

In the *HIV paradigm*, the two stages are separate. The treatment effect on the surrogate endpoint and on the true endpoint are first estimated within trial i. The treatment effect on the true endpoint, β_i, is plotted against the treatment effect on the surrogate endpoint, α_i, and a straight line fitted to these points. Since the two stages are distinct, this allows for a variety of types of endpoints and models in the first stage since it is only the estimates of β_i and α_i that are taken forward to the second stage. β_i and α_i will be estimated with precision varying across trials, as some will be smaller or larger and some treatments will have a greater or lesser effect on the true endpoint. Some measure of the precision from the first stage will therefore need to be incorporated in the second stage. The authors of the original paper (Daniels and Hughes, 1997) use what they describe as an empirical Bayesian approach with non-informative priors for estimating the fitting the models and estimating the parameters.

The data used in this thesis are from multi-arm trials with more than one regimen and therefore more than one treatment comparison within a single trial. Evaluating surrogate markers using data from multi-arm trials is not discussed extensively in the literature and neither of these approaches to surrogate marker evaluation are directly suited to this application. Some additional methodological development is therefore necessary.

The *HIV paradigm* will be the basis for the analysis in this section, taking a

frequentist rather than a Bayesian approach. For this setting, the true endpoint is poor outcome to treatment and therefore β_i will be the log odds ratio of a poor outcome between the two treatments in the treatment comparison. There are three markers which are being evaluated as surrogates in this chapter and therefore α_i will be the log odds ratio of a positive culture at each of months 1 (at a point of dichotomy at 20+), 2 and 3.

Tibaldi et al. (2003) observe that if the true and surrogate endpoint are modelled separately then it is harder to study the *individual-level* surrogacy. This is not a considerable drawback as determining and assessing *trial-level* surrogacy is of more importance in the context of this thesis.

The assumptions necessary for a meta-analysis are discussed in section 6.4.2, the methods are described in section 6.4.3 and the results are summarised in section 6.4.4. The models are repeated incorporating adjustments for baseline covariates in section 6.4.5 and fit for different subgroups of the population in section 6.4.6. The results of these analyses are discussed in section 6.4.7.

6.4.2 Meta-Analytic Assumptions

Three assumptions are necessary for a meta-analysis. These assumptions are satisfied in the data used in this thesis as detailed below.

6.4.2.1 Homogeneity of Trials

The trials that yielded the data that are used for these analyses are described in detail in Chapter 4. Several trials were not selected for inclusion in this project on the basis of clear criteria, despite data being available. Each of the trials that were selected were conducted by the TB research units of the British Medical Research Council within a period of around twenty years. Clinical and bacteriological protocols from many of these trials have been recovered from archives and it has been verified that they are very similar. The personnel involved in these trials were very similar (although the local staff were specific to the trial sites, many of which were nevertheless involved in more than one trial) and the treatment cards and data elements recorded were very similar. Patients with extra-pulmonary tuberculosis were excluded where this was detected, and one trial looking at silico-tuberculosis was not included. It is therefore reasonable to assume that all aspects of the conduct of

the trials were very similar and these data can be combined in a meta-analysis.

6.4.2.2 Comparability of Regimens

Each of the regimens included in these data were six month combinations (with various dosing schedules) of up to five of only the following six drugs: *isoniazid, rifampicin, pyrazinamide, ethambutol, streptomycin* and *thiacetazone*.

The regimens in each of these trials were therefore very similar and can reasonably be combined in this meta-analysis. Apart from thiacetazone, discontinued due to the increased risk of Stevens-Johnson syndrome in patients who are HIV positive (Fox et al., 1999), and streptomycin, rarely used today as a first-line drug since it is the only treatment in this list given intravenously (Global Alliance for TB Drug Development, 2008), these drugs are used routinely in all parts of the world to treat tuberculosis (World Health Organisation, 2003). A surrogate evaluated on a particular drug or a particular class of drugs can then only be used for that class of drugs. Different drugs for the same disease may have different surrogates. Any new drug evaluated today in a phase III trial will be evaluated in combination with some or all of the first four drugs in this list. An example is the REMoxTB trial evaluating moxifloxacin as a replacement for either ethambutol or isoniazid in the standard six month regimen 2ERHZ/4HR (see Nunn et al. (2008) for details).

Therefore, not only are the regimens very similar in each trial used in this analysis, they can reasonably be considered to be in the same class as those used in a phase III trial today, meaning that a surrogate validated on these data could reasonably be used in a trial today.
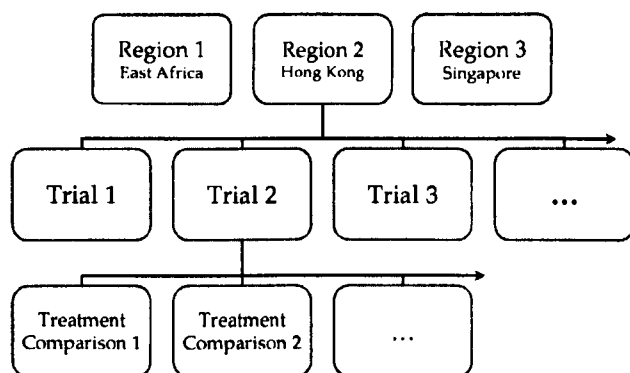
Figure 6.2: Hierarchical structure of the data.

6.4.2.3 Choice of Units

Figure 6.2 on the preceding page shows the hierarchical structure of the data. The 37 treatment comparisons (from a total of 49 regimens) are grouped in twelve trials which in turn are grouped into three geographical regions. In this setting, the unit in stage II of the analysis will be treatment comparison. The relationship between the treatment effect on a poor outcome and the treatment effect on a positive culture will be evaluated at the treatment comparison level. However, treatment comparisons are clustered within trials and trials are clustered within geographical region. The clustering within trials is important and will be addressed in the models. Differences between geographical regions will be explored in section 6.4.6.

6.4.3 Methods: Building the Model

The analysis is based on the HIV paradigm described previously. The model comprises two stages. In the first stage, the treatment effects, α_i and β_i are estimated within each trial. In the second stage, a regression line is fitted to the treatment effects α_i and β_i, and the relationship between them explored.

6.4.3.1 Stage I: Estimating the treatment effects

Across the twelve trials included in this analysis, there were varying numbers of treatment arms. The number of regimens in a trial varied from two (the first Singapore study) to eight (the fourth Hong Kong study), and therefore the number of treatment comparisons in a trial varied from one to seven. The subscript, $j = 1, \ldots, m_i$, represents the treatment comparison within a trial in question, where m_i is the number of treatment comparisons in trial i (and $\sum_{i=1}^{12} m_i = 37$).

Separate models were fitted for each trial estimating the treatment effect, β_{ij}, on the true endpoint (poor outcome) and the treatment effect, α_{ij}, on the surrogate endpoint (culture positivity) for treatment comparison j in trial i, $i = 1, \ldots, 12$. For each trial i, the two models fitted were as follows:

$$\text{logit}\left(P\left(S_{ik} = 1|Z_{ijk}\right)\right) = \mu_i + \sum_{j=1}^{m_i} \alpha_{ij} Z_{ijk} \tag{6.22}$$

$$\text{logit}(P\left(T_{ik} = 1|Z_{ijk}\right)) = \nu_i + \sum_{j=1}^{m_i} \beta_{ij} Z_{ijk} \tag{6.23}$$

where

$$Z_{ijk} = \begin{cases} 1 & \text{if patient } k \text{ was allocated treatment } j \text{ in trial } i \\ 0 & \text{if patient } k \text{ was not allocated treatment } j \text{ in trial } i \end{cases} \tag{6.24}$$

is the indicator variable denoting the treatment. μ_i and ν_i are intercept terms in the model and are estimated, but are treated as ancillary parameters and not taken forward to the second stage. These two models give us our 37 estimator pairs $(\hat{\alpha}_{ij}, \hat{\beta}_{ij})$ and variances of these estimates, $\sigma^2_{\hat{\alpha}_{ij}} = \text{Var}(\hat{\alpha}_{ij})$ and $\sigma^2_{\hat{\beta}_{ij}} = \text{Var}(\hat{\beta}_{ij})$, which are themselves estimated from the models.

6.4.3.2 Stage II: Modelling the relationship between treatment effects

Following the HIV paradigm, we fit a straight line to the pairs $(\hat{\alpha}_{ij}, \hat{\beta}_{ij})$:

$$\hat{\beta}_{ij} = \delta + \kappa \hat{\alpha}_{ij} + \varepsilon_{ij}, \tag{6.25}$$

imposing $\delta = 0$, where $\varepsilon_{ij} \sim N\left(0, \tau^2\right)$ is the between treatment comparison error. Apart from the move from a Bayesian to a frequentist framework, this is following the HIV paradigm with three differences:

1. *Imposing $\delta = 0$.* The HIV paradigm includes an intercept, δ, in this second stage. This allows for an additional test of $H_0 : \delta = 0$. A good surrogate will have a zero intercept (the fitted line goes through the origin). If the intercept is non-zero, this means that there is a proportion of the treatment effect on the true endpoint that is not captured by the surrogate—perhaps indicating that there are causal pathways that do not pass through the surrogate. Including an intercept in this model is only reasonable if all the treatment comparisons are comparing the same treatment with the same control (as is the case in the example used to illustrate the methods in Daniels and Hughes (1997)). In these data, all of the treatment comparisons are not comparing the same regimens and

171

therefore this model must be fit without an intercept, imposing $\delta = 0$. $\delta \neq 0$ would suggest that there were a residual treatment effect on the true endpoint that was not captured by the surrogate. Since all the comparisons are different, this would be a meaningless concept. It may be the case that a proportion of the treatment effect on the true endpoint is not captured by one of the three candidate surrogates being evaluated in this chapter, and this can be observed by the spread of the data about the fitted line.

2. In the simplest and most natural fitting of the model as described, the estimation procedure would include the assumption that each of the pairs $(\hat{\alpha}_{ij}, \hat{\beta}_{ij})$ are independent and of equal weight. This, of course, is not the case in this analysis, $(\hat{\alpha}_{ij}, \hat{\beta}_{ij})$ are clustered within trials as each pair with the same i are comparisons with the same control arm, and these estimates themselves are of varying precision depending on actual the number of patients and the number of poor outcomes observed in patients involved in a treatment comparison. Two changes are therefore made.

   (a) *Weighted regression*. Weighted regression has been used, with weights, w_{ij}, equal to the inverse of the mean of the variances of $\hat{\alpha}_{ij}$ and $\hat{\beta}_{ij}$:

   $$w_{ij} = \frac{2}{\sigma^2_{\alpha_{ij}} + \sigma^2_{\beta_{ij}}}, \tag{6.26}$$

   to account for the presence of heteroskedasticity. The choice of weights is largely arbitrary, but this choice means that pairs estimated with greater precision (smaller variances) are more influential in the fitting of the regression line that those estimated with poorer precision.

   (b) *Robust standard errors*. Calculating robust standard errors, rather than the standard ordinary least squares estimates of the standard errors, means that the model is *robust* to departures from the assumption of independence, the only assumption is that observations are independent within each trial. Robust standard errors have therefore been used to account for the clustering of treatment comparisons within trials.

(a) Month 1, dichotomised at 20+

(b) Month 2

(c) Month 3
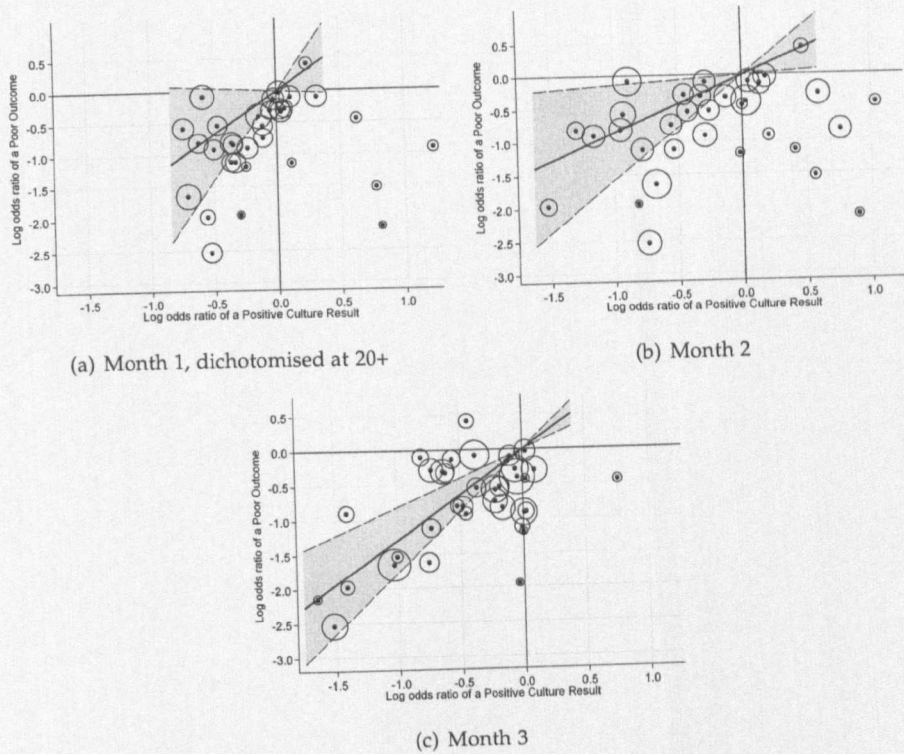
Figure 6.3: Logs odds ratio of a poor outcome plotted against log odds ratio of a positive culture. Fitted line is weighted by the precision of the estimates, and this precision is represented by the diameter of the circles around each point. A 95% confidence interval on the slope is also shown.

6.4.3.2.1 Treatment Ordering

As described in section 6.1.1, the ordering of the treatment regimens was such that the log odds ratio of a poor outcome would be negative for almost all of the comparisons. Almost all of the points plotted are therefore in one of the two lower quadrants in each of the graphs.

It is informative to look at how the points plotted in stage II are distributed about the Cartesian quadrants on the graphs. If the points lie on the horizontal axis, this indicates that there is no difference in the log odds of a poor outcome between the treatment regimens being compared and if the points lie on the vertical axis, this indicates that there is no difference in the log odds of a positive culture result between the treatment regimens being compared.

Importantly, if a point lies in either the upper left or the lower right quadrants, this indicates that the sign of the log odds ratio of a poor outcome is opposite to that of the log odds ratio of a positive culture. For such points,

173

the treatment has *opposite* effects on the surrogate to the true endpoint which is highly undesirable for a surrogate endpoint. A good surrogate will have all points lying in the upper right or lower left quadrants.

6.4.4 Results

Figure 6.3 on the previous page is a plot of the log odds ratio of a poor outcome ($\hat{\beta}_{ij}$) against the log odds ratio of a positive culture ($\hat{\alpha}_{ij}$) for each of the three candidate surrogate markers: culture positivity at month 1 with a point of dichotomy at 20+, culture positivity at month 2, and culture positivity at month 3. The fitted line is weighted by the precision of the estimates where precision is calculated as the inverse of the mean of the variances of the two estimates in each pair. Two additional dashed lines are plotted in each graph with slopes as the upper limit and lower limit of the 95% confidence interval on the estimated slope. Table 6.5 shows the slope with 95% confidence interval and the proportion of explained variation, R^2. This is the proportion of variation in the log odds ratio of a poor outcome (the treatment effect on the true endpoint) that is explained by the log odds ratio of a positive culture (the treatment effect on the candidate surrogate) and is a way of quantifying the performance of a surrogate. $R^2 = 1$ shows that the marker is a perfect surrogate, and $R^2 < 0.5$ (or even $R^2 < 0.6$) shows that that marker is a poor surrogate. The table also includes the number of treatment comparisons included in the regression and plotted on the graphs (see section 6.1.2 above). The total number of treatment comparisons available was 37. The values of α_{ij} and β_{ij} with 95% confidence intervals estimated in stage I are tabulated for each comparison in Appendix A in Tables A.2, A.3 and A.4.

Marker	Number of Comparisons	Number of Trials	Estimated Slope, κ	95% CI	R^2
Month 1[†]	32	9	1.35	(-0.10,2.80)	0.36
Month 2	33	9	0.85	(0.13,1.57)	0.36
Month 3	35	11	1.29	(0.82,1.76)	0.69

[†]With a point of dichotomy at 20+.

Table 6.5: Results of stage II of the meta-analysis for each of the three candidate surrogate markers.

6.4.4.1 Month 1, Heavily Positive

There is considerable scattering about the fitted line in Figure 6.3(a) on page 173, and the proportion of variation explained is only 0.36 (Table 6.5 on the preceding page). There is grouping around the origin showing no real difference between treatments in either rates of poor outcome or proportions of heavily positive at month 1 for some comparisons. Apart from this grouping around the origin, there are six points in the lower right quadrant indicating that the treatment direction on a poor outcome is opposite to that on the heavily culture positivity at month 1 although, for only one of these points was the log odds ratio of a positive culture statistically different from zero (the second treatment comparison in Study X, see Table A.2 on page 252). There are a number of points that lie outside and below the 95% confidence interval on the slope suggesting that there is a proportion of the treatment effect on the true endpoint in these particular treatment comparisons that is not explained by the candidate surrogate of a heavily positive culture at month 1.

6.4.4.2 Month 2

Again, there is considerable scattering about the fitted line in Figure 6.3(b) with the proportion of explained variation again low at 0.36, although the spread of the points in the horizontal direction is greater than at month 1. Compared to month 1, there are more points that lie outside and below the 95% confidence interval on the slope suggesting that there is a proportion of the treatment effect on the a poor outcome that is not captured by the 2 month culture. Not counting the grouping around the origin, there are also seven points that lie in the lower right quadrant, and for two of these points the log odds ratio of a positive culture was statistically significant (the first and third treatment comparison in Study U, see Table A.3 on page 253). The log odds ratio of a poor outcome was also statistically significant in the third comparison meaning that both the treatment effect on the month 2 culture result and on a poor outcome were statistically significant, but in opposite directions. Surprisingly, there are more points for the 2 month culture than the 1 month heavily positive culture in this lower right quadrant, corresponding to comparisons for which the treatment effect on a poor outcome is in the opposite direction to the treatment effect on the culture result.

175

6.4.4.3 Month 3

There is less scattering about the fitted line in Figure 6.3(c) than in the other two figures and this is reflected in a proportion of explained variation considerably higher at 0.69 and the narrowest 95% confidence interval on the slope. There are fewer points below the 95% confidence region around the fitted line and, excluding the clustering around the origin, there is only one point in the lower right quadrant and one in the upper left.

6.4.5 Adjusting for Baseline Risk Factors

6.4.5.1 Introduction

In an important paper looking to evaluate aspects of the serial sputum colony counts (SSCC) profile as surrogate markers (Davies et al., 2006a), the authors found that the second slope parameter in the bi-exponential model better discriminated between treatments when additional patient characteristics (in particular HIV status) were included in the model. While long-term treatment outcome was not available in this study, and therefore this second slope parameter of the SSCC analysis could not formally be evaluated as a surrogate marker, this work does suggest that failing to adjusting for patient-level baseline risk factors could show a bacteriological surrogate marker to be poorer than it actually is. The analysis described above was therefore repeated including important baseline risk factors as covariates in the model to see if the fit of the model and the performance of the surrogate marker improves.

These data are from randomised controlled trials where trial participants were randomly allocated to treatment regimens on enrolment to the trial. The trials were of varying sizes, but it is likely that in the larger trials any baseline risk factors would have been approximately evenly distributed between treatment regimens. Adjusting for baseline risk factors should not therefore greatly affect the results.

Section 3.5.1 reviews the literature identifying baseline risk factors. The baseline patient-level covariates that are recorded in these trials and therefore available in these data are evaluated as risk factors for poor outcome in section 4.5.2, see in particular Table 4.9 showing odds ratios and 95% confidence intervals for each risk factor as a predictor of poor outcome. In this analysis, each of weight, age, sex, isoniazid resistance, baseline sputum smear, baseline sputum culture, extent of cavitation and extent of disease were found individ-

176

ually to be important risk factors for poor outcome. These will therefore be included in this section as possible covariates to be included in stage I of the model.

It is known that HIV infection and rifampicin resistance are risk factors for poor outcome, but these trials were conducted in the pre-HIV era and before the widespread use of rifampicin to treat TB. It can therefore be assumed that all patients in these trials were HIV negative and rifampicin susceptible.

6.4.5.2 Methods

The models involved in the two stages are described in section 6.4.3 above. In stage I, patient-level covariates are included in the model using a backward step-wise selection method to select those that are important risk factors for poor outcome in each particular trial. All of the covariates found to be important in Chapter 5 are included either as continuous variables (weight and age), binary variables (sex, isoniazid resistance, streptomycin resistance) or categorical variables (baseline sputum smear, baseline sputum culture, radiographic extent of cavitation and radiographic extent of disease).The backward step-wise selection method used to identify the important risk factors, is repeated for each trial, and proceeds as follows:

1. Calculate the increase in the log likelihood ratio statistic in removing a single covariate from the logistic regression model including all covariates, for each of the risk factors.

2. Identify the covariate for which the removal of this covariate results in the smallest increase in the log likelihood ratio statistic that is not statistically significant at the 5% level when compared to the χ^2 distribution with one degree of freedom.

3. Repeat from step 1 with this covariate removed from the model. If the removal of each covariate individually results in an increase in the log likelihood ratio statistic that is statistically significant (showing that every covariate is important as a risk factor in the model), go to step 4 for the beginning of the forward step-wise selection procedure.

4. Calculate the decrease in the log likelihood ratio statistic in individually adding each of the covariates not already included in the model.

5. If the decrease is statistically significant after the addition of any of these covariates, include these in the model and repeat from step 1. Otherwise, the selection process in finished.

Only those covariates statistically significant as risk factors for poor outcome for that particular trial are then retained and are included in both models at stage I so that the α_{ij} and the β_{ij} are adjusted for the same set of covariates for a particular trial i. Stage II then proceeds as before.

(a) Month 1, dichotomised at 20+

(b) Month 2

(c) Month 3
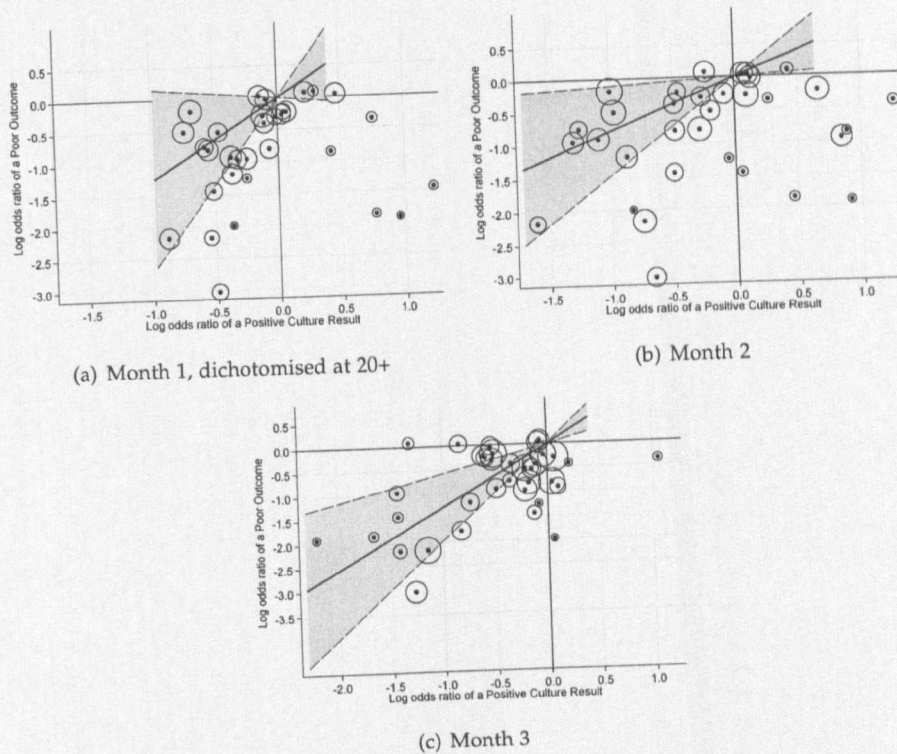
Figure 6.4: Logs odds ratio of a poor outcome plotted against log odds ratio of a positive culture. Logs odds ratio are adjusted for certain baseline risk factors. See text for details. Fitted line is weighted by the precision of the estimates, and this precision is represented by the diameter of the circles around each point. A 95% confidence interval on the slope is also shown.

6.4.5.3 Results

Figure 6.4 is a plot for each of the candidate surrogates, of the log odds ratio of a poor outcome (β_{ij}) against the log odds ratio of a positive culture (α_{ij})

where both log odds ratios are adjusted for a number of baseline risk factors. The slopes with 95% confidence intervals and the proportion of variation in the adjusted log odds ratio of a poor outcome explained by the log odds ratio of a positive culture for each marker is given in Table 6.6.

Marker	Number of Comparisons	Number of Trials	Slope, κ	95% CI	R^2
Month 1[†]	32	9	1.26	(-0.17,2.68)	0.32
Month 2	33	9	0.78	(0.10, 1.46)	0.31
Month 3	35	11	1.27	(0.57,1.98)	0.61

[†]With a point of dichotomy at 20+.

Table 6.6: Results of stage II of this analysis for each of the three candidate surrogate markers. Odds ratios calculated in stage I are adjusted for certain baseline risk factors. See text for details.

It is clear from Table 6.6 that the values of R^2 are lower on adjusting for baseline risk factors than those in the unadjusted model (Table 6.5 on page 174) for each of the candidate surrogates, and that the slopes of the fitted lines are marginally shallower in each case. One interesting difference from the unadjusted graph is that there is a slightly greater spread of points in each graph in the horizontal and vertical direction, and this is seen most in the Figure 6.4(b). What this suggests is that, as with (Davies et al., 2006a), on adjusting for important baseline covariates, some of the unexplained noise is removed and there is greater discrimination between treatments (shown by larger log odds ratios of both a poor outcome and a positive culture). Nevertheless, this doesn't greatly affect the relationship between the treatment effect on a poor outcome and the treatment effect on culture positivity, and in fact the values of R^2 are slightly reduced.

6.4.6 Subgroup Analyses

The twelve trials were conducted across three separate geographical regions (East Africa, Hong Kong and Singapore) and evaluated a variety of different regimens. All regimens were given for six months, but various combinations of different drugs were given for differing periods of time across the intensive and the continuation phase with differing weekly dosing schedules. Therefore, the three candidate surrogate markers were evaluated in this chapter across a wide variety of treatment comparisons.

(a) East Africa, month 1, dichotomised at 20+   (b) Hong Kong, month 1, dichotomised at 20+

(c) East Africa, month 2   (d) Hong Kong, month 2

(e) East Africa, month 3   (f) Hong Kong, month 3

Figure 6.5: Logs odds ratio of a poor outcome plotted against log odds ratio of a positive culture for trials from East Africa and Hong Kong. Fitted line is weighted by the precision of the estimates, and this precision is represented by the diameter of the circles around each point. A 95% confidence interval on the slope is also shown.

6.4.6.1 Analysis by Geographical Region

In Chapter 4, it was shown that the age and sex distribution differed between East Africa, Hong Kong and Singapore in addition to the severity of disease (as shown by baseline cultures and baseline radiographic extent of disease and extent of cavitation) in patients presenting for enrolment into the trials. On this basis (in addition to further evidence to suggest response to treatment might differ by geographical region (Fox et al., 1999)), the analysis described above was repeated separately for patients in East African trials, trials conducted in Hong Kong and in trials conducted in Singapore.

Figure 6.5 on the preceding page shows meta-analysis plots for each of the three candidate markers for trials from East Africa and trials from Hong Kong with the fitted line plotted. Figure 6.6 on the next page shows plots for the three candidate markers for trials from Singapore. Slopes with 95% confidence intervals and R^2 values are given in Table 6.7 on page 183. These graphs show very different results between the three geographical regions.

At months 1 and 2, the graphs restricted to data from East Africa (13 treatment comparisons in each analysis) shows very great variation about the line, $R^2 = 0.29$ and $R^2 = 0.19$ in each case and very wide confidence intervals (Figures 6.5(a) and 6.5(c)). A number of points lie outside and below the 95% confidence intervals on the slope (which themselves are very wide) indicating that these markers are failing to capture the treatment effect on a poor outcome in East African trials. This contrasts with the month 3 culture result in East Africa (evaluated across 16 treatment comparisons), with a clear linear trend in the points ($R^2 = 0.81$, Figure 6.5(e)).

In contrast to the East African graphs, a linear trend is more apparent for all months in the graphs restricted to data from Hong Kong only (15 treatment comparisons for each analysis). At month 1, there is less spread in the horizontal direction, as before, and the proportion of explained variation is reasonably high, $R^2 = 0.69$ (Figure 6.5(b)). The width of the confidence interval is also very wide. The best fit is at month 2 with a narrow 95% confidence interval around the slope, a high proportion of explained variation at $R^2 = 0.86$ (figure 6.5(d)), and no points outside the lower left quadrant, except for three which are very close to the origin. At month 3, the fit is not as good ($R^2 = 0.62$, figure 6.5(f)). The width of the 95% confidence interval around the slope is wider and there is more variation about the fitted line, although there are still no points outside the lower left quadrant.

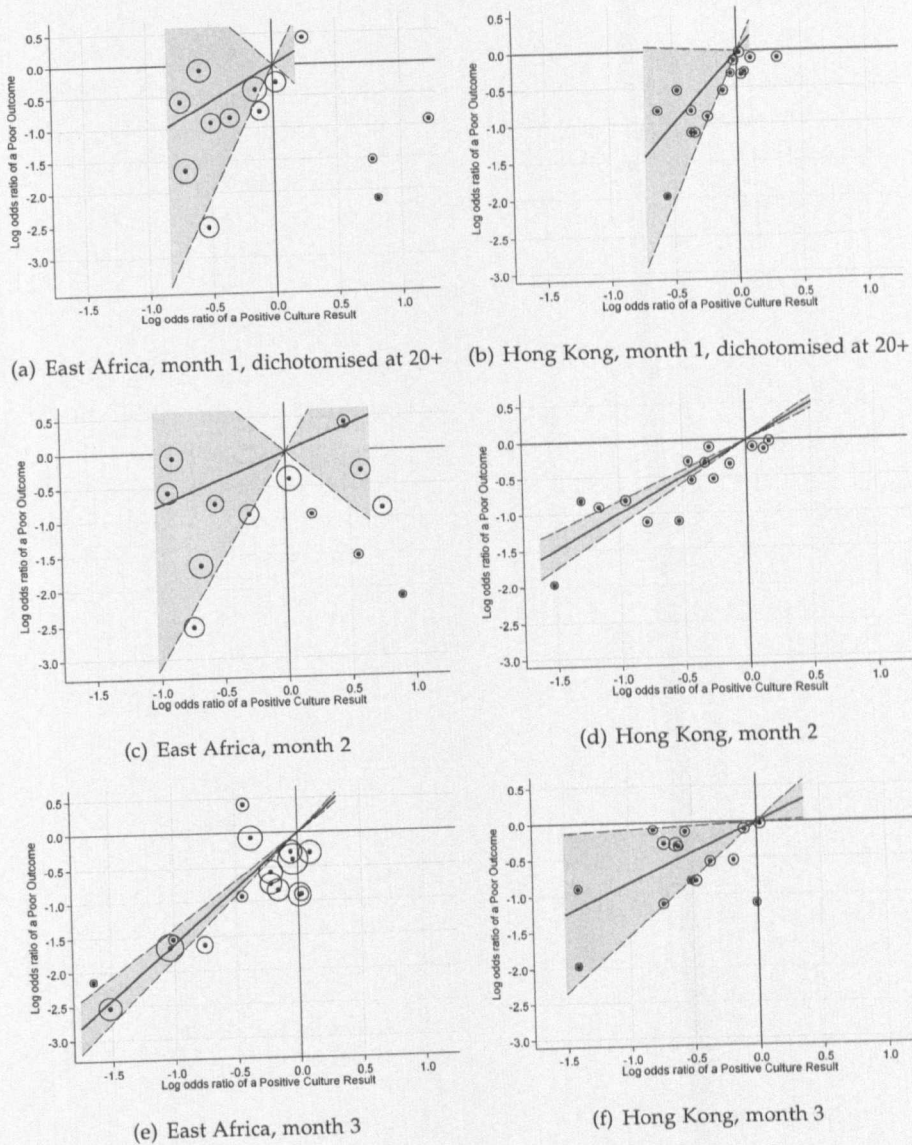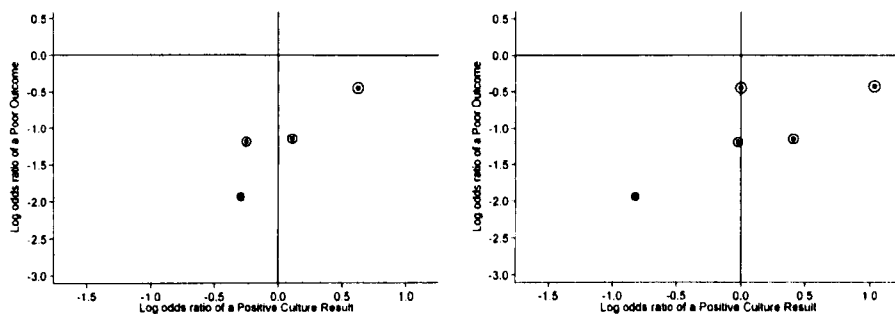(a) Singapore, month 1, dichotomised at 20+

(b) Singapore, month 2
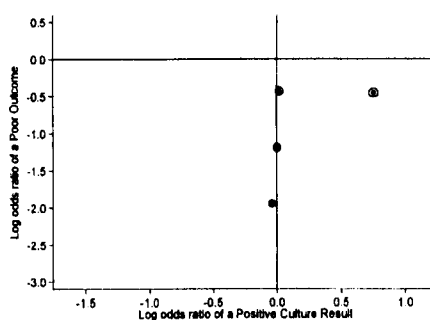
(c) Singapore, month 3

Figure 6.6: Logs odds ratio of a poor outcome plotted against log odds ratio of a positive culture for trials from Singapore. Precision is represented by the diameter of the circles around each point.

Fitting one model with different slopes for trials from Hong Kong and trials from East Africa, allows for testing for equality of slopes using a simple t-test. At month 1 and month 2, there was no evidence for a difference in slopes, p= 0.46 and p= 0.75 respectively. At month 3, there was a evidence for a difference in slopes between trials from Hong Kong and trials from East Africa, p= 0.015.

Between 13 and 16 treatment comparisons were used for evaluating each of the three candidate markers as surrogates in data restricted to East Africa trial and restricted to Hong Kong trials. No more than five treatment comparisons were available from Singapore trials for this meta-analysis, and all of these comparisons were from a single trial, the third Singapore trial (the first Singapore trial was not included for any analysis, see 6.1.2). The slope, 95% confidence interval and R^2 from the fitted line are shown in Table 6.7,

Region	Marker	Comparisons	Trials	Slope, κ	95% CI	R^2
East Africa	Month 1[†]	13	4	1.13	(-1.82,4.11)	0.29
	Month 2	13	4	0.76	(-1.57,3.09)	0.19
	Month 3	16	6	1.61	(1.38,1.83)	0.81
Hong Kong	Month 1[†]	15	4	1.98	(-0.92,4.05)	0.68
	Month 2	15	4	0.99	(0.82,1.16)	0.86
	Month 3	15	4	0.82	(0.09,1.56)	0.62
Singapore	Month 1[†]	4	1	0.19	(-4.95,5.33)[‡]	< 0.01
	Month 2	5	1	-0.06	(-2.31,2.20)[‡]	< 0.01
	Month 3	4	1	-0.52	(-5.09,4.05)[‡]	0.04

[†]With a point of dichotomy at 20+.

[‡]Robust standard errors were not used since the treatment comparisons were all from a single trial.

Table 6.7: Results of stage II of this analysis for each of the three candidate surrogate markers, by geographical region.

although the fitted line is not plotted as it is clear that a straight line through the origin is not appropriate for these data. Unlike all previous analysis, robust standard errors to account for the clustering within trial were not used as all the treatment comparisons were from the same trial.

6.4.6.2 Restriction to Regimens containing Rifampicin and Isoniazid

Rifampicin and isoniazid are the most important drugs in any combination regimen for the treatment of tuberculosis. It is for this reason that tuberculosis disease that is resistant to both isoniazid and rifampicin is identified almost as a separate disease (MDR-TB, see section 3.3.1) and requires a very different treatment strategy (World Health Organization, 2003). The trials contributing data to these analyses begin before and continued after the discovery of rifampicin. While all regimens in these trials contain isoniazid, not all contain rifampicin. Only 18 of the 37 treatment comparisons were of one rifampicin-containing regimen with another rifampicin-containing regimen. The meta-analysis was therefore repeated considering only those comparisons of regimens that *both* contained isoniazid *and* rifampicin throughout the six months of treatment. Figure 6.7 on the next page shows meta-analysis plots for each of the three candidate markers restricted to comparisons with both regimens containing isoniazid and rifampicin with the fitted line plotted. Slopes with 95% confidence intervals and R^2 values are given in Table 6.8 on the following

(a) Month 1, dichotomised at 20+

(b) Month 2

(c) Month 3
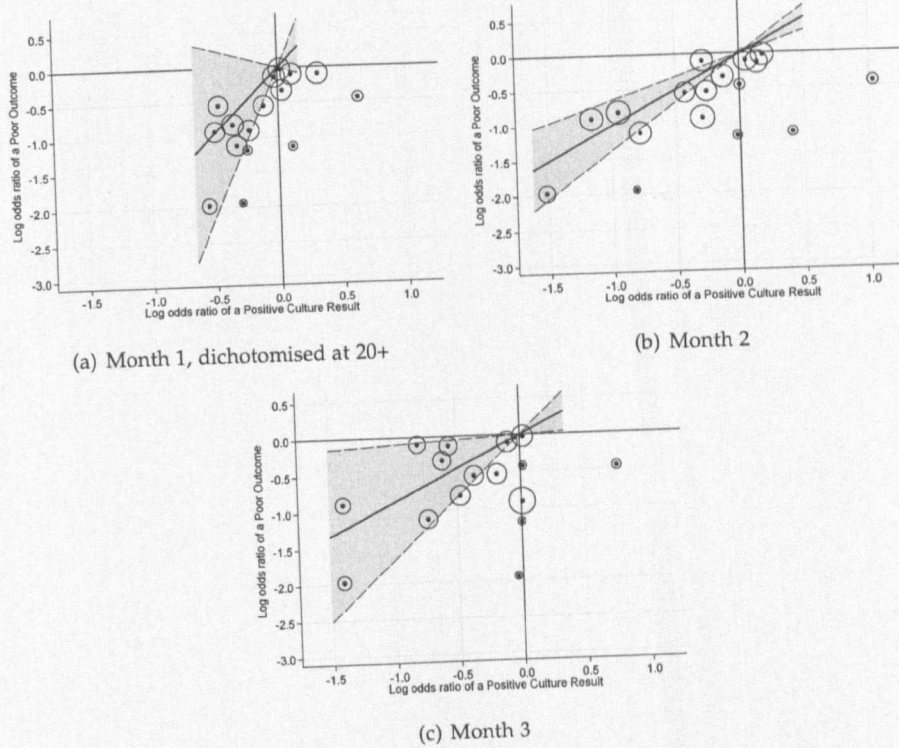
Figure 6.7: Logs odds ratio of a poor outcome plotted against log odds ratio of a positive culture, restricted to treatment comparisons for which both treatments contained isoniazid and rifampicin throughout the six month treatment period. Fitted line is weighted by the precision of the estimates, and this precision is represented by the diameter of the circles around each point. A 95% confidence interval on the slope is also shown.

page.

Comparing these graphs with those in Figure 6.3 on page 173, the results from all treatment comparisons, there are clear differences. On restricting treatment comparisons to rifampicin-containing only, the fit of the line for the month 1 and the month 2 cultures improves ($R^2 = 0.36$ to $R^2 = 0.54$ for the month 1 culture, and $R^2 = 0.36$ to $R^2 = 0.67$ for the month 2 culture) and there are fewer points in the lower right quadrant. For the month 3 culture, the fit of the line is not as good ($R^2 = 0.69$ to $R^2 = 0.46$) and poorer than for the month 2 culture with a wide 95% confidence interval on the slope. Considering only treatment comparisons of regimens containing isoniazid and rifampicin for the duration of treatment, the 2 month culture is superior to the 3 month

Marker	Number of Comparisons	Number of Trials	Estimated Slope, κ	95% CI	R^2
Month 1[†]	16	4	1.85	(-0.51,4.20)	0.54
Month 2	17	4	1.00	(0.64,1.36)	0.67
Month 3	16	4	0.88	(0.10,1.66)	0.46

[†]With a point of dichotomy at 20+.

Table 6.8: Results of stage II of this analysis for each of the three candidate surrogate markers, restricted to those treatment comparisons with both isoniazid and rifampicin given throughout both regimens in the comparison.

culture as a possible surrogate marker.

6.4.6.3 Further exploration

Figure 6.8 on the following page shows the meta-analysis plots for trials from East Africa and Hong Kong, evaluating the 2 and 3 month culture results as surrogate markers. 95% confidence intervals on the estimates of the log odds ratios are plotted as capped spikes on the plotted points. This plot allows for further assessment of the possible spread about the fitted line. In Figures 6.8(a) and 6.8(d), the confidence intervals cover a wide area indicating low precision in the estimate of the slope of the fitted line, as reflected in the low values of R^2 and the wide 95% confidence intervals around the estimates of the slopes. In Figure 6.8(b), the confidence intervals are all of a similar size and the overall spread do not detract from the linear pattern. In Figure 6.8(c), several of the points not near the origin have narrow confidence intervals, again not detracting from the overall linear pattern.

6.4.7 Discussion

Meta-analysis is essential in evaluating any marker as a surrogate endpoint. The testing of the Prentice criteria in section 6.2 was a necessary first step and the calculation of the single trial measures in section 6.3 was useful for data exploration, but it is the results of this section that will determine the value as surrogates of the three candidate markers being evaluated in this chapter.

(a) East Africa, month 2

(b) Hong Kong, month 2

(c) East Africa, month 3

(d) Hong Kong, month 3
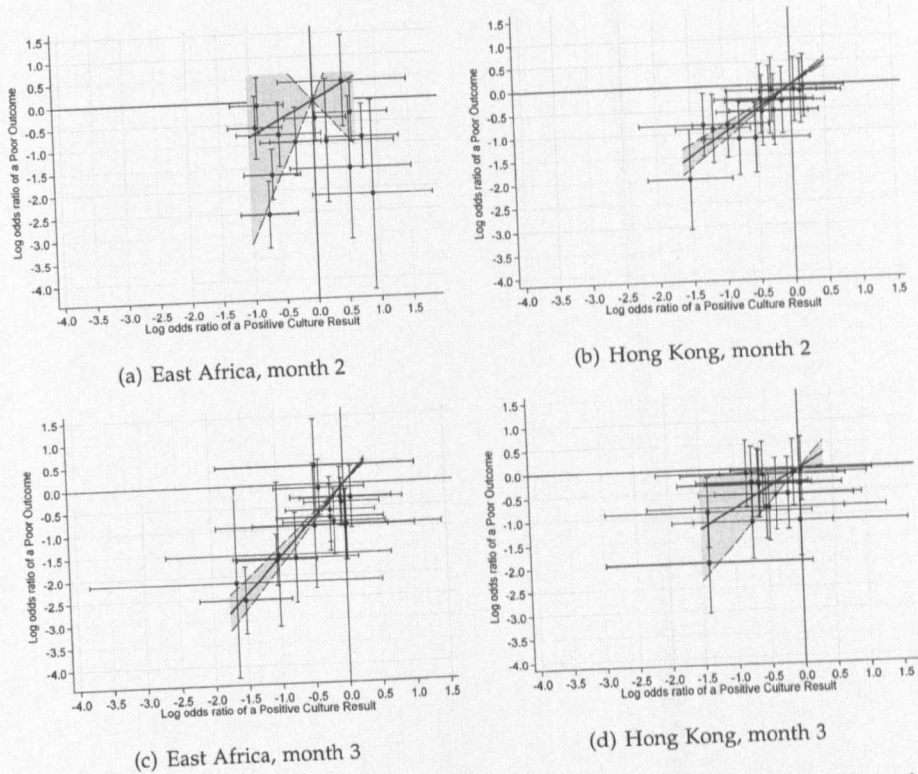
Figure 6.8: Logs odds ratio of a poor outcome plotted against log odds ratio of a positive culture for trials from East Africa and Hong Kong. Fitted line is weighted by the precision of the estimates, and this precision is represented by the diameter of the circles around each point. A 95% confidence interval on the slope and 95% confidence intervals of the logs odds ratios are also shown.

6.4.7.1 Methods

As described above, the *HIV paradigm* was recast in a frequentist framework and methods extended to evaluate each of these three candidate markers as surrogates for poor outcome. These methods are flexible, allowing for different distributional assumptions for the surrogate and true endpoints. Separating the patient-level and treatment comparison-level analyses into two stages also allows for the inclusion of patient-level covariates in stage I and for a clear graphical representation of the results of stage II. The precision of the estimates of α_{ij} and β_{ij} for each i and j is incorporated into stage II using a weighted linear regression giving greater weight to estimates made with greater precision (represented on the graphs with wider circles around these

186

points) and the clustering within trials is incorporated using robust standard errors. There are two drawbacks:

1. Using robust standard errors adjusts the standard errors of the parameters from the model for the clustering within trials, but the point estimates of the parameters are based on the assumption that each pair of points $(\alpha_{ij}, \beta_{ij})$ is assumed to be independent. This is an incorrect assumption, as treatment comparisons are clustered within trials, having common control regimens.

2. α_{ij} and β_{ij} are can only be estimated with error for each i and j. At stage II the error in β_{ij} is incorporated in the weighted regression, but the error in α_{ij} is not. The effect of wrongly assuming that the α_{ij} are estimated without error is *regression dilution bias*—there will be attenuation of the slope of the fitted line towards zero, and the slope will therefore be underestimated.

These are real drawbacks with the methods. Chapter 7 is devoted to extending these methods in an attempt to overcome these problems.

6.4.7.2 Results

Using data from a number of randomised clinical trials, Mitchison (1996) concluded that the 2 month culture was a better surrogate marker than the 3 month culture. As a result, it is commonly accepted that the 2 month culture is a valid surrogate and it is often described as such (see section 3.5.3). From these data, the overall analyses showed that both the 2 month culture and the 1 month culture dichotomised at 20+ are poor surrogate markers. In contrast, the 3 month culture is possibly a useful surrogate, although there is considerable variation about the fitted line.

6.4.7.2.1 Month 1 It is clear that, based on the results of the analysis, a heavily positive culture at month 1 can be used to give an indication of the direction of the effect of a treatment on a poor outcome, but given the low R^2, the 95% confidence interval on the slope containing 0, the grouping around the vertical axis and the points in the lower right quadrant, *a heavily positive culture at month 1 cannot be considered as a useful surrogate marker.*

6.4.7.2.2 Month 2 As with the heavily positive culture at month 1, the culture result at month 2 can be used to give an indication of the direction of the effect of a treatment on a poor outcome. Since the spread of the points in the horizontal direction is greater and the width of the 95% confidence interval on the slope is narrower, the culture result at month 2 can also be used to give a very rough indication of the size of the treatment effect on a poor outcome. Nevertheless, the low R^2 in addition to the large number of points below the 95% confidence band on the fitted line and in the lower right quadrant showing that there is a proportion of the treatment effect on a poor outcome that is not captured by the month 2 culture result show that, on the basis of this analysis of these data, *the month 2 culture result cannot be used as a surrogate marker.*

The points in the lower right quadrant could be explained by causal pathways of the disease that bypass the surrogate, but are acted on by the treatments in these comparisons.

6.4.7.2.3 Month 3 These data suggest that a larger proportion of the treatment effect is captured by the month 3 culture than by either a heavily positive culture at month 1 or a positive culture at month 2. While there is little guidance on what constitutes a good surrogate marker, Burzykowski et al. (2005) identify a surrogate for which $R^2_{trial} = 0.692$ (resulting from a slightly different analysis to that presented here) as a "moderate" surrogate endpoint. While R^2 is 0.69, there are points that do not lie near to the line, particularly the grouping around both the vertical and the horizontal axes close to the origin.

Therefore, on this analysis of these data, *the month 3 culture result could be considered to be a possible surrogate marker, but it is far from perfect.*

Looking at figure 6.3(c) on page 173, it is clear that a large treatment effect on a the 3 month culture result does not necessarily correspond to a large treatment effect on a poor outcome. Similarly, a small treatment effect on the 3 month culture result does not necessarily correspond to a small treatment effect on a poor outcome, although this is usually the case. There is far from sufficient precision, therefore, to predict the treatment effect on a poor outcome based solely on the treatment effect on the 3 month culture result and therefore cannot be used strictly as a surrogate endpoint in a phase III trial as a substitute for the final endpoint of treatment outcome.

6.4.7.2.4 Adjusting for baseline covariates In section 6.4.5, baseline covariates were included in the models in stage I to give treatment effects adjusted by important risk factors taken forward to stage II. Adjusting for baseline risk factors did marginally increase the spread in the horizontal direction suggesting that the culture results adjusted for these risk factors did discriminate better between regimens. Nevertheless, the fit of the line does not improve. The most important risk factor found by Davies et al. (2006a) in evaluating aspects of the SSCC profile as surrogate markers was HIV status which is not available in these data as the trials were conducted in the pre-HIV era. What the results in this chapter have shown is that adjusting for baseline risk factors other than HIV status (and rifampicin resistance, another covariate commonly found to be an important risk factor) do not result in any of these three candidate markers being shown to be better surrogate markers. Further discussion will therefore consider the unadjusted rather than the adjusted analysis.

6.4.7.2.5 Subgroup Analyses In East African trials, the month 3 culture was a better surrogate than the month 2 culture, but in Hong Kong trials the month 2 culture was found to be superior to the month 3 culture. For both of these markers (the 3 month culture in East African trials and the 2 month culture in Hong Kong trials), the fit was better than in the analysis on the complete data ($R^2 = 0.81$ and $R^2 = 0.89$ respectively compared to $R^2 = 0.69$ and $R^2 = 0.36$ respectively).

There are two possible reasons for this discrepancy: *differences in culture conversion rates* and *the proportion of rifampicin-containing regimens.*

Table A.1 on page 248 shows the proportion of participants that are culture positive at each monthly, separated by geographical region. Culture conversion tends to be later in East African trials (29% culture positive at month 2 and 10% still culture positive at month 3) than in the Hong Kong trials (15% culture positive at month 2 and only 4% still culture positive at month 3). This could be the reason why the 3 month culture is a superior surrogate marker in East African trials and the 2 month culture is superior in Hong Kong trials. This suggests that a marker reflecting the time-to-conversion or a marker resulting from longitudinal modelling on culture results during treatment (requiring more timepoints than available in these data) could be a more useful surrogate.

In only rifampicin containing regimens, the month 2 culture was found to be a better surrogate than the 3 month culture. One of the difficulties in

189

interpreting these results is the relationship, in these data, between the region that the trial was conducted and the regimens included in the trials. Partly because the East African trials were earlier in time than the Hong Kong and Singapore trials, and partly that the trials in Hong Kong and Singapore were less restricted by the high cost of rifampicin, trials in East Asia evaluated treatment arms that contained rifampicin throughout more often that trials in East Africa. These studies in East Asia also concentrated more on intermittent regimens which were more easily supervised and therefore more likely to be effective in the urban settings of Hong Kong and Singapore (Fox et al., 1999). The subset of treatment comparisons from trials conducted in East Asia is therefore almost the same as the subset of treatment comparisons for which both regimens contained both isoniazid and rifampicin for six months and also the subset of treatment comparisons for which both regimens involved daily dosing throughout (see Table 6.9).

Geographical Region	Isoniazid-containing	Rifampicin-containing	Daily Regimens	Total
East Africa	16 (100%)	1 (6%)	14 (88%)	16
Hong Kong	15 (100%)	11 (73%)	0 (0%)	15
Singapore	6 (100%)	6 (100%)	1 (17%)	6
Total	37 (100%)	18 (49%)	15 (41%)	37

Table 6.9: Number of treatment comparisons tabulated by geographical region and details of regimens in comparison. For each column, a treatment comparison was included if both regimens in the comparison met the criteria.

Rifampicin and isoniazid are vital for the treatment of tuberculosis and therefore it is not surprising that there might be differences in the relationship between culture results during treatment and treatment outcome. It is likely therefore that the differences between the 2 and the 3 month culture results as surrogates between East Africa and Hong Kong might be due in part to whether the regimens contained rifampicin.

6.4.7.2.6 Singapore The single treatment comparison from the first Singapore trial was not included in any analysis (see 6.1.2). This was because the first two months of treatment were exactly the same in both regimens meaning that it could not be used for evaluating either the month 1 or month 2 culture result as a surrogate marker for poor outcome. There were too few positive cultures at month 3 to estimate the treatment effect on culture positivity at

month 3 and therefore this comparison was not included in the evaluation of the month 3 culture result. Therefore, comparisons from only one trial conducted in Singapore, the third Singapore trial, were included in these analyses. It is clear from Figure 6.6 on page 182 that none of the three candidate markers of culture results at months 1,2 or 3 could be used as surrogate markers on the basis of data from this Singapore trial. While there are clear differences between rates of poor outcome on the regimens in this trial, these differences are not consistently reflected in cultures results as the points on this graph are spread either side of the vertical axis. These results are from four or five comparisons from one trial, and therefore the data are too few to draw any strong conclusions from these data, although it would be expected that the results would be similar to those in the Hong Kong studies.

This trial was unique among the twelve trials included in these analyses in three of the six arms included a combined tablet of isoniazid, rifampicin and pyrazinamide to compare patient acceptability, occurrence of adverse effects and efficacy with the same combination given as separate tablets (Singapore Tuberculosis Service/British Medical Research Council, 1991). In all other trials used in these analyses, drugs given in combination were given as separate tablets (excepting some patients in the fourth Hong Kong study, see section 4.1.1.3.1). It had been previously shown that the bioavailability of the three drugs in this combination was satisfactory (Ellard et al., 1986). It appears therefore, that a combined tablet has modes of action on the true endpoint (poor outcome) that is not captured by the surrogate marker (culture results during treatment) compared to a combination of the same drugs given separately. The results of the United States Public Health Service study 21 support this; a combined formulation was associated with a sputum conversion rate more rapid than the separate formulation, but with a similar low recurrence rate (Combs et al., 1990).

Whatever the reason, it appears that the use of a combined formulation might alter that relationship between culture results and poor outcome, though these results are based on one trial and are therefore inconclusive. This fact should be considered in future surrogate marker studies in the treatment of TB. It would be of particular interest to repeat these analyses on data from the multi-centre trial currently being conducted by the IUATLD (Study C, preliminary results recently presented at the IUATLD annual conference in October 2008 (Lienhardt et al., 2008)) comparing a combined formulation with a regimen of drugs given separately in sites in Africa, Asia and South America.

6.5 Comparison with Recent Trial Data

6.5.1 Introduction

As described in Chapter 4, all the data used for the analyses in this and previous chapters are from MRC trials conducted in East Africa and East Asia in the 1970s and 1980s, with the last trial starting enrolment in 1983. If one of the candidate markers evaluated in this thesis is accepted as a surrogate, it will be used in future trials that will be conducted nearly thirty years after the collection of the data on which the surrogate was evaluated. The advent of HIV and the spread of MDR-TB are the known changes in the TB epidemic that have occurred in this period, but it is likely that there will have been other unknown biological or pathogenic changes occurring.

For this reason, the possibility of additional clinical trial data was explored to determine whether the performance of the surrogate was different in recent trial data compared to the older data. Following the closure of the MRC TB units in 1986, very few large phase III clinical trials have been conducted that had outcome to treatment after a period of at least twelve months of follow-up as the clinical endpoint. However, data was provided by the CDC TB Trials Consortium (TBTC) from *Study 22* and by the International Union Against TB and Lung Disease (IUATLD) clinical trials team from *Study A* for use in this thesis.

6.5.2 Description of the Trial Data

6.5.2.1 US Public Health Study 22

The 22nd TB study conducted by the US Department of Public Health (leading to the formation of the CDC TB Trials Consortium (TBTC)) was a multi-centre, open-label, randomised controlled trial comparing rifapentine and isoniazid given once-weekly and rifampicin isoniazid given twice-weekly during the continuation phase (Benator et al., 2002). Patients with pulmonary TB who had completed two months of the intensive phase of the standard regimen were enrolled in the USA and Canada and randomly allocated to one of two treatment arms and assessed for failure during treatment and for bacteriological evidence of relapse every three months for twenty-four months following the end of treatment.

It was a requirement that all patients had had isoniazid, rifampicin, pyraz-

inamide and either streptomycin or ethambutol in the two-month intensive phase. The first two weeks of the intensive phase was given daily and then daily, twice-weekly or thrice-weekly thereafter. All patients effectively had the same intensive phase (with only slight variations) and the treatment differed only in the continuation phase.

Between April 1995 and November 1998, 1004 HIV-negative patients were enrolled, of whom 502 were allocated to the once-weekly regimen and 502 allocated to the twice-weekly regimen. Follow-up for the last patients was completed in March, 2001. 71 HIV-positive patients were also enrolled and analysed separately (Vernon et al., 1999).

Most of the cultures were done on liquid media, with the remainder being done on solid media. A positive culture result at two months, for example, is an indication that the patient had a positive culture either on liquid media or on solid media. Relapse and treatment failure were established on a positive culture on either solid or liquid media.

Unfortunately, since the study sites with in the USA and Canada where the rates of HIV are low, the number of HIV-positive patients in this study was small, only 71 (7%). Nevertheless, patients with HIV were included in the analysis. No trial participants had pretreatment rifampicin resistance and therefore there were no cases of MDR-TB. 497 patients allocated to isoniazid and rifapentine daily and 494 allocated to isoniazid and rifampicin twice-weekly in the continuation phase were included for the analysis (after excluding 31 with only three drugs in the intensive phase and 53 with extrapulmonary TB).

6.5.2.2 IUATLD Study A

The first TB study conducted by the International Union against TB and Lung Disease (IUATLD) was an open-label, randomised controlled clinical trial conducted in sites across Africa and Asia. Patients were allocated to either (i) the standard six month regimen, 2EHRZ /4HR; (ii) the WHO-recommended eight month regimen without rifampicin in the continuation phase, 2EHRZ /6HE; or (iii) the same regimen with the intensive phase given thrice-weekly, 2(EHRZ)₃ /6HE. Patients were assessed at the end of treatment and sputum samples taken at 3, 6 and 12 months after the end of treatment to assess outcome of treatment.

Between March, 1998 and December, 2001, 1335 patients were randomised

to one of the three treatment regimens. 466 patients allocated to 2(EHRZ)₃ /6HE, 456 allocated to 2EHRZ /6HE and 433 allocated to the control regimen 2EHRZ /4HR were included in the analysis. This included 127 (10%) who were HIV positive and 25 (1.9%) with rifampicin resistance, of which 20 (1.5%) also had isoniazid resistance and were therefore cases of MDR-TB.

6.5.3 Methods

The meta-analysis described in section 6.4.3 was repeated including these data from Study 22 (one treatment comparison) and Study A (two treatment comparisons). Following the principles described in section 6.1.1, the regimen with the least efficacy, that with the highest proportion of poor outcomes, was identified as the 'control' regimen. This was the once-weekly rifapentine arm in Study 22 and the eight-month thrice-weekly regimen in Study A.

In Study A, sputum samples were only taken at two months after the start of treatment and therefore only the two month culture is available in these data to evaluate as a surrogate marker. The third regimen had a different intensive phase (EHRZ was given thrice weekly) and was therefore selected as the control. This was so that the two regimens denoted as *experimental* had the first two months of treatment different to the control.

In Study 22, only the three month culture will be evaluated as a surrogate marker. Sputum samples were not taken at one month during treatment and therefore only the two month and the three month cultures are available in these data to evaluate as surrogate markers. Since Study 22 was designed to compare two regimens with different continuation phases, patients were enrolled after they had completed effectively the same intensive phase of the standard regimen. The 2 month culture result could not therefore capture any treatment difference in poor outcome and cannot therefore be a surrogate in this treatment comparison.

Therefore, the data from Study A can be included with the rest of the data used in this thesis and the meta-analysis repeated to evaluate the *two month culture result as a surrogate marker*, and the data from Study 22 can be included with the rest of the data used in this thesis and the meta-analysis repeated to evaluate to *three month culture result as a surrogate marker*.

Trial	Treatment Comparison	α_{ij}		β_{ij}	
		Estimate	95% CI	Estimate	95% CI
Evaluating the 2 month culture					
Study A	1	-0.61	(-0.96, -0.26)	-0.31	(-0.77, 0.15)
	2	-0.37	(-0.71, -0.03)	-1.11	(-1.68, -0.54)
Evaluating the 3 month culture					
Study 22	1	-0.10	(-0.80, 0.60)	-0.36	(-0.77, -0.05)

Table 6.10: Results of stage I of the meta-analysis for the data from Study A and Study 22 evaluating the month 2 and the month 3 culture results as surrogates.

6.5.4 Results

Table 6.10 shows the log odds ratios of poor outcome, β_{ij}, with 95% confidence intervals and log odds ratios of 2 or 3 month culture results, α_{ij}, with 95% confidence intervals for the treatment comparisons from Study A and Study 22. These are plotted together with the results from the rest of the data (section 6.4.4) to explore how the strength of the 2 month and the 3 month culture results as surrogate markers changes after the inclusion of more recent data.
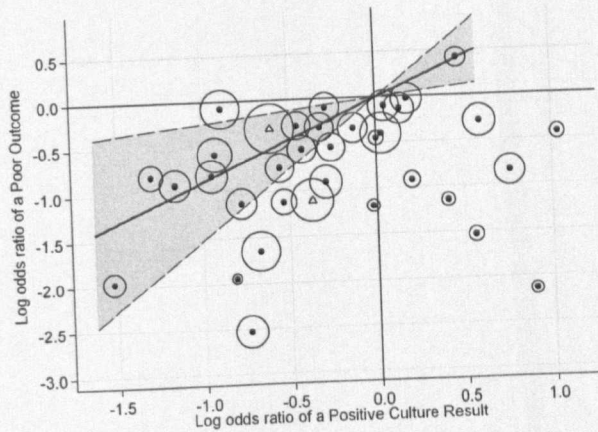
Figure 6.9: Log odds ratio of a poor outcome plotted against log odds ratio of a two month positive culture result. The triangles correspond to the two treatment comparisons from Study A.

Figure 6.9 shows the plot of the data evaluating the 2 month culture as a surrogate marker with the points representing the treatment comparisons from Study A plotted as triangles, and Figure 6.10 on the following page the
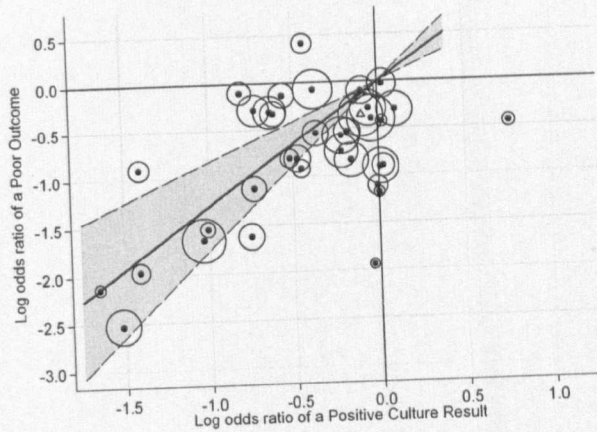
Figure 6.10: Log odds ratio of a poor outcome plotted against log odds ratio of a three month positive culture result. The triangle corresponds to the treatment comparison from Study 22.

plot of the data evaluating the 3 month culture as a surrogate marker with the point representing the treatment comparison from Study 22 plotted as a triangle.

Marker	Number of Comparisons	Number of Trials	Estimated Slope, κ	95% CI	R^2
Month 2	35	10	0.87	(0.24,1.51)	0.37
Month 3	36	12	1.29	(0.83,1.76)	0.69

Table 6.11: Results of stage II of this analysis for each of the three candidate surrogate markers, restricted to those treatment comparisons with both isoniazid and rifampicin given throughout both regimens in the comparison. α_{ij} is the log odds ratio of a positive culture result and β_{ij} is the log odds ratio of a poor outcome.

Table 6.11 shows the slope with 95% confidence interval and proportion of variation explained, R^2, for the fitted line in each graph. Comparing table 6.11 with table 6.5 on page 174, the slopes and the proportion of explained variation remain largely unchanged. The two points representing the treatment comparisons from Study A in figure 6.9 and the point representing the treatment comparison from Study 22 in figure 6.10 are largely concordant with the rest of the data and lie close to the line. This is reflected as the proportions of explained variation in table 6.11 are unchanged from that reported in 6.5.

6.5.5 Discussion

Unfortunately very little clinical trial data are available from the last ten years. Only two treatment comparisons from one trial were available to evaluate the two month culture result as a surrogate marker and one treatment comparison from a second trial was available to evaluate the three month culture result as a surrogate marker. This is one difficulty in the evaluation of surrogate markers for poor outcome to treatment for TB, that most of the data available is thirty years old or more. Work is currently being undertaken to produce standards for TB clinical trial data with the aim to promote sharing of data from future clinical trials[1]. Nevertheless, the two month culture result comparison in the two treatment comparisons from Study A and the three month culture result comparison in the single treatment comparison from Study 22 were not inconsistent with the results from the historical data. The treatment effect on the culture result and the treatment effect on poor outcome were in the same direction for all three treatment comparisons and the plotted points lay close to the fitted lines.

Unlike the main data used in the analyses in this thesis, a small number of patients with HIV or pretreatment rifampicin resistance were included in these recent data. It is encouraging that, even with these patient included in the analysis, the results were comparable with those from historical data, although numbers were too small to explore how HIV and rifampicin resistance affect the use of monthly culture results as surrogate markers.

In Study 22, a positive culture at month 3 was defined as a positive culture on either solid or liquid media, unlike culture result in all other studies which were on solid media only. Despite this difference, these results from Study 22 were not inconsistent with those from the rest of the data.

6.6 Discussion and Conclusion

6.6.1 Summary

In Chapter 5, culture results at each of months 1, 2, 3 and 4 were evaluated as prognostic markers for poor outcome, exploring the effect of varying the point of dichotomy. Three markers, the two month culture result, the one

---

[1]CDISC Cardiovascular and Tuberculosis Data Standards http://www.cdisc.org /standards/cardio/index.html Retrieved 23 Apr 2009.

month culture result dichotomised at 20+ and the three month culture result, were then selected for evaluation as surrogate markers with results presented in this chapter.

Prentice laid the framework for surrogate marker evaluation in 1989 with his criteria (Prentice, 1989), an article that has since received nearly 500 citations to date[2]. It was shown in Chapter 5 that *all* of culture results at months 1, 2, 3 or 4 whatever the point of dichotomy were strongly associated with poor outcome (Figure 5.2 on page 136). It was subsequently shown in section 6.2 that, in the majority of treatment comparisons, each of the markers failed to satisfy the Prentice criteria. Of 13 treatment comparisons where the treatment had a statistically significant effect on poor outcome, the three month culture result did not satisfy the Prentice criteria in any. Similarly, of 12 treatment comparisons where the treatment had a statistically significant effect on poor outcome, the one month culture (dichotomised at 20+) did not satisfy the Prentice criteria in 11 (88%) and the two month culture did not satisfy the Prentice criteria in 9 (75%). None of the three markers satisfied the Prentice criteria in all treatment comparisons and therefore, on the basis of the Prentice criteria, none of these three markers can be considered to be surrogate markers.

The use of hypothesis testing in the Prentice criteria does mean that only failed surrogates can be identified; there is no framework for accepting a true surrogate (as failure to reject the null hypothesis could be due to insufficient information rather than it necessarily being correct). One way to incorporate this would be to use the approach developed for *equivalence trials*. The treatment effect could be shown to be fully captured by the surrogate when the 95% confidence interval of the treatment effect on the true endpoint adjusted for the surrogate endpoint lies completely within the interval $(-\theta, +\theta)$ for some sufficiently small θ. This could be included as an additional step if the Prentice criteria have been verified to determine whether the information available is sufficient to determine surrogacy or whether more data is required.

In section 6.3, single trial measures developed to estimate the proportion of treatment effect explained by the surrogate were used to evaluate the three candidate markers as surrogates. There was great variation between treatment comparisons with point estimates lying outside of the interval [0,1] and wide

---

confidence intervals sometimes containing the whole interval [0,1]. It is difficult to draw any conclusions from these results which expose the deficiencies in the methods rather than possible deficiencies in the markers.

In section 6.4 a two stage meta-analytic approach was developed to evaluate each of the three candidate markers as surrogates for poor outcome. It is clear from the literature review in Chapter 2 than meta-analytic methods are superior to single trial methods as the surrogate marker is evaluated across a variety of treatment comparisons rather than just one. The methods developed produce a useful graphical visualisation of the results that help demonstrate the performance of the surrogate in a clearly accessible way. These meta-analytic methods were applied to the data to determine the relationship between the treatment effect on the candidate surrogate endpoints and the treatment effect on poor outcome. There were clear differences between trials conducted in Hong Kong and trials conducted in East Africa. The two month culture result was a superior surrogate in Hong Kong trials and the three month culture was superior in East African trials. The results from Hong Kong trials were similar to those from treatment comparisons of regimens both containing rifampicin making it difficult to separate the effect of geographical region from the effect of rifampicin.

Data from two recent TB clinical trials were included with these data and the results presented in section 6.5. With only two additional treatment comparisons for evaluating the two month culture result as a surrogate and one for evaluating the three month culture result as a surrogate, the results were not inconsistent with those using only the older trials.

6.6.2 Conclusions

International guidelines covering Statistical Principles for Clinical Trials state that 'the strength of the evidence for surrogacy depends upon (i) the biological plausibility of the relationship, (ii) the demonstration in epidemiological studies of the prognostic value of the surrogate for the clinical outcome and (iii) evidence from clinical trials that treatment effects on the surrogate correspond to effects on the clinical outcome' (International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals For Human Use, 1998).

Culture results directly reflect the number of causative bacilli in a patient's sputum which in turn reflects the bacillary load in the patient's lung. There-

fore, there is good biological plausibility that culture results during treatment are likely to be useful surrogates for poor outcome to treatment.

The data used in this thesis have showed that culture results are not useful markers for predicting outcome for individual patients, but there is a good association with poor outcome at the population level. Using the Prentice criteria for evaluating the culture results at 1, 2 or 3 months did not give encouraging results, although some recent authors have been giving less weight to the criteria (e.g. Burzykowski, 2008).

The results from the meta-analysis, on the other hand, are more encouraging. In the East African trials, the month 3 culture result was shown to be a good surrogate ($R^2 = 0.81$), in the Hong Kong trials, the month 2 culture result was shown to be a good surrogate ($R^2 = 0.86$). It is likely that this discrepancy is due either to the higher proportion of rifampicin-containing comparisons in the Hong Kong trials or to the delayed culture conversion rates in the East African trials.

The quality of the surrogate can be seen from the graphical representation of the results. The methods used to calculate the fitted line and the proportion of explained variation, R^2, have several drawbacks which will be discussed and explored in the next chapter.

Further discussion of these results and areas for future research are found in Chapter 8.

Chapter 7

An Extension of the Two-Stage Modelling Approach

7.1 Introduction

In Chapter 6, three candidate markers, the culture result at month 1 dichotomised at 20+, the month 2 culture and the month 3 culture, were evaluated as surrogate markers. As well as testing the Prentice criteria and calculating single trial measures, the main part of the analysis was devoted to the two-stage meta-analytic analysis based loosely on the *HIV paradigm* (section 6.4). The results of these analyses showed the month 2 culture to be superior to the month 3 culture in Hong Kong studies, with the reverse true in East African studies. These analyses were straightforward to perform with no complicated modelling required beyond logistic regression and weighted linear regression with robust standard errors. Nevertheless, there were two obvious methodological deficiencies.

Firstly, it was assumed that the treatment effect on the surrogate endpoint, α_{ij}, was measured without error in the second stage of the analysis, whereas only the estimate, $\hat{\alpha}_{ij}$, is available. This overly simplistic assumption leads to two errors: (*i*) *attenuation bias* and (*ii*) over-estimating the precision of the estimates of the slope parameter, κ (Carroll and Stefanski, 1995). Therefore,

ignoring the variation in the explanatory variable ($\hat{\alpha}_{ij}$) causes attenuation of the estimate of the slope towards zero and over-precise parameter estimates.

Secondly, robust standard errors were used to correct for clustering within trial which is a more straightforward technique to use, but is a 'cruder' approach than using a random effects model. This robust standard errors approach for correcting for clustering only adjusts the standard errors of the parameters, the point estimates themselves are not adjusted for the clustering. A random effects model will give point estimates of the parameters, as well as the standard errors of these parameters, that better account for the hierarchical structure, and it is therefore a preferable approach, although the alternative may be considered to be sufficient when the between trial correlation is not of interest (as is the case in these data).

The *Belgian paradigm* (see section 6.4.1) does account for the error in estimating both the treatment effect on the true endpoint as well as the treatment effect on the surrogate endpoint. However, those authors involved in the development of the Belgian paradigm used a simpler situation of two arm studies and have not considered the case of multi-arm trials with multiple treatment comparisons within a trial. Therefore these methods have not yet been extended to account for the clustering of treatment comparisons within trials (this is an area of future research for the group based at the Center for Statistics, Universiteit Hasselt, Diepenbeek, Belgium[1]). It is also the case that these methods were developed for Gaussian true and surrogate endpoints. For binary or categorical true or surrogate endpoints, the methods are not straightforward and there are sometimes reported problems with convergence.

In this chapter, the meta-analytic methods used in Chapter 6 will be extended in an attempt to overcome these problems, combining some ideas from the Belgian paradigm.

In section 7.2 the model used in Chapter 6 is developed, introducing a simplification of the Belgian paradigm proposed by Tibaldi et al. (2003) and extending this to non-normal true and surrogate endpoints and to multi-arm trials. The properties of two new approaches are explored using a simulation study and compared with that used in Chapter 6 in section 7.3. These are then applied to the trial data in section 7.4 and the chapter concludes with discussion in section 7.5.

---

[1] Personal communication from Professor Geert Molenberghs, Director of the Center for Statistics at Hasselt University.

7.2 Model Development

7.2.1 Stijnen's Approach

7.2.1.1 Model Detail

Within the Belgian paradigm, Tibaldi et al. (2003) note that even when both the true and surrogate endpoints are normally distributed, the fitting of the linear mixed models involved often turns out to be 'surprisingly difficult'. They propose a number of simplifications along what they describe as three *dimensions*—the trial, the endpoint and the measurement error dimensions.

The *Trial Dimension* relates to whether the effect of trial differences are treated as fixed or random effects; the *Endpoint Dimension* relates to whether the true and the surrogate endpoints are modelled jointly or separately (in a manner similar to the HIV paradigm); and the *Measurement Error Dimension* relates to how measurement error in the treatment effects is accounted for. It is important to note that these authors are not considering multi-arm trials, they only have two-arm trials in mind and therefore the problem of clustering within a trial is not addressed.

Taking the case when the true and surrogate endpoints (represented by T and S respectively) are modelled separately and therefore mixed effects are unnecessary (thus determining the trial and the endpoint dimension) the following models are used in the first stage:

$$S_{ik} = \mu_i + \alpha_i Z_{ik} + \varepsilon_{S_{ik}} \tag{7.1}$$

$$T_{ik} = \nu_i + \beta_i Z_{ik} + \varepsilon_{T_{ik}}, \tag{7.2}$$

where μ_i and ν_i, are the trial-specific intercepts and α_i and β_i the slope (treatment effect) parameters, i denotes the trial and k denotes the trial participant using notation consistent within this thesis. $\varepsilon_{S_{ik}}$ and $\varepsilon_{T_{ik}}$ are the residual error terms assumed to be independent. In the second stage, the authors fit the following model:

$$\hat{\beta}_i = \lambda_0 + \lambda_1 \hat{\mu}_i + \lambda_2 \hat{\alpha}_i + \varepsilon_i. \tag{7.3}$$

The authors include the estimate of the intercept $\hat{\mu}_i$ in this model at the second stage. It is not clear that the estimate of the mean of the surrogate in the

control regimen ($\hat{\mu}_i$) can contribute information to predict the treatment effect on the true endpoint, but the authors have chosen to carry this term forward into this second stage. λ_0 and λ_2 are therefore similar to δ and κ defined in section 6.4.3, adjusted for the mean of the surrogate in the control regimen.

The authors suggest three choices for the measurement error dimension:

1. Fit the simple linear model assuming all of the error terms, ε_i, are independent and identically distributed with equal variance σ^2. This ignores the fact that $\hat{\alpha}_i$ and $\hat{\beta}_i$ are only estimates of α_i and β_i and will be estimated with precision varying with i (depending on trial size and distributions of S and T in each trial). This is therefore an unsatisfactory approach.

2. Fit the linear model using weighted regression, weighting each observation according to trial size. This is similar to the approach developed in section 6.4.3 above, except that the weights were the precision of the estimates defined as the inverse of the mean of the variances of $\hat{\alpha}_{ij}$ and $\hat{\beta}_{ij}$, rather than trial size. It is the more precise estimates of α_i and β_i that should carry more weight in this regression in the second stage, and this precision does not necessarily correspond to trial size. Either way, this approach will account for some of the heterogeneity of variation, but not all since $\hat{\alpha}_i$ are (falsely) assumed to be exact estimates of α_i.

3. The authors introduce a third approach which overcomes the problems of the other two, and refer to this as *Stijnen's Approach*, which is described below.

From stage I, the following is assumed to be true:

$$
\begin{pmatrix} \hat{\mu}_i \\ \hat{\alpha}_i \\ \hat{\beta}_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_i \\ \alpha_i \\ \beta_i \end{pmatrix}, C_i \right),
\tag{7.4}
$$

that is $\hat{\mu}_i$, $\hat{\alpha}_i$ and $\hat{\beta}_i$ are estimates of μ_i, α_i and β_i respectively where C_i is the covariance matrix of the estimates. If $\varepsilon_{S_{ik}}$ and $\varepsilon_{T_{ik}}$ are assumed to be independent the matrix C_i will be diagonal. This approach also allows for T_{ik} and S_{ik} to be modelled jointly in which case C_i will not be diagonal.

Stage II differs from the second approach to measurement error above with the addition of an intermediate step. A further assumption is that each trial-

204

specific parameter is a trial-specific realization of what the authors call *true overall treatment effects*:

$$\begin{pmatrix} \mu_i \\ \alpha_i \\ \beta_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \alpha \\ \beta \end{pmatrix}, \Sigma\right),$$ (7.5)

where Σ is the covariance matrix of the trial-specific parameters as estimates of the true treatment effects.

The resulting model is then:

$$\begin{pmatrix} \hat{\mu}_i \\ \hat{\alpha}_i \\ \hat{\beta}_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \alpha \\ \beta \end{pmatrix}, \Sigma + C_i\right).$$ (7.6)

The advantage of Stijnen's approach is that the covariance matrices C_i can be estimated in the first stage, and held fixed in the second stage, while the true underlying values μ, α, and β with the covariance matrix Σ are estimated while the variation in the estimates $\hat{\mu}_i$, $\hat{\alpha}_i$ and $\hat{\beta}_i$ about the trial-specific parameter values μ_i, α_i and β_i is properly accounted for.

This then is an important simplification of the Belgian Paradigm, and is not restricted to normal true and surrogate endpoints, as long as the estimates of the treatment effects on the true and surrogate endpoints can be assumed to follow normal distributions.

7.2.1.2 Multi-Arm Extension

One obvious drawback of using Stijnen's Approach for the data used in this thesis is that it is assumed that treatment comparisons are independent, that is, each trial was only evaluating one experimental treatment. In the data used in this thesis, many of the trials are multi-arm introducing a second level in the data hierarchy, and the subscript j, yielding $\hat{\alpha}_{ij}$ and $\hat{\beta}_{ij}$ as estimates of the treatment effect comparing treatment j with the control regimen (treatment 0) on the surrogate and true endpoints respectively (see section 6.4.3).

An additional intermediate step can then be introduced in the second stage. The approach described here is an extension of the methods introduced in Tibaldi et al. (2003). $\hat{\mu}_{ij}$ is not considered in this extension.

Let the following be true:

$$\begin{pmatrix} \hat{\alpha}_{ij} \\ \hat{\beta}_{ij} \end{pmatrix} \sim N \left(\begin{pmatrix} \alpha_{ij} \\ \beta_{ij} \end{pmatrix}, C_{ij} \right). \tag{7.7}$$

Parameter estimates are assumed to be estimates of the treatment-comparison-specific parameters rather than the trial-specific parameters with covariance matrix C_{ij}. Then:

$$\begin{pmatrix} \alpha_{ij} \\ \beta_{ij} \end{pmatrix} \sim N \left(\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix}, D_i \right), \tag{7.8}$$

the treatment-comparison-specific parameters are assumed to be realizations of the trial-specific parameters with covariance matrix D_i which in turn are assumed to be realizations of the true underlying parameters:

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \sim N \left(\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \Sigma \right). \tag{7.9}$$

Overall, this is:

$$\begin{pmatrix} \hat{\alpha}_{ij} \\ \hat{\beta}_{ij} \end{pmatrix} \sim N \left(\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, C_{ij} + D_i + \Sigma \right). \tag{7.10}$$

Σ can be estimated along with the parameters α and β holding both C_{ij}, as the within-trial (between-treatment comparison) variation and D_i as the between-trial variation, fixed. Both C_{ij} and D_i have been estimated in the previous stage.

7.2.1.3 Discussion

The extension described above takes into account the clustering of treatment comparisons within trial while keeping the advantages of Stijnen's Approach over the other two approaches described above.

The only drawback is the assumption that there exists *true overall treatment effects*. That is, there exists a true treatment effect on the surrogate endpoint, α, of which each trial-specific treatment effect, α_i, is an estimate and a true treatment effect on the true endpoint, β, of which each trial-specific treatment effect, β_i, is an estimate. This assumption may be valid when the experimental and the control treatments are the same for each trial i, but is not true when the treatment effects are from treatment comparisons comparing different experimental treatments with different control treatments, as is the case in the

206

data used for this thesis. In these data, it cannot be correct that there is an underlying α and β, but rather that there is an underlying relationship between α_i and β_i (if the marker in question is in fact a surrogate marker). On this basis, the underlying assumption is also invalidated on extending multiple arms which is perhaps why the authors themselves did not consider this extension. A different approach is therefore needed.

7.2.2 Alternatives to Stijnen's Approach

7.2.2.1 Introduction

It is clear that the meta-analytic approach used in Chapter 6 above did not fully account for all the variation or the heterogeneity in the estimates of the treatment effect, although the weighted linear regression with robust standard errors is a definite improvement over simple linear regression assuming equal variances. There is a real advantage in being able to use a method such as linear regression; it is available in all statistical packages, is conceptually straightforward to understand with a meaningful proportion of explained variation, and can be used and the results can be interpreted easily by non-statisticians such as clinicians and researchers. The key question is therefore: how much bias is introduced in using the simplified model described in Chapter 6? If the bias introduced is minimal, then the results described in Chapter 6 are reliable and the methods can reasonably be used to evaluate surrogate markers for response to treatment for tuberculosis.

Different elements of the HIV paradigm and Stijnen's approach in the Belgian paradigm are useful for evaluating surrogate markers in the context of data arising from studies evaluated in this thesis. Several different methods, incorporating these elements, will therefore be proposed in this chapter that deal with the heterogeneity of variances and the error in the estimates of the treatment effects. In the next section, each of these methods will be compared using a simulation study.

7.2.2.2 Model Detail

Stage I is common in all these methods proposed below and is identical to that described in section 6.4.3. This yields the estimates $\hat{\alpha}_{ij}$ and $\hat{\beta}_{ij}$ and variances of these estimates, $\sigma^2_{\alpha_{ij}} = \text{Var}(\hat{\alpha}_{ij})$ and $\sigma^2_{\beta_{ij}} = \text{Var}(\hat{\beta}_{ij})$, which are themselves estimated from the models. This follows the HIV paradigm as well as Stij-

nen's approach outlined above. The differences between the methods is the approach taken in Stage II. As a reminder, κ is the parameter describing the slope of the regression of $\hat{\beta}_{ij}$ on $\hat{\alpha}_{ij}$, τ^2 is the residual variance (the variance of $\hat{\beta}_{ij}$ conditional on $\hat{\alpha}_{ij}$) and R^2 is the proportion of variation in $\hat{\beta}_{ij}$ explained by $\hat{\alpha}_{ij}$. There are a number of different options:

1. *Simple linear regression.* This is the simplest approach, but also the approach which makes the most false assumptions. The varying precision and clustering within trials of the estimates is not accounted for and this method will therefore be discounted.

2. *Weighted linear regression with robust standard errors.* This is the approach outlined in Chapter 6. This is straightforward to implement, is quick to run and is an extension of linear regression that can be understood easily. As described in section 7.1, there are two drawbacks with this: the error in estimating α_{ij} is ignored, and only the standard errors of the parameters, not the parameters themselves, are adjusted for the clustering within trial. While this is a simple method to use and will therefore be the first method evaluated in the simulation study, these drawbacks may affect the results and therefore other approaches must be explored. This is described as **Method 1**.

3. *Weighted linear regression with correction for attenuation bias.* Carroll and Stefanski (1995) give a simple form of regression calibration for correcting for attenuation bias in the presence of measurement error in the explanatory variable. They assume *classical additive measurement error model* where $w = x + u$ is observed rather than the true explanatory variable x and u is independent of x, has $u \sim N(0, \sigma_u^2)$. The ordinary least squares estimator β_w from the regression of y on w is not a consistent estimate of β_x, the estimator from the regression of y on x, resulting in attenuation bias. It can be shown that $\beta_w = \lambda_r \beta_x$. The exact form of λ_r is derived in Appendix B for the case of regression of y on w with no intercept (λ_r^* in equation B.26 on page 258):

$$\lambda_r^* = \frac{\sigma_x^2 + \bar{x}^2}{\sigma_x^2 + \sigma_u^2 + \bar{x}^2} < 1, \qquad (7.11)$$

where x has mean $\bar{x}$ and variance σ_x^2. λ_r^* is called the *attenuating factor* or the *Reliability Ratio*. The simple method for correcting for attenuation

proposed by these authors involves estimating this factor and scaling the estimate β_w to yield a more accurate estimate of β_x. Given, $\hat{\sigma}_u^2$, an estimate of σ_u^2 derived from elsewhere, The reliability ratio can be estimated by:

$$\hat{\lambda}_r = \frac{\hat{\sigma}_w^2 - \hat{\sigma}_u^2 + \bar{w}^2}{\hat{\sigma}_w^2 + \bar{w}^2}, \tag{7.12}$$

where $\hat{\sigma}_w^2$ is the sample variance of w and $\bar{w}$ is the sample mean.

In the problem specific to this thesis, α_{ij} is estimated with error by $\hat{\alpha}_{ij}$ where the classical additive measurement error model is assumed with known variance $\sigma_u^2 = \sigma_{\hat{\alpha}_{ij}}^2 = \mathrm{Var}(\hat{\alpha}_{ij})$. Rather than a single error variance describing the error in $\hat{\alpha}_{ij}$ for all i and j, the variance is different for each observation. The mean will be used in the reliability ratio. The regression calibration estimate of κ, $\hat{\kappa}_{RC}$ is therefore:

$$\hat{\kappa}_{RC} = \frac{\mathrm{Var}_{ij}(\hat{\alpha}_{ij}) + \overline{\hat{\alpha}_{ij}}^2}{\mathrm{Var}_{ij}(\hat{\alpha}_{ij}) - \overline{\sigma_{\hat{\alpha}_{ij}}^2} + \overline{\hat{\alpha}_{ij}}^2} \hat{\kappa}_{naive}, \tag{7.13}$$

where $\hat{\kappa}_{naive}$ is the naive estimate of κ from the weighted linear regression model and where:

$$\overline{\sigma_{\hat{\alpha}_{ij}}^2} = \frac{\sum_{i,j} \sigma_{\hat{\alpha}_{ij}}^2}{\sum_i m_i}, \tag{7.14}$$

$$\overline{\hat{\alpha}_{ij}} = \frac{\sum_{i,j} \hat{\alpha}_{ij}}{\sum_i m_i} \tag{7.15}$$

and where $\mathrm{Var}_{ij}(\hat{\alpha}_{ij})$ is the variance in $\hat{\alpha}_{ij}$ across i and j in contrast to $\sigma_{\hat{\alpha}_{ij}}^2$ which is the variance in $\hat{\alpha}_{ij}$ for a specific i and j:

$$\mathrm{Var}_{ij}(\hat{\alpha}_{ij}) = \frac{1}{\sum_i m_i} \sum_{i,j} \left[\left(\hat{\alpha}_{ij} - \frac{\sum_{i,j} \hat{\alpha}_{ij}}{\sum_i m_i} \right)^2 \right]. \tag{7.16}$$

This approach does not adjust the estimates of the variance in $\hat{\kappa}$ or the estimates of τ^2, but the same attenuation factor can be used to adjust R^2 (see Appendix B). This will be included as the second method in the simulation study. This is described as **Method 2**.

4. *Random effects model.* A random effects model better deals with the clustering of treatment comparisons within trial than merely using robust

estimates of the standard errors, as the estimates themselves are also adjusted for the clustering. However, this model assumes that α_{ij} and β_{ij} are estimated without error and therefore does not account for the differing precisions of the estimates. The trials are of varying sizes and have different recurrence rates depending on the treatment regimens being compared as well as a number of other trial-specific factors. Therefore the precisions of the estimates of the treatment effects will differ greatly and any model that does not take this into account is of limited value. An application of the random effects model that incorporated weighting by precisions of estimates would be advantageous, but such a method was not accessible for use in this thesis and therefore this approach will not be explored further.

5. *SIMEX Algorithm.* The SIMEX algorithm was first proposed by Cook and Stefanski (1994) and is a method using *SIM*ulation followed by *EX*trapolation to remove the attenuation bias caused by error in the explanatory variables in a model. The implementation of the SIMEX algorithm in this thesis is based on that described by Carroll and Stefanski (1995). This is a method complementing regression calibration described in point 3 above that is more computationally intensive. However, it is more suited to the problem at hand since in this implementation, it allows for the measurement error to vary by observation. The key idea of SIMEX is to attempt to determine the effect of the measurement error on the final model estimates of interest using simulations. The SIMEX algorithm assumes the additive measurement error model where $\hat{\alpha}_{ij} = \alpha_{ij} + u_{ij}$ is observed rather than α_{ij} and where $u_{ij} \sim N(0, \sigma^2_{\alpha_{ij}})$.

In the *simulation step* of the algorithm, additional independent measurement errors with variance $\zeta_1 \sigma^2_{\alpha_{ij}}$ are generated and added to each $\hat{\alpha}_{ij}$. The resulting dataset can be thought of as a *contaminated dataset*. This is repeated with successively larger measurement error added to the data, such that $\zeta_1 < \zeta_2 < \cdots < \zeta_m < \zeta_{m+1} < \cdots$. For the mth dataset, the measurement error variance for each data point is therefore assumed to be $(1 + \zeta_m)\sigma^2_{\alpha_{ij}}$. In the *estimation step*, estimates of the coefficient κ are obtained in the usual way from each of the contaminated datasets. These two steps are then repeated to yield a large number of contaminated datasets and coefficients κ for each ζ_m. The mean of the estimate of κ for each level of contamination can then be plotted against the level of

contamination, ζ_m.

An illustration of this approach is shown in Figure 7.1. The estimates of the coefficients derived from the contaminated simulated datasets are plotted as solid circles and the naive estimate derived from the original dataset, corresponding to $\zeta_m = 0$, is plotted as a hollow circle. Extrapolation of the line fitted to these points back to $\zeta_m = -1$ yields the SIMEX estimate, κ_{SIMEX}. If the measurement error variance for a data point is $(1 + \zeta_m)\sigma^2_{\alpha_{ij}}$ then the measurement error variance at $\zeta_m = -1$ should be equal to $(1 + -1)\sigma^2_\alpha = 0$ and the SIMEX estimate, κ_{SIMEX}, should be an unbiased estimator of the parameter, κ, that would had been obtained had the α_{ij} been known exactly without error.
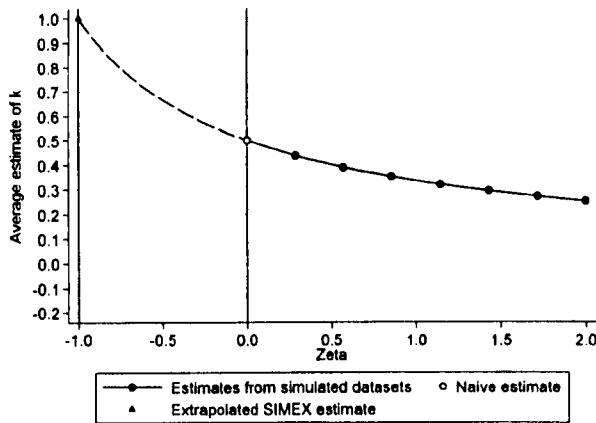
Figure 7.1: An illustration of estimates deriving from the SIMEX algorithm.

The same approach can be used to derive SIMEX estimates of τ^2 and R^2. As in point 3, this approach does not adjust the estimates of the variance in $\hat{R}$, but does correct for the attenuation bias in the estimate of κ. This will be included as the third method in the simulation study. This is described as **Method 3**.

7.2.3 Summary

Stijnen's approach is a useful simplification of the Belgian paradigm, but is not entirely suited to the application in this thesis. Three alternatives have been proposed that will be compared using a simulation study:

- **Method 1.** Weighted linear regression with robust standard errors as used in Chapter 6 with no correction for attenuation bias,

- **Method 2.** Weighted linear regression with correction for attenuation bias using the reliability ratio, and

- **Method 3.** The SIMEX algorithm with weighted linear regression to remove the effect of measurement error on the final parameters.

7.3 Simulation Study

7.3.0.1 Objective

In a simulation study, data can be simulated assuming a variety of different scenarios of known underlying distributional parameters. Statistical methods that are being evaluated can be applied to the data and the estimates of the parameters compared with the known true parameter values.

The objective of this simulation study was to compare the three methods on how accurately they estimate model parameters, focusing on the slope parameter, κ, and the variance of the estimate of the treatment effect on the true endpoint conditional on the estimate of the treatment effect on the surrogate endpoint, τ^2.

7.3.1 Methods

Using the approach presented by (Burton et al., 2006), a simulation study was designed and carried out to evaluate the three different approaches. Several datasets were simulated under different assumptions in order to compare the effect of the three methods on stronger and weaker surrogate endpoints. Each method was applied to all datasets and the results compared using performance measures of bias, accuracy and coverage.

7.3.1.1 Simulation data assumptions and inputs

Based on the discussion in section 7.2, it was assumed that there was an underlying intercept, δ, and slope, κ, that described the linear relationship between the effect of treatment on the surrogate endpoint and the effect of treatment on the true endpoint. Within each trial, there were assumed to be trial-specific

parameters describing this relationship, $\delta = \delta_i$ and $\kappa = \kappa_i$ which were sampled from the following distribution:

$$\begin{pmatrix} \delta_i \\ \kappa_i \end{pmatrix} \sim MVN \left(\begin{pmatrix} \delta \\ \kappa \end{pmatrix}, \begin{pmatrix} \sigma_\delta^2 & \rho_{\delta\kappa}\sigma_\delta^2\sigma_\kappa^2 \\ \cdot & \sigma_\kappa^2 \end{pmatrix} \right). \tag{7.17}$$

Under this model, δ_i and κ_i were assumed to be multivariate normal with variances σ_δ^2 and σ_κ^2 respectively and correlation $\rho_{\delta\kappa}$. These three parameters were held constant.

The structure and specific parameters values used to simulate the data was motivated by the analysis of the data used in this thesis and, in particular, the results from the analysis in Chapter 6, corresponding to method 1 to be evaluated in this chapter.

The treatment effect on the surrogate endpoint α_{ij} and on the true endpoint β_{ij} for treatment comparison j in trial i were sampled from the following distributions:

$$\alpha_{ij} \sim N\left(\mu_{\alpha_i}, \sigma_{\alpha_i}^2\right), \tag{7.18}$$

$$\beta_{ij} \sim N\left(\delta_i + \kappa_i\alpha_{ij}, \tau^2\right). \tag{7.19}$$

Here, μ_{α_i} and $\sigma_{\alpha_i}^2$ were trial specific parameters describing the effects of the treatment on the surrogate endpoint in trial i. τ^2 is the variance of β_{ij} conditional on α_{ij}. To give a range of values of α_{ij} and β_{ij} corresponding to treatments of varying efficacy, the trial specific parameters, μ_{α_i} and $\sigma_{\alpha_i}^2$ were sampled from the following distributions:

$$\mu_{\alpha_i} \sim N\left(0.4, 0.25\right), \tag{7.20}$$

$$\ln\left(\sigma_{\alpha_i}^2\right) \sim N\left(-1.2, 0.04\right). \tag{7.21}$$

The distributions of these parameters were selected to yield values of α_{ij} that were similar to those calculated from the real data.

To mimic the real data, each simulated dataset contained data from twelve trials each with a number of treatment arms ranging from 2 to 8 with around 100 individuals allocated to each treatment arm. The number of treatment

arms in each trial, m_i, was sampled from the binomial distribution:

$$(m_i - 2) \sim Bi \left(n = 8, p = 0.25 \right), \tag{7.22}$$

defined in such a way that the minimum number was 2 (corresponding to one treatment comparison), the median number was 4 (corresponding to three treatment comparisons) and 50% of the trials had 3, 4 or 5 regimens with the possibility of up to 10 regimens. Varying the number of treatment arms in each trial better reflects the real data combined from multiple trials.

The number of individuals allocated to regimen j in trial i, n_{ij}, was sampled from the normal distribution:

$$n_{ij} \sim N \left(100, 64 \right), \tag{7.23}$$

defined such that the mean number of individuals allocated to a single regimen was 100 with 99% of regimens having within 75 and 125 individuals.

True and surrogate responses (T_{ijk} and S_{ijk} respectively for individual $k = 1, \ldots, n_{ij}$ allocated to regimen $j = 0, \ldots m_i$ in trial $i = 1, \ldots, 12$) were then sampled independently from logistic distributions such that:

$$\text{logit} \left(P(S_{ijk} = 1) \right) = \mu_i + \alpha_{ij}, \tag{7.24}$$

$$\text{logit} \left(P(T_{ijk} = 1) \right) = \nu_i + \beta_{ij}, \tag{7.25}$$

where $\alpha_{i0} = 0$ and $\beta_{i0} = 0$. The parameters μ_i and ν_i are sampled from the following distributions:

$$\mu_i \sim N \left(1.4, 1.21 \right), \tag{7.26}$$

$$\nu_i \sim N \left(1.7, 0.49 \right). \tag{7.27}$$

Parameter	Value(s) taken
$\rho_{\delta\kappa}$	0.83
σ_κ^2	0.05
κ	0, 0.2, 0.5, 1.0, 1.5, 2.0
τ^2	0.1, 0.3, 1.0, 2.0

Table 7.1: Values of parameters used to simulate the data. These values are taken from the analysis of the real data in Chapter 6.

All remaining parameters values in the distributions defined above were estimated from the actual data so as to simulate data that represents real data. Table 7.1 shows the parameter values used to simulate the data. These estimates are based on the analysis presented in Chapter 6 evaluating the culture result at 3 months as a surrogate marker. κ and τ^2 were estimated from the data to be 1.29 and 0.30 respectively, and were varied to simulate a total of 24 different scenarios with surrogate markers of varying strength. The value of σ_δ^2 cannot be estimated from the data since δ is constrained to be 0, but was given the value 0.25.

7.3.1.2 Performance measures for evaluating different methods using simulation

Burton et al. (2006) group measures for evaluating different methods using simulation into three categories.

- *Bias*. The *bias*, *B*, is 'the deviation in an estimate from the true quantity' (Burton et al., 2006) and is defined as follows for an estimator $\hat{x}$ of the true quantity x_{true}:

$$B\left(\hat{x}\right) = \frac{1}{N_S} \sum_{s=1}^{N_S} \hat{x}_s - x_{true}, \qquad (7.28)$$

where N_S simulations have been performed and $\hat{x}_s$ is the estimator of x from the sth simulation dataset. When comparing the bias of a statistical method across scenarios where x_{true} varies, the *percentage bias* $(B(\hat{x})/x_{true})$ or the *standardised bias* $(B(\hat{x})/\,\text{SE}(\hat{x}))$ can be a more useful measures of bias where $\text{SE}(\hat{x})$ is the empirical standard error of the estimate of interest over all simulations.

- *Accuracy*. The *mean squared error* (*MSE*) is a useful measure of accuracy of the estimate incorporating measures of bias and of variability:

$$\text{MSE}(\hat{x}) = B(\hat{x}) + (\text{SE}(\hat{x}))^2. \qquad (7.29)$$

- *Coverage*. Since confidence intervals are commonly constructed to show the precision of an estimator, it is important to assess *coverage*. The coverage of a confidence interval is 'the proportion of times that the obtained confidence interval contains the true specified parameter value' (Burton et al., 2006), that is the proportion of times that the interval

$[\hat{x}_s - z_{1-\alpha/2}\,\mathrm{SE}(\hat{x}_s), \hat{x}_s + z_{1-\alpha/2}\,\mathrm{SE}(\hat{x}_s)]$ contains x_{true} for $s = 1, \ldots, N_s$, where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ percentile of the standard normal distribution, $z_{1-\alpha/2} = \Phi^{-1}(1 - \alpha/2)$, corresponding to the $100(1 - \alpha)\%$ confidence interval. Here $\mathrm{SE}(\hat{x}_s)$ is the estimate of the standard error of $\hat{x}_s$ from the sth simulated dataset. The coverage should be close to the level of confidence interval in question. For example, the 95% confidence intervals constructed should contain the true value in 95% of the simulated datasets.

The authors suggest a possible criterion for assessing adequate coverage. The coverage should not lie outside of approximately two standard errors of $(1 - \alpha)$ where an approximate formula is $\mathrm{SE}(1 - \alpha) = \sqrt{\alpha(1 - \alpha)/N_s}$. Table 7.2 shows the intervals within which an acceptable coverage should lie for a number of different confidence intervals calculated assuming 2000 simulations.

If the coverage is unacceptable, there are two possibilities: *over-coverage* and *under-coverage*. 'Over-coverage, where the coverage rates are above 95 per cent [for a 95% confidence interval], suggests that the results are too conservative as more simulations will not find a significant result when there is a true effect thus leading to a loss of statistical power with too many type II errors. In contrast, under-coverage, where the coverage rates are lower than 95 per cent, is unacceptable as it indicates over-confidence in the estimates since more simulations will incorrectly detect a significant result, which leads to higher than expected type I errors' (Burton et al., 2006).

Confidence Interval	Acceptable Coverage
50%	(0.478, 0.522)
75%	(0.731, 0.769)
90%	(0.887, 0.913)
95%	(0.940, 0.960)
97.5%	(0.968, 0.982)
99%	(0.986, 0.994)

Table 7.2: Intervals within which an acceptable coverage would lie for different confidence levels for 2000 simulated datasets.

7.3.1.3 Determining the number of simulations

Burton et al. (2006) emphasise the importance of giving a clear rationale for the number of simulations performed in the simulation study in the same way that a sample size calculation would always be performed before the start of a clinical trial to determine the required number of participants. The number of simulations for this simulation study was determined by simulating 10,000 datasets for a single scenario and observing the bias in the estimated parameters calculated from increasing numbers of simulations.

Using the same design described above for $\kappa_{true} = 1.5$ and $\tau^2_{true} = 0.3$ (to match the values estimated from the data), 10,000 datasets were simulated. The analysis described in Chapter 6, method 1 (the method that was expected to be the most biased method), was performed on each of these simulated datasets to yield estimates $\hat{\kappa}_s$ and $\hat{\tau}^2_s$ of κ and τ^2 respectively, $s = 1, \ldots, 10000$. The bias and mean squared error in each of $\hat{\kappa}$ and $\hat{\tau}^2$ were calculated using increasing numbers of simulations to evaluate how these measures varied as the number of simulations used increased. Figure 7.2 on the following page shows the biases in $\hat{\kappa}$ and τ^2 ($B_S(\hat{\kappa})$ and $B_S(\hat{\tau}^2)$) and the mean square errors in $\hat{\kappa}$ and $\hat{\tau}^2$. The number of simulations used to calculate the estimates is on the horizontal axis and is shown on the log scale since it is the accumulation of information that is of interest rather than increasing numbers of simulations.

From all of the graphs, there is wide variability initially with the lines beginning to settle from around 500 simulations. The change in all of the measures is small as the number of simulations doubles from 1000 to 2000 and even smaller as the numbers more than doubles from 2000 to 5000. It was therefore decided that 2000 simulations would be sufficient to assess the relative performances of the different statistical methods.

7.3.2 Results

2000 datasets for each of the 24 different scenarios were simulated and stage 1 of the analysis was applied to each dataset. Stage 1 (yielding $\hat{\alpha}_{ij}$ and $\hat{\beta}_{ij}$) is common to all three methods. The total computer time to simulate the 48000 datasets was 3 hours and 53 minutes and the total time to complete stage 1 was 12 hours and 22 minutes. All computer times for these simulations are from a standard laptop with 2 GB of RAM and a dual core processor each of 1.20 GHz and are given for a rough comparison of computation time required
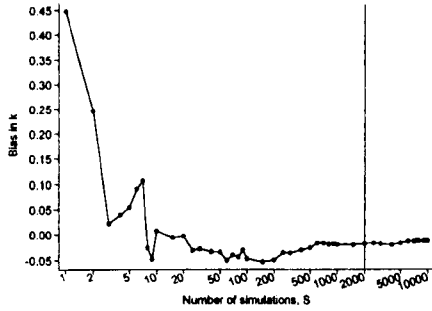
(a) Bias in $\hat{\kappa}$, $B_S(\hat{\kappa})$, by number of simulations (b) Bias in $\hat{\tau}^2$, $B_S(\tau^2)$, by number of simulations
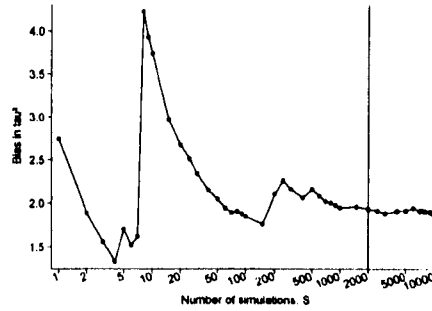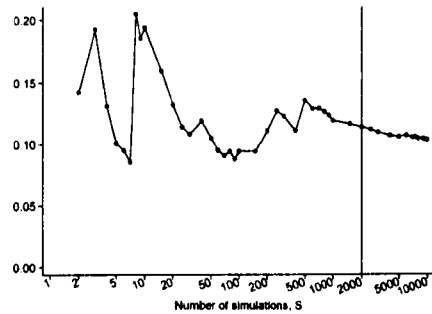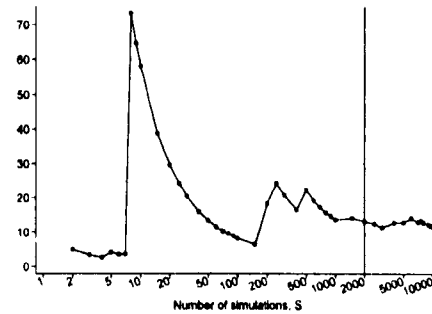
(c) Mean squared error in $\hat{\kappa}$ (d) Mean squared error in $\hat{\tau}^2$

Figure 7.2: Estimates of bias and mean square error in $\hat{\kappa}$ and $\hat{\tau}^2$ for increasing numbers of simulations.

for each method.

7.3.2.1 Method 1: No correction for attenuation bias.

A weighted linear regression analysis with robust standard errors was applied to the results of stage 1 as described above. The total run time was 3 hours and 15 minutes.

7.3.2.1.1 Bias Figure 7.3 on the next page shows graphs of bias and percentage bias in $\hat{\kappa}$ and $\hat{\tau}^2$ for different values of κ_{true} and τ^2_{true} (in the keys of the graphs, t^2 corresponds to τ^2_{true}).

It is clear from all graphs that the bias in estimating both parameters is not insubstantial, except for low values of κ. For low values of κ, the slope of the regression line of β_{ij} on α_{ij} is nearly flat and the association between α_{ij} and β_{ij} is very weak, corresponding to a very weak surrogate. The bias in κ crosses

218

(a) Bias in $\hat{\kappa}$, $B_S(\hat{\kappa})$

(b) Percentage bias in $\hat{\kappa}$

(c) Bias in $\hat{\tau}^2$, $B_S(\hat{\tau}^2)$

(d) Percentage bias in $\hat{\tau}^2$
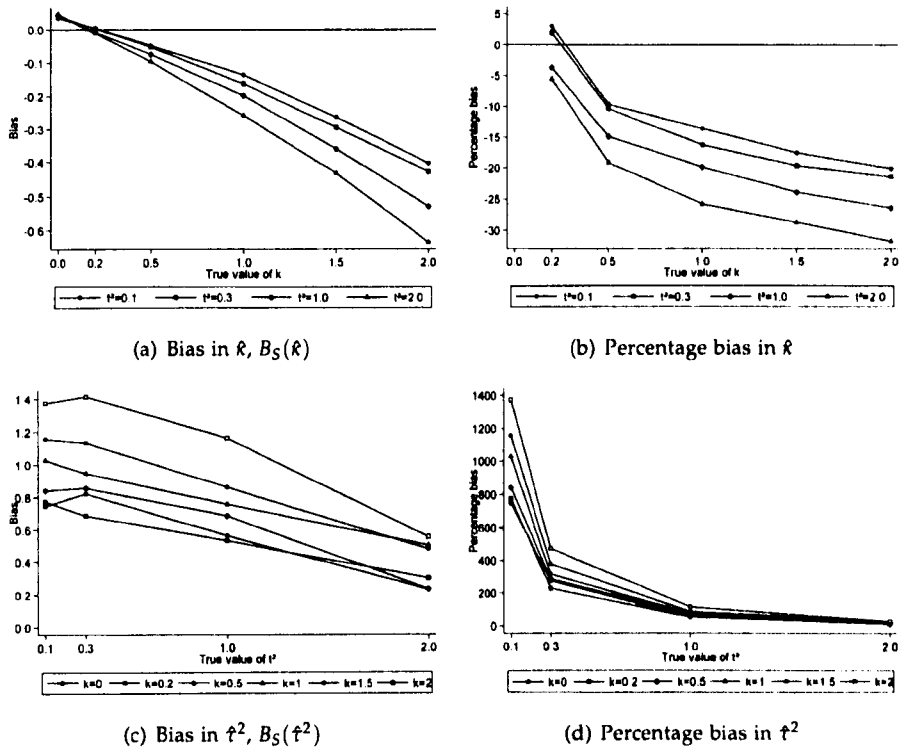
Figure 7.3: Bias and percentage bias in $\hat{\kappa}$ and $\hat{\tau}^2$, Method 1.

zero for low κ_{true} and therefore the method correctly estimates no association when there is none.

The bias in $\hat{\kappa}$ is negative implying κ is being underestimated and increases in magnitude as κ_{true} increases. The bias in $\hat{\tau}^2$ is positive and decreases as τ^2_{true} increases. At $\kappa_{true} = 1.5$ and $\tau^2_{true} = 0.3$ (similar to values calculated from the analysis in Chapter 6 evaluating the month 3 culture result as a surrogate), the bias in $\hat{\kappa}$ is -0.29 or -20% and the bias in τ^2_{true} is 1.13 or 378%.

7.3.2.1.2 Accuracy Figure 7.4 on the following page shows graphs of mean squared error in $\hat{\kappa}$ and $\hat{\tau}^2$ for different values of κ_{true} and τ^2_{true}.

The MSE in $\hat{\kappa}$ is small for small values of κ_{true}, but increases as κ_{true} increases. The mean squared error in $\hat{\tau}^2$ is 5 to 10 times greater with no discernable relationship with τ^2_{true}. At $\kappa_{true} = 1.5$ and $\tau^2_{true} = 0.3$, the MSE in $\hat{\kappa}$ is 0.17 and the MSE in $\hat{\tau}^2$ is 10.49.

(a) Mean squared error in $\hat{k}$ (b) Mean squared error in $\hat{t}^2$
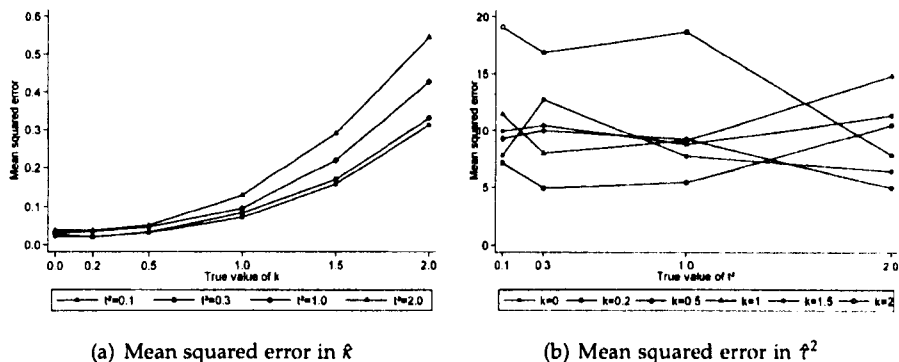
Figure 7.4: Mean squared error in $\hat{k}$ and $\hat{t}^2$, Method 1.

7.3.2.1.3 Coverage

The parameter κ and its standard error are estimated for each simulated dataset and from this the 95% confidence interval can be calculated. κ is the most important parameter of interest as it describes the slope of the linear relationship between the treatment effect on the true and the surrogate endpoints. Figure 7.5(a) shows the graph of coverage of 90% and 95% confidence intervals around $\hat{k}$ for different values of κ_{true} and τ^2_{true}.
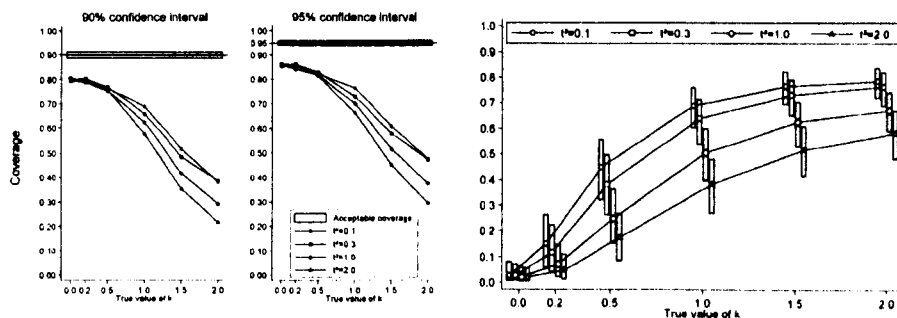
(a) Coverage of 90% and 95% confidence intervals around $\hat{k}$ after 2000 simulations. The shaded regions show the region within which an acceptable coverage would lie.

(b) Distribution of estimates of R^2. The plotted point is the median for that scenario and the grey bars show the inter-quartile range.

Figure 7.5: Coverage of confidence intervals around $\hat{k}$ and distribution of R^2, Method 1.

It is clear that there is considerable under-coverage. None of the points plotted are within the acceptable coverage region and the under-coverage is more severe for higher κ_{true} and τ^2_{true}. This means that the type I error rate is inflated and there is over-confidence in the estimate of κ.

220

7.3.2.1.4 Estimating R^2 Figure 7.5(b) on the preceding page shows the distribution of the values of the proportion of explained variation, R^2, in the treatment effect on the true endpoint that is explained by the treatment effect on the surrogate endpoint. The plotted points corresponds to the median and the grey bar shows the inter-quartile range for each combination of κ_{true} and τ^2_{true}.

For low κ_{true}, corresponding to a poor surrogate, R^2 is also low reflecting this. R^2 is also lower for higher τ^2_{true} corresponding to greater variability in β_{ij} conditional on α_{ij}. As κ_{true} increases and τ^2_{true} decreases, R^2 also increases reaching what appears to be a plateau around $R^2 = 0.8$. The short grey bars indicate that the calculated values of R^2 do not vary a great deal around the median.

7.3.2.2 Method 2: Correction for Attenuation using the Reliability Ratio

A weighted linear regression analysis with robust standard errors was applied to the results of stage 1 as described above. The reliability ratio was calculated for each dataset and the estimates of κ and R^2 were scaled accordingly. The total run time was 4 hours and 9 minutes.

7.3.2.2.1 Bias Figure 7.6 on the next page shows graphs of bias and percentage bias in $\hat{\kappa}$ for different values of κ_{true} and τ^2_{true}. The equivalent values from method 1 without correction for attenuation are shown on the graph in pale grey for comparison. Since only $\hat{\kappa}$ (in addition to R^2) was scaled by the reliability ratio, the estimates $\hat{\tau}^2$ remain unchanged from the weighted linear regression analysis with no correction for attenuation (method 1) and are therefore not presented here.

Comparing this with those from method 1, the overall pattern is very similar except that the bias is reduced a little as expected. The bias in $\hat{\kappa}$ is negative and increases in magnitude as κ_{true} increases. At $\kappa_{true} = 1.5$ and $\tau^2_{true} = 0.3$, the bias in $\hat{\kappa}$ is -0.16 or -11% compared to -20% from method 1.

7.3.2.2.2 Accuracy Figure 7.7(a) on page 223 shows graph of mean squared error in $\hat{\kappa}$ for different values of κ_{true} and τ^2_{true}. The mean squared error is a function of $\hat{\kappa}$ which was been scaled, and therefore the mean squared error is different to that resulting from method 1 (shown on the same graph in pale grey).

(a) Bias in $\hat{\kappa}$, $B_S(\hat{\kappa})$ (b) Percentage bias in $\hat{\kappa}$
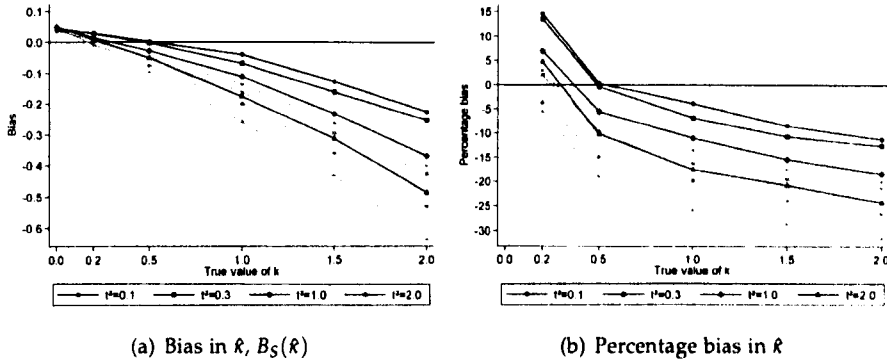
Figure 7.6: Bias and percentage bias in $\hat{\kappa}$, Method 2.

Comparing this that from method 1, the shape is very similar and, as above, the mean squared error is slightly smaller. At $\kappa_{true} = 1.5$ and $\tau^2_{true} = 0.3$, the MSE in $\hat{\kappa}$ is 0.13.

7.3.2.2.3 Coverage

Figure 7.7(b) on the next page shows the graph of coverage of 90% and 95% confidence intervals around $\hat{\kappa}$ for different values of κ_{true} and τ^2_{true}. The coverage from method 1 is shown in pale grey. It is clear that there is still considerable under-coverage, even after correction for attenuation bias. For $\kappa_{true} \leq 0.5$ (corresponding to weak surrogacy) the coverage is better before correcting for attenuation, but for $\kappa_{true} > 0.5$ the coverage improves after correction for attenuation. None of the points are within the acceptable coverage region and the under-coverage is more severe for higher κ_{true} and τ^2_{true}. This means that the type I error rate is inflated and there is over-confidence in the estimate of κ.

7.3.2.2.4 Estimating R^2

Figure 7.7(c) on the following page shows the distribution of the values of the proportion of explained variation, R^2. The plotted points corresponds to the median and the grey bar shows the inter-quartile range for each combination of κ_{true} and τ^2_{true}. Corresponding statistics from method 1 are shown in pale grey.

For low κ_{true}, corresponding to a poor surrogate, R^2 is also low reflecting this. R^2 is also lower for higher τ^2_{true} corresponding to greater variability in β_{ij} conditional on α_{ij}. As κ_{true} increases and τ^2_{true} decreases, R^2 also increases reaching what appears to be a plateau just below $R^2 = 0.9$. The short grey bars indicate that the calculated values of R^2 do not vary a great deal around
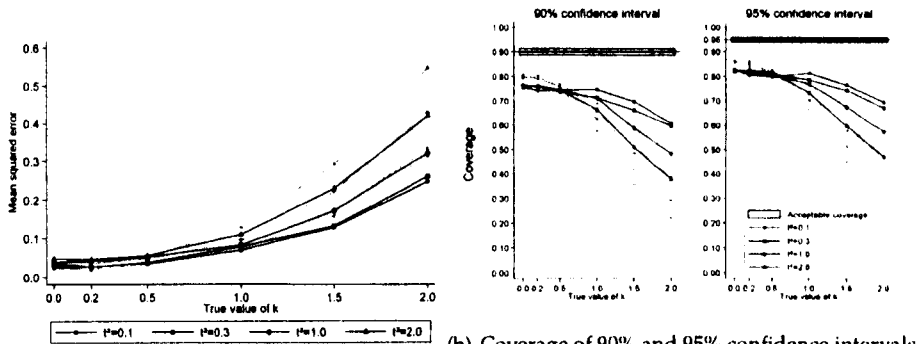
(a) Mean squared error in $\hat{k}$ after correction for attenuation using the reliability ratio.

(b) Coverage of 90% and 95% confidence intervals around $\hat{k}$ after 2000 simulations and after correction for attenuation using the reliability ratio. The shaded regions show the region within which an acceptable coverage would lie.
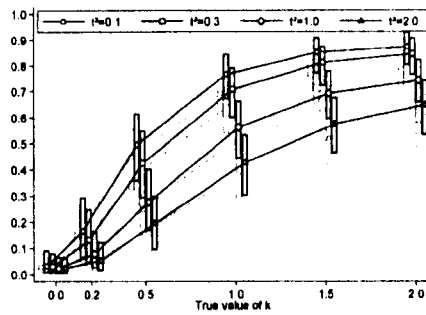
(c) Distribution of estimates of R^2 after correction for attenuation using the reliability ratio. The plotted point is the median for that scenario and the grey bars show the inter-quartile range.

Figure 7.7: Mean squared error in $\hat{k}$, coverage of confidence intervals around $\hat{k}$ and distribution of R^2, Method 2.

the median. The R^2 values are greater than those resulting from method 1 as they have been scaled by the reliability ratio which is always less than 1.

7.3.2.3 Method 3: Using the SIMEX algorithm.

The SIMEX algorithm was applied following the weighted linear regression analysis with robust standard errors was applied to the results of stage 1 as described above. The total run time was approximately 115 hours. The steps involved in implementing the SIMEX algorithm are as follows.

1. Calculate the naïve estimates of κ, $\text{Var}(\kappa)$, τ^2 and R^2.

2. Simulate 400 contaminated datasets with measurement error added to the $\hat{\alpha}_{ij}$ for each of $\zeta_m = 0.5, 1.0, 1.5, 2.0$. Only 400 datasets were used due to computational limitations. However, this compares favourably with the number used in examples in Carroll and Stefanski (1995).

3. Calculate estimates of κ, $\text{Var}(\kappa)$, τ^2 and R^2 for each of the contaminated datasets.

4. Calculate the 90% trimmed mean of each estimate at each value of $\zeta_m = 0.5, 1.0, 1.5, 2.0$. The 90% trimmed mean is calculated by taking the mean of values that lie above the 5th percentile and below the 95th percentile. In a very small number of these contaminated datasets, the calculated estimates were found to be either unfeasibly large or unfeasibly small and therefore the trimmed mean, excluding these outliers, was used rather than the simple mean.

   This then yields four estimates of each of the four parameters (κ, $\text{Var}(\kappa)$, τ^2 and R^2) for each value of ζ_m in addition to the naive estimates corresponding to $\zeta_m = 0$.

5. For each of κ, $\text{Var}(\kappa)$, τ^2 and R^2, the resulting estimates were plotted against ζ_m and a quadratic line fitted to these data to yield four different quadratic functions for each parameter. Carroll and Stefanski (1995) state that the quadratic extrapolant function is sufficiently complex in most cases. The SIMEX estimate for each parameter is then calculated as the point on the fitted line where $\zeta_m = -1$.

These steps are repeated for each of the 2000 simulations for each of the 24 scenarios and measures of bias, accuracy and coverage estimated as for methods 1 and 2.

Results from method 3 are presented in a similar way to as before and will be shown alongside those from method 2 or method 1 whichever is a more useful comparison.

7.3.2.3.1 Bias Figure 7.8 on the next page shows graphs of bias and percentage bias in $\hat{\kappa}$ and $\hat{\tau}^2$ for different values of κ_{true} and τ_{true}^2. For the bias and percentage bias in $\hat{\kappa}$, the same from method 2 correcting for attenuation bias using the reliability ratio are shown in pale grey to allow comparison. For the bias and percentage bias in $\hat{\tau}^2$, the same from method 1 are shown in pale grey to allow comparison.

224

(a) Bias in $\hat{k}$, $B_S(\hat{k})$

(b) Percentage bias in $\hat{k}$

(c) Bias in $\hat{t}^2$, $B_S(\hat{t}^2)$

(d) Percentage bias in $\hat{t}^2$
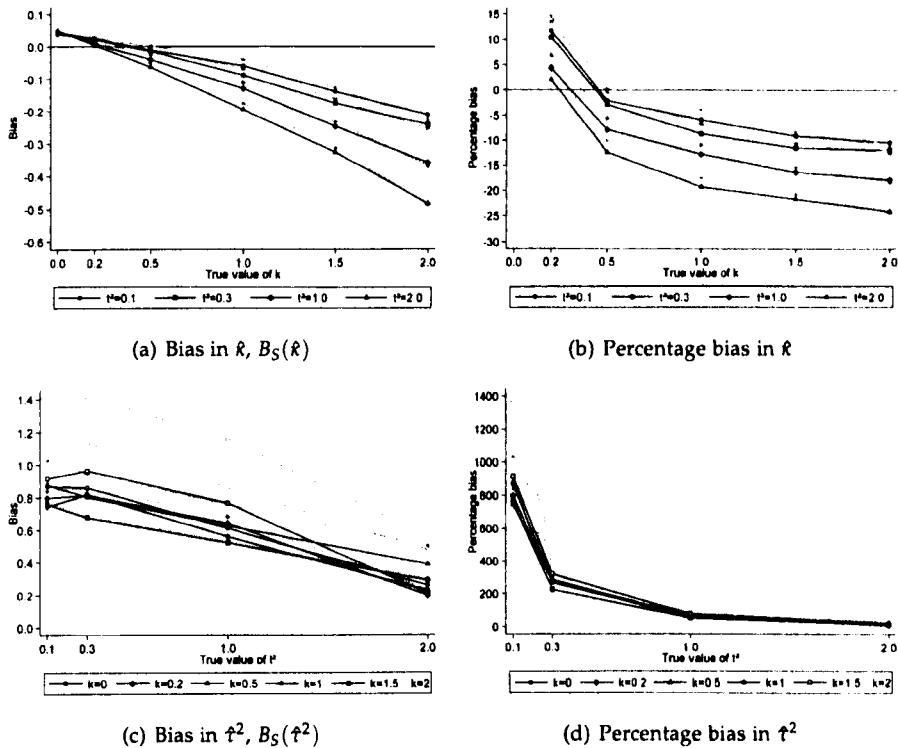
Figure 7.8: Bias and percentage bias in $\hat{k}$ and $\hat{t}^2$, Method 3.

The bias in $\hat{k}$ is less than in method 1 (not shown on this figure), but is only slightly greater than that resulting from method 2 (the grey lines in the figure). The SIMEX algorithm does not reduce the bias any more than a simple correction for attenuation bias using the reliability ratio.

The bias in $\hat{t}^2$ is also slightly less than that in method 1 (the grey lines in the figure).

The bias in $\hat{k}$ is negative and increases in magnitude as κ_{true} increases. The bias in $\hat{t}^2$ is positive and decreases as τ^2_{true} increases. At $\kappa_{true} = 1.5$ and $\tau^2_{true} = 0.3$, the bias in $\hat{k}$ is -0.17 or -12% and the bias in τ^2_{true} is 0.86 or 287%.

7.3.2.3.2 Accuracy Figure 7.9 on the following page shows graphs of mean squared error in $\hat{k}$ and $\hat{t}^2$ for different values of κ_{true} and τ^2_{true}. For the MSE in $\hat{k}$, the results from method 2 are shown in pale grey and for the MSE in $\hat{t}^2$, the results from method 1 are shown in pale grey to allow comparison.

The MSE in $\hat{k}$ is less than in method 1 (not shown in this figure), but is

(a) Mean squared error in $\hat{\kappa}$ (b) Mean squared error in $\hat{\tau}^2$
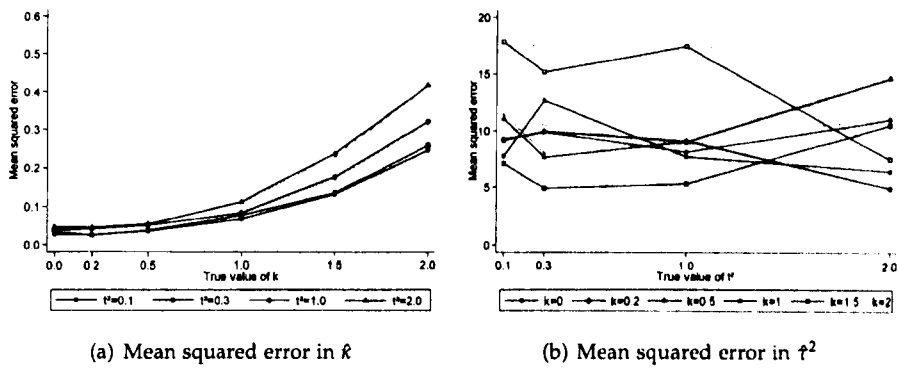
Figure 7.9: Mean squared error in $\hat{\kappa}$ and $\hat{\tau}^2$, Method 3.

marginally greater than that resulting from method 2 (the grey lines in the figure). The SIMEX algorithm does not reduce the MSE any more than a simple correction for attenuation bias using the reliability ratio. The MSE in $\hat{\kappa}$ is small for small values of κ_{true}, but increases as κ_{true} increases.

The MSE in $\hat{\tau}^2$ is also very slightly less than in method 1 (the grey lines in the figure).

The mean squared error in $\hat{\tau}^2$ is 5 to 10 times greater with no discernable relationship with τ^2_{true}. At $\kappa_{true} = 1.5$ and $\tau^2_{true} = 0.3$, the MSE in $\hat{\kappa}$ is 0.14 and the MSE in $\hat{\tau}^2$ is 9.90.

7.3.2.3.3 Coverage

Figure 7.10(a) on the next page shows the graph of coverage of 90% and 95% confidence intervals around $\hat{\kappa}$ for different values of κ_{true} and τ^2_{true}. The coverage from method 2 is shown in pale grey.

It is clear that there is considerable under-coverage after using the SIMEX algorithm and coverage is not improved over method 2. As with the other parameters described above, the coverage using method 3 lies somewhere between that using method 2 and that using method 1.

None of the points are within the acceptable coverage region and the under-coverage is more severe for higher κ_{true} and τ^2_{true}. This means that the type I error rate is inflated and there is over-confidence in the estimate of κ.

7.3.2.3.4 Estimating R^2

Figure 7.10(b) on the following page shows the distribution of the values of the proportion of explained variation, R^2, in the treatment effect on the true endpoint that is explained by the treatment effect
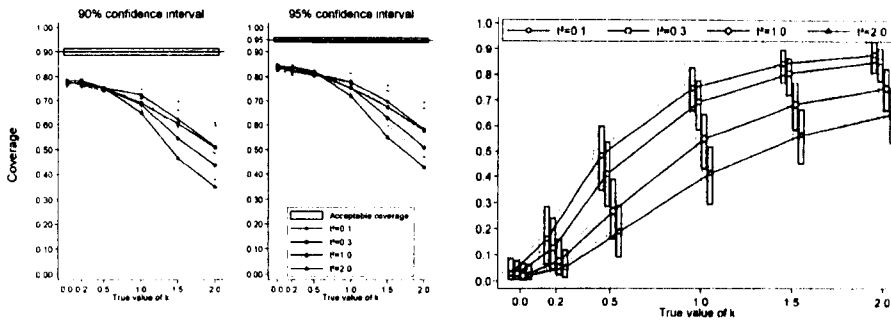
(a) Coverage of 90% and 95% confidence intervals around $\hat{\kappa}$ after 2000 simulations. The shaded regions show the region within which an acceptable coverage would lie.

(b) Distribution of estimates of R^2. The plotted point is the median for that scenario and the grey bars show the inter-quartile range.

Figure 7.10: Coverage of confidence intervals about $\hat{\kappa}$ and distribution of R^2, Method 3.

on the surrogate endpoint. The plotted points corresponds to the median and the grey bar shows the inter-quartile range for each combination of κ_{true} and τ^2_{true}. Corresponding statistics from method 2 are shown in pale grey.

For low κ_{true}, corresponding to a poor surrogate, R^2 is also low reflecting this. R^2 is also lower for higher τ^2_{true} corresponding to greater variability in β_{ij} conditional on α_{ij}. As κ_{true} increases and τ^2_{true} decreases, R^2 also increases reaching what appears to be a plateau around $R^2 = 0.85$. The short grey bars indicate that the calculated values of R^2 do not vary a great deal around the median. The values of R^2 found after using the SIMEX algorithm are marginally lower than those after correcting for measurement error using the reliability ratio.

7.3.2.4 Summary

Table 7.3 on the next page summarises the results from the simulation study comparing the three methods. Statistics are shown for the scenario $\kappa_{true} = 1.5$ and $\tau^2_{true} = 0.3$ as these are similar values to those calculated from the analysis in Chapter 6 evaluating the month 3 culture result as a surrogate. Method 2, correction for attenuation using the reliability ratio, did not affect the estimates of τ^2 and therefore the bias and MSE in $\hat{\tau}^2$ are not calculated for Method 2.

Method 3 improves on Method 1 in all of the summary statistics and

227

Statistic	Method 1	Method 2	Method 3
Bias in $\hat{\kappa}$	−0.29	−0.16	−0.17
Percentage bias in $\hat{\kappa}$	−20%	−11%	−12%
Bias in $\hat{\tau}^2$	1.13	-	0.86
Percentage bias in $\hat{\tau}^2$	378%	-	287%
MSE of $\hat{\kappa}$	0.17	0.13	0.14
MSE of $\hat{\tau}^2$	10.49	-	9.90
95% coverage of $\hat{\kappa}$	0.58	0.74	0.68
90% coverage of $\hat{\kappa}$	0.48	0.66	0.61
Median R^2	0.73	0.81	0.81

Table 7.3: Statistics resulting from the simulation study comparing the three methods at $\kappa_{true} = 1.5$ and $\tau^2_{true} = 0.3$. These parameter values are similar to those calculated from the analysis in Chapter 6 evaluating the month 3 culture result as a surrogate.

Method 2 again improves on Method 3, but this improvement is marginal.

7.4 Application to Trial Data

Methods 2 and 3 were applied to the trial data used in this thesis to explore how the results compare to those from Chapter 6. Table 7.4 show the estimate of the slope, κ, of the regression line and the proportion of explained variation, R^2, on applying each of methods 2 and 3 to the trial data used in this thesis. The results from Chapter 6 (method 1) are shown for comparison.

Marker	Statistic	Method 1 (95% CI)	Method 2	Method 3
Month 1	Slope, κ	1.35 (-0.10,2.80)	2.47	1.84
Culture[†]	R^2	0.36	0.67	0.49
Month 2	Slope, κ	0.85 (0.13,1.57)	1.40	1.07
Culture	R^2	0.36	0.58	0.45
Month 3	Slope, κ	1.29 (0.82,1.76)	-8.02	2.11
Culture	R^2	0.69	-4.29[‡]	1.13[‡]

[†]With a point of dichotomy at 20+.
[‡]These values for the proportion of explained variation are clearly inadmissable.

Table 7.4: Comparing the estimates of the slope, κ, of the regression line and the proportion of explained variation, R^2, on applying each of the three methods to the trial data used in this thesis.

For the month 1 (dichotomised at 20+ colonies) and the month 2 cultures, both methods scaled the proportion of explained variation up from 0.36 to

0.49 and 0.45 for method 3 and 0.67 and 0.58 for method 2.

For the month 3 culture, methods 2 and 3 yield inadmissible values for R^2 and the value of κ calculated from method 2 is clearly incorrect.

Table 7.4 shows the results of methods 2 and 3 applied to subsets of the trial data: East African trials, Hong Kong trials and comparisons for which both treatments contained rifampicin. It is clear from this table, that both method 2 and method 3 frequently yield uninterpretable values for R^2.

Subgroup	Marker	Statistic	Method 1	Method 2	Method 3
East African Trials	Month 1 Culture[†]	Slope, κ	1.14	1.54	1.44
		R^2	0.29	0.39	0.35
	Month 2 Culture	Slope, κ	0.76	1.06	0.92
		R^2	0.19	0.26	0.23
	Month 3 Culture	Slope, κ	1.61	3.94	2.28
		R^2	0.81	2.00[‡]	1.14[‡]
Hong Kong Trials	Month 1 Culture[†]	Slope, κ	1.98	5.20	3.18
		R^2	0.68	1.79[‡]	1.08[‡]
	Month 2 Culture	Slope, κ	0.99	1.25	1.21
		R^2	0.86	1.08[‡]	1.04[‡]
	Month 3 Culture	Slope, κ	0.82	5.58	1.43
		R^2	0.62	4.22[‡]	1.07[‡]
Comparisons of Rifampicin -containing Regimens	Month 1 Culture[†]	Slope, κ	1.85	9.71	2.85
		R^2	0.54	2.86[‡]	0.81
	Month 2 Culture	Slope, κ	1.00	1.93	1.28
		R^2	0.67	1.30[‡]	0.85
	Month 3 Culture	Slope, κ	0.88	-1.19	1.71
		R^2	0.46	-0.62[‡]	0.87

Table 7.5: Comparing the estimates of the slope, κ, of the regression line and the proportion of explained variation, R^2, on applying each of the three methods to subgroups of the trial data used in this thesis.

[†]With a point of dichotomy at 20+.
[‡]These values for the proportion of explained variation are clearly inadmissable.

7.5 Discussion

7.5.1 Simulation Study

As discussed in section 7.1, the methods used in Chapter 6 to evaluate culture results during treatment as surrogate markers had two main drawbacks.

Firstly, the treatment effect on the surrogate marker, α_{ij}, was estimated by $\hat{\alpha}_{ij}$ with error. However, this error was not accounted for in the model, leading to attenuation bias and over-estimation of the precision of the slope of the regression lines. Secondly, the precision of the parameter estimates was adjusted for clustering of treatment comparisons within trial using robust standard errors, but the point estimates themselves were calculated under the assumption of no clustering. Two methods, motivated by Stijnen's Approach (which itself proved to be inappropriate for this thesis due to multiple treatment comparisons), were selected to overcome the first drawback. The first involved scaling the estimate of the slope and of the proportion of explained variation by a reliability ratio to account for the attenuation (described as *Method 2*). The second involved using the SIMEX algorithm, simulating contaminated datasets and using extrapolation in an attempt to remove the effect of the measurement error (described as *Method 3*). Using a simulation study, these two methods were compared with the method used in Chapter 6 (described as *Method 1*) to determine the bias and precision of the model parameters estimates.

- *Method 1*. The simulation study was useful to evaluate the two stage method used in Chapter 6. It was found that there was considerable bias in the estimates of κ and τ^2; the mean squared error in estimating both was large and there was under-coverage in the confidence intervals around κ. An important advantage of this method is that it yields a clear graphical display of the behaviour of the surrogate marker. This means that, even if the estimates of κ and of R^2 are likely to be unreliable, a rough idea of the behaviour of the markers can be determined from these graphs.

- *Methods 2 and 3*. Both simple correction with the reliability ratio and the more computationally intensive SIMEX algorithm yield estimates of κ that are less biased and have lower mean square errors. The coverage of confidence intervals around κ is improved (for $\kappa > 0.5$), although there is still considerable undercoverage. The estimates from the SIMEX algorithm are marginally more biased than those using the correction with reliability ratio and the coverage is poorer. Since the SIMEX algorithm takes substantially more computer time (approximately 115 hours compared to 4 hours and 9 minutes using the reliability ratio for the whole simulation study) correction for attenuation with the reliability ratio is to be preferred.

230

Correction for attenuation with the reliability ratio does reduce bias and improve coverage, but the bias is not completely removed and there is still unacceptable under-coverage. Since two different methods for correcting for attenuation (the SIMEX algorithm and the reliability ratio) yielded roughly similar estimates of κ, it appears that the bias and the under-coverage is not due to attenuation caused by measurement error in $\hat{\alpha}_{ij}$ alone and there are other sources of bias in these methods.

None of the methods described here adjust the point estimates of κ for the clustering within trial, they merely incorporate robust standard errors that take account of this clustering in the standard errors. It is likely that this remaining bias may be due to the clustering of treatment comparisons within trials which has not been properly accounted for.

7.5.2 Application to Trial Data

For the month 3 culture, methods 2 and 3 yield inadmissible values for R^2 and the value of κ calculated from method 2 is clearly incorrect. In the subgroup analyses, methods 2 and 3 frequently yielded inadmissible values for R^2, most often when the R^2 resulting from method 1 was greater than around 0.65. It is likely that one of the reasons for this is the large variance in estimating α_{ij} ($\sigma^2_{\alpha_{ij}} = \mathrm{Var}(\hat{\alpha}_{ik})$) compared to the variance of $\hat{\alpha}_{ij}$ across trials ($\mathrm{Var}_{ij}(\hat{\alpha}_{ij})$).

For the month 1 culture result the variance of $\hat{\alpha}_{ij}$ across all trials i and treatment comparisons j is $\mathrm{Var}_{ij}(\hat{\alpha}_{ij}) = 0.203$ and the mean of $\sigma^2_{\alpha_{ij}}$ across all trials i and treatment comparisons j is smaller, 0.103. For the month 2 culture result $\mathrm{Var}_{ij}(\hat{\alpha}_{ij}) = 0.389$ and the mean of $\sigma^2_{\alpha_{ij}}$ is 0.201, again smaller. For the month 3 culture result $\mathrm{Var}_{ij}(\hat{\alpha}_{ij}) = 0.277$ and the mean of $\sigma^2_{\alpha_{ij}}$ is 0.550, considerably larger. The measurement error in estimating the quantities α_{ij} when evaluating the month 3 culture result as a surrogate is larger than the spread of α_{ij} across trials. This is due to the small number of individuals culture positive at month 3. This effectively means that the noise in α_{ij} is greater than the information available and therefore attempting to remove this error by method 2 or method 3 yields results that are uninterpretable.

Similarly, the measurement error in estimating α_{ij} is larger or only slightly smaller than the variance of $\hat{\alpha}_{ij}$ across all treatment comparisons in Hong Kong trials for each of the three markers and in East African trials for month 3 culture result, leading to R^2 values over 1 in each case for methods 2 and 3. Interestingly, the same is again true when including only rifampicin-containing

treatment comparisons, but method 3 gives sensible values for R^2 whereas method 2 does not. This suggests the method 3 is more robust to large measurement error, but it is clear that neither method is very reliable for use on these data.

7.5.3 Conclusions

These two extensions (adjustment using the reliability ratio and extrapolation using the SIMEX algorithm) do not remove all of the bias in estimating the slope parameter κ. Application of these methods to the trial data used in this thesis have yielded some values of R^2 and κ that are uninterpretable. The application of these two methods are therefore not appropriate without further work into the causes of (i) the residual bias and (ii) the inadmissible R^2 values.

Further discussion of these results and areas for future research are found in Chapter 8.

Chapter 8

Discussion and Conclusions

8.1 Introduction

In a recent issue of Statistical Methods in Medical Research, the following important point is made:

> 'Upon identification of a possible surrogate endpoint, in order to gain widespread acceptance for that endpoint, multiple groups must be convinced of its validity. These parties include practising clinicians, statisticians, regulatory bodies and other researchers. The statistical knowledge of these groups is variable, ranging from knowledgeable (regulatory bodies, statisticians) to often very limited (clinicians). Based on these considerations, it is our opinion that a single statistic, graphic, or theory of surrogate endpoints will not be adequate to satisfy everyone who must accept the validity of a proposed surrogate endpoint.'

> (Green et al., 2008)

Formal methods for the evaluation of surrogate markers have been 'the subject of intensive research over the past decades' (from the editorial in the same issue, Burzykowski (2008)) and there is not yet any clear consensus on the best and simplest methods for evaluating surrogate markers. Green and colleagues are therefore right to recommend the use of a variety of methods.

A prognostic marker is used to predict disease outcome in an individual and is used as a clinician's tool. In contrast, a surrogate marker is used as a

233

substitute for the true endpoint in a clinical trial and is therefore more of a trialist's tool. Unlike the disease area of HIV, for example, there is currently no reliable prognostic marker for the tuberculosis TB clinician, and there is certainly no reliable surrogate marker for poor outcome to treatment for TB for the trialist. Despite limited evidence, both the World Health Organisation (WHO) and the International Union against TB (IUATLD) recommend extending the intensive phase of two months of treatment by an additional month if a patient has a positive smear at two months (see section 3.5.2), although it is expected that this recommendation will be removed from the fourth edition of the WHO TB treatment guidelines due for publication in the summer of 2009. There is considerably less evidence for surrogacy, and yet some authors implicitly or explicitly make claims for the use of the two month culture result as a surrogate endpoint. There have been several phase II studies conducted in the past few years evaluating the benefit of moxifloxacin in the treatment of TB that have used culture conversion at eight weeks as the primary outcome of the trial. Examples include the Phase II study of the OFLOTUB consortium (Rustomjee et al., 2008b) and those recently conducted by the CDC TBTC[1]. The most recent such study evaluating the addition of Moxifloxacin to the standard regimen was published in the Lancet in April 2009. The authors were quick to point out that the statistically significant results based on the eight week culture results '[do] not prove the efficacy of moxifloxacin to shorten tuberculosis treatment' (Conde et al., 2009). Nevertheless, a letter in the same issue commending these authors describe the eight week culture status as 'a hallmark in assessing regimen efficacy'.

There is therefore an urgent need for clear evidence to inform decisions regarding the use of culture results during treatment as prognostic and surrogate markers. It is in this context that the results presented in this thesis must be placed.

8.1.1 The MRC Clinical Trial Data

Evaluation of surrogate markers requires data from multiple clinical trials where both the surrogate marker and the true endpoint are measured. The data used in this thesis are from twelve TB treatment clinical trials conducted by the MRC in East Africa and East Asia in the 1970s and 1980s. 37 treatment

---

[1]Completed and ongoing TBTC Studies. http://www.cdc.gov/tb/tbtc/projects.htm. Retrieved 23 Apr 2009.

comparisons including 6974 trial participants were included for use in these analyses. These data are unique—no other TB clinical trial data of such quantity and quality exist—and provide an unprecedented opportunity to evaluate culture results during treatment as prognostic and surrogate markers.

8.1.2 Other Markers

The objectives of this thesis were not restricted to culture results during treatment as prognostic and surrogate markers. However, following the review of the literature in Chapter 3, it was clear that the markers that had the most potential *and* for which a reasonable amount of data were available for evaluation were culture results during treatment. These markers were therefore selected and evaluated as prognostic and surrogate markers.

The only clinical trial data of a sufficiently large quantity (derived from multiple trials and a variety of treatment comparisons) to evaluate any markers as surrogates pertains to smear or culture results during treatment. Since smear results are less specific and commonly less sensitive, culture results are the only markers that can be evaluated as surrogates with data currently available today.

Not only are there no other data currently available, it is highly unlikely that similar data on another marker will be available for the evaluation of a surrogate marker in the near future. The REMoxTB trial will provide data on a host of possible surrogate markers (serial sputum colony counts, repeated days to positivity from liquid media and others) in the context of a clinical trial, and other phase III trials currently being conducted or due to start in the next few years will add to this. However, the earliest that even data from one trial could be available is not likely to be before 2012 and most likely considerably later.

8.2 Prognostic Markers

In Chapter 5, culture results at each of months 1, 2, 3 and 4 during treatment were evaluated as prognostic markers exploring the effect of varying the point of dichotomy. Table 8.1 summarises the main findings from Chapter 5 for each marker at points of dichotomy of 1 and 20+.

It is clear that there was a strong association between a positive culture at each of months 1, 2, 3 and 4 and poor outcome. A patient with a positive

235

	Point of Dichotomy					
	1			20+		
	OR (95% CI)	TPF	FPF	OR (95% CI)	TPF	FPF
Month 1	3.0 (2.9, 3.1)	80%	56%	2.9 (2.9, 3.0)	55%	30%
Month 2	3.7 (3.6, 3.8)	40%	15%	5.2 (4.9, 5.6)	13%	3%
Month 3	6.8 (6.6, 7.0)	19%	4%	13.5 (12.5, 14.6)	8%	1%
Month 4	8.7 (8.2, 9.2)	15%	2%	28.6 (24.4, 33.5)	10%	1%

Table 8.1: Odds Ratios (labelled OR) with 95% confidence intervals (labelled 95% CI), True Positive Fraction (TPF) and False Positive Fraction (FPF) at different months for three different points of dichotomy.

culture at two months has an odds of poor outcome 3.7 times that of a patient with a negative culture at two months. A patient with a heavily positive (20+ or more) culture at month 3 has an odd of poor outcome 13.5 times that of a patient with a negative or sparsely positive culture. Nevertheless, the low values of TPF for a point of dichotomy of 1 show that while culture results give an indication of likelihood of poor outcome, they can not reliably be used to predict the outcome for a particular individual. Only 40% of patients with a positive culture at two months have a poor outcome of treatment and only 19% of patients with a positive culture at three months have a poor outcome.

On the basis of the these data, culture results during treatment are useful for identifying groups of patients that are at risk of having a poor outcome to treatment, but cannot be used for reliably predicting poor outcome for an individual.

8.3 Surrogate Markers

The meta-analysis approach developed in Chapter 6 is methodologically more rigorous than the single trial methods and therefore more weight is placed on those results. The meta-analysis consists of two stages. In stage I, logistic regression is used to estimate α_{ij}, the log odds ratio of a positive culture result for each treatment comparison (the treatment effect on the surrogate) and β_{ij}, the log odds ratio of poor outcome for each treatment comparison (the treatment effect on the true endpoint). Stage II involves regression of β_{ij} on α_{ij} using weighted linear regression with the weights being the inverse of the mean of the variances of $\hat{\alpha}_{ij}$ and $\hat{\beta}_{ij}$ and robust standard errors used to adjust for clustering of treatment comparisons within trials.

236

In the East African trials, the month 3 culture result was shown to be a good surrogate ($R^2 = 0.81$), and in the Hong Kong trials, the month 2 culture result was shown to be a good surrogate ($R^2 = 0.86$). These values for R^2 are high, but the discrepancy between East Africa and East Asia is difficult to interpret, and is most likely due to a combination of two factors: 1) the longer times to conversion in East African trials, and 2) the higher proportion of rifampicin-containing regimens in the Hong Kong trials. Since rifampicin is the most important drug in the standard regimen used today, results from rifampicin-containing regimens are likely to be more generalisable. Delayed culture conversion rates could be due to poorer adherence, different strains of mycobacteria, more extensive cavitation, more advanced disease due to later diagnosis or geographic differences in climate, culture or genetics. Detailed exploration of these differences is beyond the scope of this thesis and is an important area for future research.

It is interesting to observe that the first prognostic study in TB also found differences between one trial conducted in Hong Kong and two conducted in East Africa (Aber and Nunn, 1978). The authors found that the 2 and the 3 month culture results were good predictors of recurrence in the East African trials, but not in the Hong Kong trial where the 2 month smear was found to be a good predictor. This difference mirrors the differences found in the results in this thesis and are perhaps not surprising as data from these three trials is included in the data used in this thesis.

While further work is needed to understand the differences between East African and Hong Kong trials, the results are encouraging. They suggest that culture results during treatment (whether at two months or at three months) may capture a significant proportion of the treatment effect on treatment outcome. The time to culture conversion, or an aspect of the longitudinal profile of culture results during the first few weeks of treatment (such as Serial Sputum Colony Counting, see section 3.5.5) may prove to be better surrogates than binary culture results. These could not be evaluated as surrogates using these data, since cultures were not taken frequently enough during the first few weeks of treatment. Nevertheless, the results presented in this thesis are inconclusive and further work is required evaluating culture results during treatment as surrogate markers.

8.3.1 Choice of Control Arm

As described in section 6.1.1, most of the trials that yielded data included in the analyses in this thesis did not have a pre-specified 'control regimen' and so the arm in each trial with the highest proportion of poor outcomes was selected as the nominal 'control regimen' for treatment comparisons.

It is not expected that this largely arbitrary choice would affect the results, indeed this choice was made to provide the largest differences in proportions of poor outcomes possible in each treatment comparison, and therefore provide the largest amount of information possible for evaluation of surrogate markers. Nevertheless, it would be informative to repeat the analyses using a different strategy to select the nominal control. The control could be selected as the regimen with the lowest proportion of poor outcomes, or could be selected entirely at random. If the results remained broadly similar across different control regimen choice strategies, then it would be appropriate to conclude that the choice of control arm as used in this study did not affect the results.

8.4 Methodological Extension

8.4.1 Summary of Results and Conclusions

There was previously no work on methodology specifically for the evaluation of surrogate endpoints in the context of multi-arm trials yielding multiple treatment comparisons. The two stage method developed in Chapter 6 based on the HIV paradigm takes account of the multi-arm trials common in TB treatment research. These meta-analytic methods can be used for evaluating a binary surrogate endpoint for a binary true endpoint using data from trials with multiple treatment comparisons although they suffered from some drawbacks.

In modelling the relationship between the estimates of the treatment effect on the true endpoint and the treatment effect on the surrogate endpoint (stage II of the meta-analysis), there are several complexities which must be taken account of. Merely considering the points as independent with equal variance and using simple linear regression will result in biased and over-precise estimates of the slope parameter. Each pair of points corresponds to a single treatment comparison which are clustered within trials—some trials with

only two regimens corresponding to a single treatment comparison, and some trials with up to eight regimens corresponding to up to seven treatment comparisons. In addition, the treatment effects themselves are estimates derived from models in stage I of the meta-analysis and are estimated with varying precision depending on the number of patients included in the comparison and the number of patients that experienced a poor outcome. In Chapter 6 and in the two extensions in Chapter 7, (i) weighted regression was used to account for the varying precision in the estimates with (ii) robust standard errors used to account for the clustering of treatment comparisons within trial. This approach is an improvement on simple linear regression, but is not without drawbacks.

Two extensions of the meta-analytic approach developed in Chapter 6 were proposed in Chapter 7 with the aim of removing the bias resulting from the error in estimating the slope of the fitted line. The first method, correction for attenuation using the reliability ratio, involves scaling the slope parameter by a simple ratio to remove the bias. The reliability ratio is calculated from the spread of the estimates of treatment effects on the surrogate endpoint, $\hat{\alpha}_{ij}$, across the treatment comparisons and from the variances of these estimates. Standard errors are not adjusted, although the reliability ratio can be used to scale the proportion of explained variation, R^2. The second method, the SIMEX algorithm, uses simulation and extrapolation to estimate the slope parameter and standard error with the effect of the error in estimating α_{ij} removed.

These two methods were compared to that used in Chapter 6 using a simulation study. This simulation study was large, involving data with a complex hierarchical structure requiring several distributional assumptions. The results of this study demonstrated that there was considerable bias in estimating the slope parameter κ and under-coverage in confidence intervals around κ in the method used in Chapter 6. Both of the extensions resulted in better coverage and reduced bias, but this improvement was not substantial. Applying these two extensions to the trial data resulted, in some cases, with incorrect estimates of R^2 greater than 1 and unreasonable estimates of the slope.

It is likely that the problems with these methods are a result of a combination of three issues:

1. *Clustering of treatment comparisons within trials.* The two methods are used to remove attenuation bias due to the imprecision of the estimates $\hat{\alpha}_{ij}$ of

α_{ij} in stage II of the meta-analysis approach. Robust standard errors are used to adjust the standard errors for the clustering of treatment comparisons within trials, but the point estimates are not adjusted for clustering. The bias and undercoverage in using these two extensions is likely to be due in part to the fact that they do not make adjustment for this clustering.

2. *Post-hoc corrective methods.* Both of the extensions described in Chapter 7 are used to adjust model estimates for bias caused by measurement error. In this way, these two extensions are corrective for the error caused by using an incorrect model in stage II of the meta-analysis. It is assumed that the error in estimating α_{ij} is additive, in that the estimate $\hat{\alpha}_{ij} = \alpha_{ij} + u_{ij}$ where u_{ij} is the error assumed to be normally distributed with zero mean and variance σ_u^2. The better way to deal with the problem of observing the estimate $\hat{\alpha}_{ij}$ rather than the actual value α_{ij}, would be to use a model that properly takes account for this measurement error, rather that adjusting for it in a post-hoc way.

3. *Specification of the simulation data.* The first step in conducting a simulation study is to simulate realistic data under a variety of scenarios. The methods being evaluated in the simulation study will then be applied to these data and the parameter estimates compared with the known underlying parameters of the distribution. In the simulation study in Chapter 7, certain assumptions were made about the joint distribution of the true and surrogate endpoints. These assumptions were based on what might be expected of the hierarchical relationships between culture results and poor outcome and the analysis on the MRC trial data in Chapter 6. Assumptions were made about the distribution of the trial-specific intercept and slope, δ_i and κ_i, the treatment-comparison-specific treatment effects, α_{ij} and β_{ij}, and the patient-specific values for the true and surrogate endpoints, T_{ijk} and S_{ijk}. Values were chosen for 19 different distributional parameters used in simulating the data.

The approach used to simulate the data in the study was based on the results of the meta-analysis in Chapter 6 and on a theoretical expectation of the relationship between culture results during treatment and poor outcome. There are different ways of modelling the data, and further work is needed in determining which is the most suitable model for the

simulation data. One approach would be to constrain $\delta_i = \delta = 0$, rather than drawing δ_i from the bivariate distribution

$$\delta_i \sim MVN\left(\begin{pmatrix} \delta \\ \kappa \end{pmatrix}, \begin{pmatrix} \sigma_\delta^2 & \rho_{\delta\kappa}\sigma_\delta^2\sigma_\kappa^2 \\ & \sigma_\kappa^2 \end{pmatrix}\right), \qquad (8.1)$$

as has been done in Chapter 7 (see equation 7.17 on page 213). Other changes that could simplify the data structure could include constraining each trial to have the same number of regimens ($m_i = m$, where $m > 2$) or making changes to the values that α_{ij} and β_{ij} could take (see section 7.3.1.1).

8.4.2 Implications of Estimation Errors in $\hat\alpha$ and $\hat\beta$

Both $\hat\alpha_{ij}$ and $\hat\beta_{ij}$ are estimates with associated variances $\sigma_{\alpha_{ij}}^2$ and $\sigma_{\beta_{ij}}^2$ respectively. The error in $\hat\beta_{ij}$ as an estimate of β_{ij} is not incorporated into the model, except in the weights w_{ij} which are the inverse of the mean of $\sigma_{\alpha_{ij}}^2$ and $\sigma_{\beta_{ij}}^2$. Also, since α_{ij} and β_{ij} are estimated separately in Stage I, the assumption that $\hat\alpha_{ij}$ and $\hat\beta_{ij}$ are independent is inherent and so there is no adjustment for this correlation $\rho_{\alpha\beta}$. It is likely that this assumption is not valid as α_{ij} and β_{ij} are estimated from the same individual trial participants. $\rho_{\alpha\beta}$ could be estimated using bootstrap methods. Wrongly assuming $\rho_{\alpha\beta} = 0$ could lead to bias in the estimates of R^2 in stage II. Due to the choice of treatment ordering, almost all estimates of $\hat\alpha_{ij}$ and $\hat\beta_{ij}$ are negative and the slopes of the lines of best fit are positive, under the assumption $\rho_{\alpha\beta} = 0$. If, for the pairs of points plotted, the $\rho_{\alpha\beta}$ are predominantly negative and therefore the major axes of the elliptical confidence regions are roughly perpendicular to the line, the spread of points about the line is in fact greater than if $\rho_{\alpha\beta} = 0$. The estimated R^2 will therefore be more diluted and will be *smaller* than the true R^2. Conversely, if the $\rho_{\alpha\beta}$ are predominantly positive and therefore the major axes of the confidence regions are roughly parallel to the line of best fit, the estimated R^2 will be less diluted than expected and could be slightly *larger* than the true R^2. It is likely that most of the $\rho_{\alpha\beta}$ will be positive and the effect of assuming $\rho_{\alpha\beta} = 0$ may therefore be minimal.

241

8.4.3 Conclusions

The meta-analytic methods developed within the Belgian paradigm are simple for true and surrogate Gaussian endpoints, as the joint distribution is bivariate Gaussian which can be modelled without much difficulty. For binary true and surrogate endpoints (as is the case for evaluating culture results during treatment as surrogate markers for poor outcome), the methods are considerably more complex and involve the use of copulas to describe the joint distribution. Stijnen's approach (described in section 7.2.1) is a development of these methods, moving away from a joint model allowing for binary true and surrogate endpoints in a two stage approach. However, while a useful simplification, this approach could not be used for trials with more than one treatment comparison per trial.

The meta-analytic methods developed in Chapter 6 can be used for evaluating a binary surrogate endpoint for a binary true endpoint using data from trials with multiple treatment comparisons although it suffered from some drawbacks described above. The two extensions described in Chapter 7 did remove some of the bias, but there remain reservations about their use in evaluating surrogate markers.

Nevertheless, while there are drawbacks in stage II of the meta-analytic approach used in Chapter 6, the graphical representation clearly demonstrates whether there is a relationship between the treatment effect on the true endpoint and the treatment effect on the surrogate endpoint. The exact values for the slope parameter, κ, the 95% confidence interval around κ, and the proportion of explained variation, R^2, should not be relied on until further methods are developed, but it is clear that the two month culture result in studies in Hong Kong and the three month culture result in East African studies are good surrogate endpoints.

8.5 Future Work

8.5.1 Further Analyses of these Data

The data used in this thesis contain information about 6,974 patients given one of 49 different treatment regimens in twelve clinical trials conducted in East Africa and East Asia across two decades.

These data are invaluable in learning more about prognostic and surrogate

endpoints for poor outcome in the treatment of TB. In addition to numerous other secondary analyses of clinical trial data, such rich data provide plenty of opportunity for further evaluation of prognostic and surrogate markers exploring effects of treatment adherence, time to poor outcome, smear results during treatment and drug resistance patterns post-randomisation. These were all beyond the scope of this thesis but could yield interesting results.

As new statistical methods for evaluating surrogate markers relevant to the disease area of TB are developed, these data could again be used for evaluating culture results as surrogate markers. Though there are TB treatment clinical trials currently being conducted with more to start in the near future, it is unlikely that such a large quantity of data will be available for the purpose of formally evaluating surrogate markers for at least a decade or more.

8.5.2 Further Simulation Studies

8.5.2.1 Simplifying the Data Structure

Further simulation studies could be conducted to identify the exact causes for the bias in the parameter estimates found in Chapter 7. As described earlier, the data structure used was fairly complex. This was necessary to closely model the data from the MRC TB trials as analysed in Chapter 6. Nevertheless, simpler structures could be used to evaluate the methods and identify the source of bias. The process could then be repeated with added complexity, working towards the full data structure as described in Chapter 7. Strategies could include: using only two-arm trials, setting $\delta = 0$, drawing the α_{ij} and β_{ij} from simpler distributions with narrower ranges and just using a fixed effects model (removing the random effect). If this simpler structure results in smaller biases, the process of adding each component sequentially should allow for diagnosis of the cause of the large biases and could provide a means for adjusting the model to remove these biases.

8.5.2.2 Estimating R^2_{true}

Further simulation studies would also allow the calculation of the true R^2 value for each of the scenarios which would allow for comparison of the estimated with the true R^2 values. This could be done by regressing the true value of β_{ij} on α_{ij} to give an estimate of the true R^2 value to compare with the estimates of R^2 calculated in Chapter 7. Unfortunately, the original simulated

data was not retained from the simulation study and calculation of the true R^2 values for each scenario used in Chapter 7 is not possible, nor is it possible to report confidence intervals on the estimates of the R^2 values. The standard error of the estimates on R^2 could be calculated using bootstrap techniques. Failing to interpret the estimates of the R^2 values without reference to either any measure of uncertainty in their estimation or the true values, as was done in Chapter 7, is a limitation. A point estimate without a corresponding measure of uncertainty (such as a 95% confidence interval) can falsely suggest a high level of precision. An R^2 of 0.81 suggests a good surrogate, but if the 95% confidence interval on this value is $(0.47, 0.96)$ then there is in fact insufficient evidence to conclude that the marker in question is a good or a poor surrogate. Similarly, if the method yields an R^2 of 0.35 in a simulation study when the true R^2 is 0.81, deficient methods are indicated rather than a deficient surrogate.

8.5.3 Other Markers

In section 3.5.5, Serial Sputum Colony Counts (SSCC) taken during the first eight weeks of treatment were introduced as being a candidate surrogate. Using data from one phase II trial, it has been shown that the second slope parameter in a bi-exponential mixed effects model distinguishes well between two four-drug treatment regimens (Rustomjee et al., 2008b). At the time of publication of the results of this phase II study, patients had not been followed up for long-term treatment outcome, so there was no evidence to determine whether or not this parameter reflects differences in rates of poor outcome. There is a possibility that these patients will be followed up for poor outcome, and this evidence may yet become available.

Table 3.2 on page 94 and 3.3 on page 95 contain a complete list of possible biomarkers for tuberculosis resulting from the joint TDR/EC expert consultation mentioned in section 3.5.7. Any number of these could prove to be effective surrogate markers for outcome to treatment for tuberculosis but, as with SSCC, not enough data is available to formally evaluate any of these markers as surrogates.

The results from this thesis have demonstrated that culture result do capture some of the treatment effect and it is likely that some aspect of the culture results during treatment could prove to be useful surrogate endpoints. Further work is needed to determine whether bi-exponential modelling of SSCC,

survival analysis of time to culture negativity, some other approach or a completely different biomarker will be a useful surrogate. It is expected that some data will be available from trials over the next few years for use in beginning this process.

8.5.4 The Meta-Analytic Methods

A better solution than the two extensions proposed in Chapter 7 to the drawbacks of the meta-analysis developed in Chapter 6 is required. There are two areas for further research. Firstly, different approaches to modelling the joint distribution of the binary true and surrogate endpoints are needed. Secondly, a full random effects model is needed to properly account for the hierarchical structure of the data with trials with multiple treatment comparisons. These are important areas of future research. Reliable statistical methods must be in place to be used for evaluating new candidate surrogate markers as the data become available.

8.5.5 Individual-level Surrogacy

The trial-level surrogacy of a marker corresponds to the degree to which the treatment effect on the marker can be used to predict the treatment effect on the true endpoint. The individual-level surrogacy of marker corresponds to the association between the marker and the true endpoint adjusted for the treatment. Tibaldi et al. (2003) observe that if the true and surrogate endpoints are modelled separately (as has been done in the meta-analysis in this thesis), it is harder to study the individual-level surrogacy.

Trial-level surrogacy is of most importance to trialists and has therefore been the focus of the evaluation of culture results during treatment as surrogates for poor outcome in this thesis. Nevertheless, there has been much work in recent years developing robust methods for assessing individual-level surrogacy (including applications of information theory, see section 2.8.2). These methods came too late for use in this thesis, but exploring the individual-level surrogacy of culture results as surrogates for poor outcome is another area for future research.

245

8.6 Summary of Conclusions

- *Chapter 5.* A strong association was found between culture results during treatment and poor outcome. Nevertheless, culture results were not good patient-specific predictors of poor outcome with low sensitivities and specificities.

- *Chapter 6.* The two month culture was found to be a good surrogate marker using data from trials conducted in Hong Kong and the three month culture was found to be a good surrogate marker using data from East African studies. This is an indication that culture results during treatment do capture some of the treatment effect, and more work is needed in understanding the differences between the Hong Kong and East African trials.

- *Chapter 7.* The meta-analytic methods for evaluating surrogate markers in this thesis included a graphical representation that permitted a clear visual evaluation of the surrogate. The methods involved in modelling the relationship between the treatment effect on the true endpoint and the treatment effect on the surrogate endpoint were deficient, and these deficiencies were not satisfactorily overcome with the two extensions proposed. More work is needed in developing a more appropriate model.

Appendix A

Additional Figures and Tables

Region	Culture Result	Baseline	Month					
			1	2	3	4	5	6
East Africa	Non missing	3179	2842	2873	2856	2863	2845	2889
	Positive	3179 (100%)	1988 (70%)	829 (29%)	292 (10%)	158 (6%)	139 (5%)	111 (4%)
	Negative	0 (0%)	854 (30%)	2044 (71%)	2564 (90%)	2705 (94%)	2706 (95%)	2778 (96%)
Hong Kong	Non missing	3229	3010	2984	2963	3092	3075	3070
	Positive	3229 (100%)	1391 (46%)	445 (15%)	113 (4%)	78 (3%)	61 (2%)	59 (2%)
	Negative	0 (0%)	1619 (54%)	2539 (85%)	2850 (96%)	3014 (97%)	3014 (98%)	3011 (98%)
Singapore	Non missing	504	484	480	482	482	483	490
	Positive	504 (100%)	246 (51%)	34 (7%)	7 (1%)	5 (1%)	9 (2%)	9 (2%)
	Negative	0 (0%)	238 (49%)	446 (93%)	475 (99%)	477 (99%)	474 (98%)	481 (98%)
Overall	Non missing	6912	6336	6337	6301	6437	6403	6449
	Positive	6912 (100%)	3625 (57%)	1308 (21%)	412 (7%)	241 (4%)	209 (3%)	179 (3%)
	Negative	0 (0%)	2711 (43%)	5029 (79%)	5889 (93%)	6196 (96%)	6194 (97%)	6270 (97%)

Table A.1: Rates of culture positivity during treatment by geographical region.

(a) Proportion of Treatment Effect, PTE

(b) Risk Reduction, RR

(c) F

(d) Relative Effect, RE

(e) Adjusted Association, AA

Figure A.1: Single Trial Summary measures and 95% bootstrap confidence intervals for the month 1 culture. Horizontal dashed lines show the region [0,1] in which a proportion should lie. Confidence limits are truncated at +3.0 and -2.0. Confidence intervals are plotted in order of trial along the horizontal axis.

(a) Proportion of Treatment Effect, PTE

(b) Risk Reduction, RR

(c) F
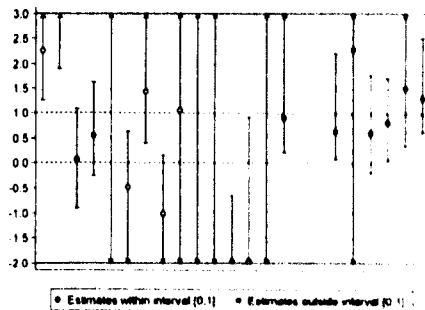
(d) Relative Effect, RE
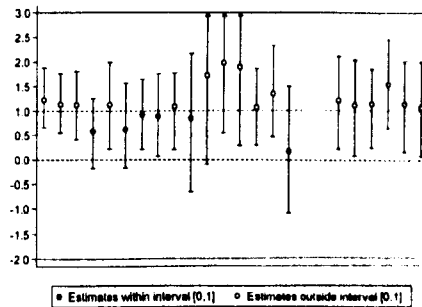
(e) Adjusted Association, AA

Figure A.2: Single Trial Summary measures and 95% bootstrap confidence intervals for the month 2 culture. Horizontal dashed lines show the region [0,1] in which a proportion should lie. Confidence limits are truncated at +3.0 and -2.0. Confidence intervals are plotted in order of trial along the horizontal axis.

250

(a) Proportion of Treatment Effect, PTE

(b) Risk Reduction, RR

(c) F
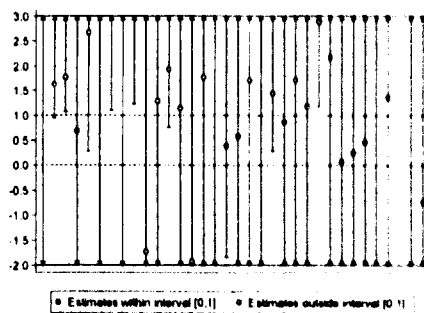
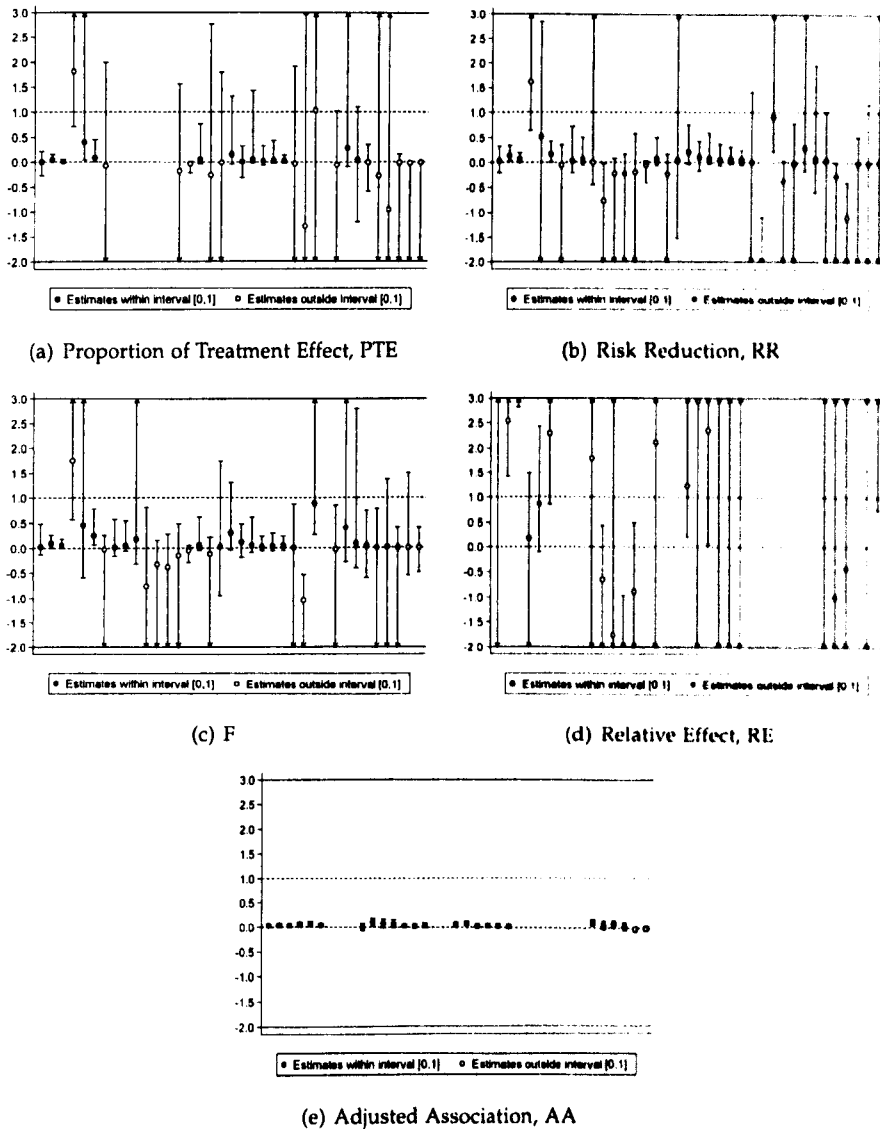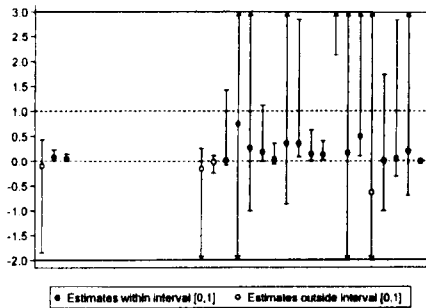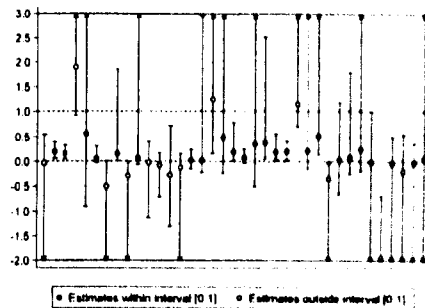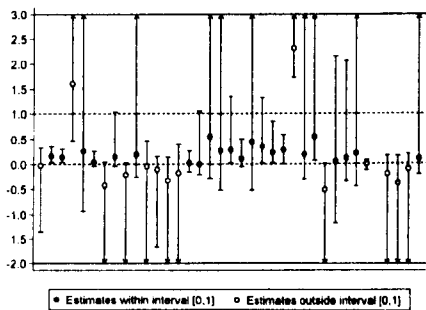(d) Relative Effect, RE
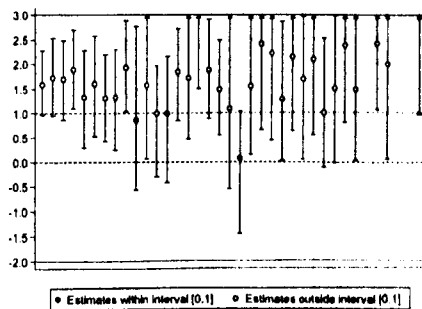
(e) Adjusted Association, AA

Figure A.3: Single Trial Summary measures and 95% bootstrap confidence intervals for the month 3 culture. Horizontal dashed lines show the region [0,1] in which a proportion should lie. Confidence limits are truncated at +3.0 and -2.0. Confidence intervals are plotted in order of trial along the horizontal axis.
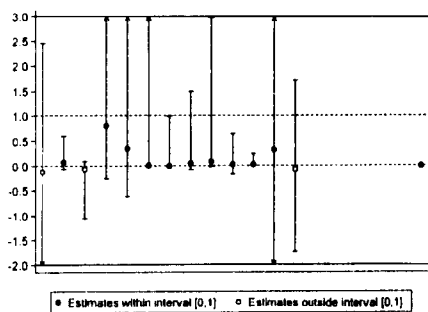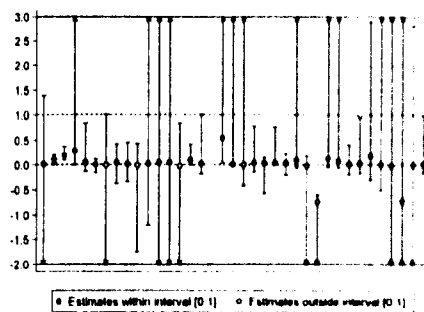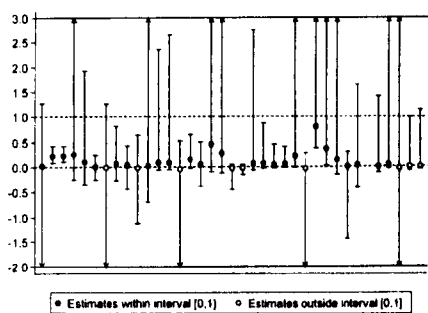
251

Trial	Treatment Comparison	α_{ij} Estimate	α_{ij} 95% CI	β_{ij} Estimate	β_{ij} 95% CI
STUDY R (East Africa)	1	-0.14	(-0.64, 0.35)	-0.4	(-0.90, 0.11)
	2	-0.71	(-1.15, -0.26)	-1.65	(-2.21, -1.08)
	3	-0.53	(-0.98, -0.08)	-2.53	(-3.27, -1.78)
STUDY T (East Africa)	1	-0.58	(-0.99, -0.18)	-0.08	(-0.70, 0.54)
	2	-0.74	(-1.16, -0.33)	-0.58	(-1.29, 0.13)
	3	-0.5	(-0.91, -0.09)	-0.91	(-1.68, -0.14)
STUDY U (East Africa)	1	0.02	(-0.53, 0.57)	-0.29	(-0.94, 0.37)
	2	-0.11	(-0.67, 0.44)	-0.74	(-1.47, -0.02)
	3	-0.35	(-0.91, 0.22)	-0.84	(-1.56, -0.11)
STUDY X (East Africa)	1	0.24	(-0.70, 1.18)	0.42	(-0.60, 1.43)
	2	1.23	(0.37, 2.09)	-0.92	(-2.30, 0.46)
	3	0.78	(-0.09, 1.65)	-1.53	(-3.13, 0.06)
	4	0.82	(-0.07, 1.70)	-2.14	(-4.25, -0.02)
STUDY Y (East Africa)	1*	0.47	(0.01, 0.92)	-0.3	(-0.98, 0.37)
	2*	0.23	(-0.23, 0.68)	-1.62	(-2.62, -0.62)
TANZ (East Africa)	1*	-0.45	(-1.15, 0.26)	-0.9	(-1.69, -0.12)
HONG KONG 1	1	0.06	(-0.62, 0.74)	-0.29	(-1.04, 0.46)
	2	-0.05	(-0.72, 0.61)	-0.31	(-1.04, 0.42)
HONG KONG 2	1	-0.62	(-1.23, -0.01)	-0.81	(-1.48, -0.14)
	2	-0.37	(-0.97, 0.24)	-1.12	(-1.87, -0.37)
HONG KONG 3	1	-0.36	(-0.75, 0.03)	-0.81	(-1.53, -0.09)
	2	-0.24	(-0.63, 0.16)	-0.9	(-1.63, -0.16)
	3	-0.33	(-0.72, 0.06)	-1.12	(-1.91, -0.33)
	4	-0.56	(-0.96, -0.16)	-1.97	(-3.04, -0.89)
HONG KONG 4	1	0.02	(-0.43, 0.47)	-0.03	(-0.70, 0.65)
	2	0.11	(-0.34, 0.56)	-0.1	(-0.79, 0.59)
	3	0.31	(-0.13, 0.76)	-0.1	(-0.79, 0.59)
	4	-0.03	(-0.48, 0.43)	-0.13	(-0.82, 0.56)
	5	0.03	(-0.41, 0.48)	-0.33	(-1.04, 0.38)
	6	-0.47	(-0.94, 0.00)	-0.53	(-1.29, 0.22)
	7	-0.11	(-0.56, 0.34)	-0.55	(-1.30, 0.21)
SINGAPORE 1	1*	-0.06	(-0.69, 0.57)	-1.53	(-3.74, 0.68)
SINGAPORE 3	1*	0.25	(-0.67, 1.16)	-0.43	(-1.76, 0.90)
	2	0.63	(-0.26, 1.52)	-0.45	(-1.78, 0.88)
	3	0.11	(-0.82, 1.04)	-1.14	(-2.80, 0.51)
	4	-0.25	(-1.23, 0.73)	-1.19	(-2.84, 0.47)
	5	-0.3	(-1.28, 0.68)	-1.94	(-4.09, 0.22)

Table A.2: Results of stage I of the meta-analysis for the candidate surrogate marker of heavy culture positivity at month 1. Starred (*) treatment comparisons are not included in stage II of the analysis as the treatments given in the first month are the same in both regimens in these comparisons. See section 6.1.2 on page 152

Trial	Treatment Comparison	α_{ij} Estimate	95% CI	β_{ij} Estimate	95% CI
STUDY R (East Africa)	1	0.03	(-0.46, 0.52)	-0.4	(-0.90, 0.11)
	2	-0.68	(-1.13, -0.22)	-1.65	(-2.21, -1.08)
	3	-0.74	(-1.20, -0.29)	-2.53	(-3.27, -1.78)
STUDY T (East Africa)	1	-0.89	(-1.31, -0.46)	-0.08	(-0.70, 0.54)
	2	-0.93	(-1.36, -0.50)	-0.58	(-1.29, 0.13)
	3	-0.29	(-0.69, 0.11)	-0.91	(-1.68, -0.14)
STUDY U (East Africa)	1	0.59	(0.01, 1.17)	-0.29	(-0.94, 0.37)
	2	-0.56	(-1.23, 0.11)	-0.74	(-1.47, -0.02)
	3	0.76	(0.18, 1.34)	-0.84	(-1.56, -0.11)
STUDY X (East Africa)	1	0.47	(-0.54, 1.48)	0.42	(-0.60, 1.43)
	2	0.2	(-0.86, 1.26)	-0.92	(-2.30, 0.46)
	3	0.56	(-0.40, 1.52)	-1.53	(-3.13, 0.06)
	4	0.9	(-0.05, 1.85)	-2.14	(-4.25, -0.02)
STUDY Y (East Africa)	1*	0.21	(-0.29, 0.71)	-0.3	(-0.98, 0.37)
	2*	-0.14	(-0.67, 0.38)	-1.62	(-2.62, -0.62)
TANZ (East Africa)	1*	-0.91	(-1.83, 0.00)	-0.9	(-1.69, -0.12)
HONG KONG 1	1	-0.45	(-1.13, 0.22)	-0.29	(-1.04, 0.46)
	2	-0.32	(-0.97, 0.33)	-0.31	(-1.04, 0.42)
HONG KONG 2	1	-1.3	(-2.25, -0.34)	-0.81	(-1.48, -0.14)
	2	-0.54	(-1.33, 0.26)	-1.12	(-1.87, -0.37)
HONG KONG 3	1	-0.95	(-1.47, -0.43)	-0.81	(-1.53, -0.09)
	2	-1.16	(-1.72, -0.60)	-0.9	(-1.63, -0.16)
	3	-0.78	(-1.28, -0.29)	-1.12	(-1.91, -0.33)
	4	-1.52	(-2.15, -0.90)	-1.97	(-3.04, -0.89)
HONG KONG 4	1	0.18	(-0.42, 0.78)	-0.03	(-0.70, 0.65)
	2	-0.29	(-0.96, 0.38)	-0.1	(-0.79, 0.59)
	3	0.05	(-0.57, 0.67)	-0.1	(-0.79, 0.59)
	4	0.14	(-0.47, 0.74)	-0.13	(-0.82, 0.56)
	5	-0.13	(-0.76, 0.50)	-0.33	(-1.04, 0.38)
	6	-0.26	(-0.92, 0.40)	-0.53	(-1.29, 0.22)
	7	-0.43	(-1.10, 0.25)	-0.55	(-1.30, 0.21)
SINGAPORE 1	1*	-0.07	(-2.05, 1.91)	-1.53	(-3.74, 0.68)
SINGAPORE 3	1	1.03	(-0.20, 2.27)	-0.43	(-1.76, 0.90)
	2	0	(-1.45, 1.45)	-0.45	(-1.78, 0.88)
	3	0.41	(-0.93, 1.74)	-1.14	(-2.80, 0.51)
	4	-0.02	(-1.47, 1.42)	-1.19	(-2.84, 0.47)
	5	-0.82	(-2.56, 0.93)	-1.94	(-4.09, 0.22)

Table A.3: Results of stage I of the meta-analysis for the candidate surrogate marker of culture positivity at month 2. Starred (*) treatment comparisons are not included in stage II of the analysis as the treatments given in the first two months are the same in both regimens in these comparisons. See section 6.1.2 on page 152

Trial	Treatment Comparison	α_{ij}		β_{ij}	
		Estimate	95% CI	Estimate	95% CI
STUDY R (East Africa)	1	-0.04	(-0.60, 0.52)	-0.4	(-0.90, 0.11)
	2	-1.03	(-1.63, -0.43)	-1.65	(-2.21, -1.08)
	3	-1.52	(-2.20, -0.83)	-2.53	(-3.27, -1.78)
STUDY T (East Africa)	1	-0.38	(-1.05, 0.30)	-0.08	(-0.70, 0.54)
	2	-0.22	(-0.88, 0.45)	-0.58	(-1.29, 0.13)
	3	0.01	(-0.62, 0.64)	-0.91	(-1.68, -0.14)
STUDY U (East Africa)	1	-0.06	(-0.81, 0.70)	-0.29	(-0.94, 0.37)
	2	-0.22	(-1.02, 0.58)	-0.74	(-1.47, -0.02)
	3	-0.16	(-0.93, 0.61)	-0.84	(-1.56, -0.11)
STUDY X (East Africa)	1	-0.43	(-1.92, 1.05)	0.42	(-0.60, 1.43)
	2	-0.45	(-1.94, 1.03)	-0.92	(-2.30, 0.46)
	3	-1.01	(-2.69, 0.68)	-1.53	(-3.13, 0.06)
	4	-1.65	(-3.83, 0.53)	-2.14	(-4.25, -0.02)
STUDY Y (East Africa)	1	0.1	(-0.67, 0.86)	-0.3	(-0.98, 0.37)
	2	-0.75	(-1.69, 0.18)	-1.62	(-2.62, -0.62)
TANZ (East Africa)	1	0.02	(-1.38, 1.43)	-0.9	(-1.69, -0.12)
HONG KONG 1	1	-0.72	(-1.57, 0.12)	-0.29	(-1.04, 0.46)
	2	-0.63	(-1.44, 0.17)	-0.31	(-1.04, 0.42)
HONG KONG 2	1	-0.52	(-2.33, 1.29)	-0.81	(-1.48, -0.14)
	2	-0.01	(-1.63, 1.61)	-1.12	(-1.87, -0.37)
HONG KONG 3	1	-0.48	(-1.61, 0.66)	-0.81	(-1.53, -0.09)
	2	-1.4	(-2.96, 0.16)	-0.9	(-1.63, -0.16)
	3	-0.73	(-1.95, 0.48)	-1.12	(-1.91, -0.33)
	4	-1.41	(-2.97, 0.16)	-1.97	(-3.04, -0.89)
HONG KONG 4	1	0.02	(-1.05, 1.09)	-0.03	(-0.70, 0.65)
	2	-0.1	(-1.21, 1.01)	-0.1	(-0.79, 0.59)
	3	-0.8	(-2.17, 0.57)	-0.1	(-0.79, 0.59)
	4	-0.56	(-1.81, 0.69)	-0.13	(-0.82, 0.56)
	5	-0.61	(-1.86, 0.63)	-0.33	(-1.04, 0.38)
	6	-0.18	(-1.30, 0.93)	-0.53	(-1.29, 0.22)
	7	-0.37	(-1.53, 0.80)	-0.55	(-1.30, 0.21)
SINGAPORE 1	1	N/A[†]		-1.53	(-3.74, 0.68)
SINGAPORE 3	1	0.02	(-2.78, 2.82)	-0.43	(-1.76, 0.90)
	2	0.76	(-1.68, 3.19)	-0.45	(-1.78, 0.88)
	3	N/A[†]		-1.14	(-2.80, 0.51)
	4	0	(-2.80, 2.80)	-1.19	(-2.84, 0.47)
	5	-0.04	(-2.84, 2.76)	-1.94	(-4.09, 0.22)

[†]No patients allocated to one of the regimens in these comparisons had a positive culture at three months and therefore the log odds ratio of a positive culture at three months could not be estimated. See section 6.1.2 on page 152

Table A.4: Results of stage I of the meta-analysis for the candidate surrogate marker of culture positivity at month 3.

Appendix B

Deriving the Reliability Ratio

B.1 Parameter Estimation

B.1.1 Standard Simple Linear Regression

Consider the equation for simple linear regression of y on x:

$$y_i = \alpha_x + \beta_x x_i + \varepsilon_i, \tag{B.1}$$

where ε_i is the error term is assumed to be $\varepsilon_i \sim N(0, \sigma^2)$, $i = 1, \ldots, n$. The least squares estimation of the parameters α_x and β_x requires minimizing the sum of squares of the *residuals*, ε_i:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \alpha_x - \beta_x x_i)^2. \tag{B.2}$$

Taking the derivative with respect to α and β_x gives the following two estimating equations:

$$\sum_{i=1}^{n} y_i - n\alpha_x - \beta_x \sum_{i=1}^{n} x_i = 0, \tag{B.3}$$

$$\sum_{i=1}^{n} x_i y_i - \alpha_x \sum_{i=1}^{n} x_i - \beta_x \sum_{i=1}^{n} x_i^2 = 0, \tag{B.4}$$

which can be solved in α_x and β_x to give:

$$\hat{\beta}_x = \frac{n\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\text{Cov}(x_i, y_i)}{\text{Var}(x_i)},$$

$$(B.5)$$

$$\hat{\alpha}_x = \bar{y} - \hat{\beta}_x \bar{x}, \qquad (B.6)$$

where $\bar{x}$ and $\bar{y}$ are the means of x and y.

Now, suppose that x was observed with error so that $w = x + u$ was observed rather than x where $u \sim N(0, \sigma_u^2)$ is independent of both x and y (this is the classical additive measurement error model). Regression of y on w and least squares estimation yields the following two estimating equations:

$$\sum_{i=1}^n y_i - n\alpha_w - \beta_w \sum_{i=1}^n (x_i + u_i) = 0, \qquad (B.7)$$

$$\sum_{i=1}^n (x_i + u_i) y_i - \alpha_w \sum_{i=1}^n (x_i + u_i) - \beta_w \sum_{i=1}^n (x_i + u_i)^2 = 0, \qquad (B.8)$$

and therefore:

$$\hat{\beta}_w = \frac{\text{Cov}(x_i + u_i, y_i)}{\text{Var}(x_i + u_i)} \qquad (B.9)$$

$$= \frac{\text{Cov}(x_i, y_i) + \text{Cov}(u_i, y_i)}{\text{Var}(x_i) + \text{Var}(u_i)} \qquad (B.10)$$

$$= \frac{\text{Cov}(x_i, y_i)}{\text{Var}(x_i) + \sigma_u^2}, \quad \text{since } \text{Cov}(u_i, y_i) = 0 \qquad (B.11)$$

$$= \frac{\text{Var}(x_i)}{\text{Var}(x_i) + \sigma_u^2} \hat{\beta}_x. \qquad (B.12)$$

Measurement error in the explanatory variable x therefore results in a biased estimate of the parameter β, reduced by a factor λ_r, the reliability ratio, where:

$$\lambda_r = \frac{\text{Var}(x_i)}{\text{Var}(x_i) + \sigma_u^2}. \qquad (B.13)$$

B.1.1.1 Simple Linear Regression with no Intercept

Now consider the equation for simple linear regression of y on x with no intercept (constraining $\alpha_x = 0$):

$$y_i = \beta_x^* x_i + \epsilon_i, \qquad (B.14)$$

where ϵ_i is the error term as before. The least squares estimation of the parameter β_x requires minimizing the sum of squares of the *residuals*, ϵ_i:

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_x^* x_i)^2. \qquad (B.15)$$

Taking the derivative with respect to β_x gives the following estimating equation:

$$\sum_{i=1}^{n} x_i y_i - \beta_x^* \sum_{i=1}^{n} x_i^2 = 0, \qquad (B.16)$$

and therefore:

$$\hat{\beta}_x^* = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} = \frac{\text{Cov}(x_i, y_i) + \bar{x}\bar{y}}{\text{Var}(x_i) + \bar{x}^2}. \qquad (B.17)$$

Now, as before, introduce measurement error in x, such that $w = x + u$ is observed rather than x, remembering that the mean of u is assumed to be zero. Regression of y on w with no intercept yields the following estimating equation:

$$\sum_{i=1}^{n} (x_i + u_i) y_i - \beta_w^* \sum_{i=1}^{n} (x_i + u_i)^2 = 0, \qquad (B.18)$$

and therefore:

$$\hat{\beta}_w^* = \frac{\sum_{i=1}^{n} (x_i + u_i) y_i}{\sum_{i=1}^{n} (x_i + u_i)^2} = \frac{\sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i u_i}{\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} u_i^2 + 2 \sum_{i=1}^{n} x_i u_i}. \qquad (B.19)$$

Now:

$$\text{Cov}(u_i, x_i) = \text{Cov}(u_i, y_i) = 0, \quad \text{due to independence} \tag{B.20}$$

$$= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(u_i - \bar{u}) \tag{B.21}$$

$$= \frac{1}{n} \sum_{i=1}^{n} x_i u_i + \bar{x}\bar{u} - \frac{1}{n}\bar{u} \sum_{i=1}^{n} x_i - \frac{1}{n}\bar{x} \sum_{i=1}^{n} u_i \tag{B.22}$$

$$= \frac{1}{n} \sum_{i=1}^{n} x_i u_i, \quad \text{since } E(u) = 0. \tag{B.23}$$

Similarly, $\sum_{i=1}^{n} y_i u_i = 0$. Therefore:

$$\hat{\beta}_w^* = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} u_i^2} \tag{B.24}$$

$$= \frac{\text{Cov}(x_i, y_i) + \bar{x}\bar{y}}{\text{Var}(x_i) + \bar{x}^2 + \sigma_u^2 + \bar{u}} \tag{B.25}$$

$$= \lambda_r^* \beta_x^* \quad \text{where } \lambda_r^* = \frac{\text{Var}(x_i) + \bar{x}^2}{\text{Var}(x_i) + \sigma_u^2 + \bar{x}^2} \tag{B.26}$$

Therefore, the reliability ratio for linear regression with no intercept, λ_r^*, is different to that for standard simple linear regression, λ_r. Carroll and Stefanski (1995) show that λ_r can be estimated by:

$$\hat{\lambda}_r = \frac{\hat{\sigma}_w^2 - \hat{\sigma}_u^2}{\hat{\sigma}_w^2}, \tag{B.27}$$

where $\hat{\sigma}_u^2$ is an estimate of σ_u^2 and $\hat{\sigma}_w^2$ is the sample variance of the observed w_i. Since $E(u) = 0$, it follows that λ_r^* can be estimated by:

$$\hat{\lambda}_r^* = \frac{\hat{\sigma}_w^2 - \hat{\sigma}_u^2 + \bar{w}^2}{\hat{\sigma}_w^2 + \bar{w}^2}, \tag{B.28}$$

where $\bar{w}$ is the sample mean of the observed w_i.

B.2 Proportion of Explained Variation

B.2.1 Standard Simple Linear Regression

Considering the same equation for simple linear regression of y on x:

$$y_i = \alpha_x + \beta_x x_i + \varepsilon_i, \tag{B.29}$$

the proportion of explained variation in, R^2, is defined as:

$$R_x^2 = \frac{SS_{reg}}{SS_{tot}}, \tag{B.30}$$

where SS_{reg} is the *regression sum of squares* and SS_{tot} is the *total sum of squares*:

$$SS_{reg} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2, \tag{B.31}$$

$$SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2, \tag{B.32}$$

where $\hat{y}_i$ is the predicted value of y_i from the model, $\hat{y}_i = \hat{\alpha}_x + \hat{\beta}_x x_i$. Let $\text{Var}(y_i) = \frac{1}{n} \sum_{i=1}^{n}(y_i - \bar{y})$ be the variance of y_i. Therefore:

$$R_x^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{B.33}$$

$$= \frac{\sum_{i=1}^{n}(\hat{\alpha}_x + \hat{\beta}_x x_i - \bar{y})^2}{n\,\text{Var}(y_i)} \tag{B.34}$$

$$= \frac{\sum_{i=1}^{n}(\bar{y} - \hat{\beta}_x \bar{x} + \hat{\beta}_x x_i - \bar{y})^2}{n\,\text{Var}(y_i)} \quad \text{from B.6,} \tag{B.35}$$

$$= \frac{\sum_{i=1}^{n} \hat{\beta}_x^2 (x_i - \bar{x})^2}{n\,\text{Var}(y_i)} \tag{B.36}$$

$$= \left(\frac{\text{Cov}(x_i, y_i)}{\text{Var}(x_i)} \right)^2 \frac{n\,\text{Var}(x_i)}{n\,\text{Var}(y_i)} \quad \text{from B.5} \tag{B.37}$$

$$= \frac{\text{Cov}(x_i, y_i)^2}{\text{Var}(x_i)\,\text{Var}(y_i)}. \tag{B.38}$$

Now, as before, suppose that x was observed with error so that $w = x + u$ was observed rather than x where $u \sim N(0, \sigma_u^2)$ is independent of both x and y (this is the classical additive measurement error model). Now, $\hat{y} = \hat{\alpha}_w + \hat{\beta}_w(x_i + u_i)$, SS_{tot} is unchanged, but $SS_{reg(w)} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{\alpha}_w + $

259

$\hat{\beta}_w(x_i + u_i) - \bar{y})^2$. Therefore:

$$R_w^2 = \frac{\sum_{i=1}^n (\bar{y} - \hat{\beta}_w(\bar{x} + \bar{u}) + \hat{\beta}_w x_i + \hat{\beta}_w u_i - \bar{y})^2}{n\,\text{Var}(y_i)} \tag{B.39}$$

$$= \frac{\sum_{i=1}^n \hat{\beta}_w^2 (x_i - \bar{x})^2 + \sum_{i=1}^n \hat{\beta}_w^2 (u_i - \bar{u})^2}{n\,\text{Var}(y_i)} \tag{B.40}$$

$$= \left(\frac{\text{Cov}(x_i, y_i)}{\text{Var}(x_i) + \sigma_u^2} \right)^2 \frac{n\,(\text{Var}(x_i) + \sigma_u^2)}{n\,\text{Var}(y_i)} \quad \text{from B.11} \tag{B.41}$$

$$= \frac{\text{Cov}(x_i, y_i)^2}{(\text{Var}(x_i) + \sigma_u^2)\,\text{Var}(y_i)} \tag{B.42}$$

$$= \lambda_r R_x^2. \tag{B.43}$$

The proportion of explained variation is therefore biased by the same ratio as the estimate of the slope parameter.

B.2.1.1 Simple Linear Regression with no Intercept

Now consider the equation for simple linear regression of y on x with no intercept (constraining $\alpha_x = 0$):

$$y_i = \beta_x^* x_i + \epsilon_i, \tag{B.44}$$

where ϵ_i is the error term as before. The proportion of explained variation in, R^2, is defined as before:

$$R_x^{2*} = \frac{SS_{reg}^*}{SS_{tot}^*}, \tag{B.45}$$

but the SS_{reg} and SS_{tot} are redefined:

$$SS_{tot}^* = \sum_{i=1}^n y_i^2, \tag{B.46}$$

$$SS_{reg}^* = \sum_{i=1}^n \hat{y}_i^2. \tag{B.47}$$

Therefore:

$$R_w^{2*} = \frac{\sum_{i=1}^{n} \hat{y}_i^2}{\sum_{i=1}^{n} y_i^2} \tag{B.48}$$

$$= \frac{\sum_{i=1}^{n} \hat{\beta}_x^{*2} x_i^2}{n\left(\mathrm{Var}(y_i) + \bar{y}^2\right)} \tag{B.49}$$

$$= \left(\frac{\mathrm{Cov}(x_i, y_i) + \bar{x}\bar{y}}{\mathrm{Var}(x_i) + \bar{x}^2}\right)^2 \left(\frac{\sum_{i=1}^{n} x_i^2}{n\left(\mathrm{Var}(y_i) + \bar{y}^2\right)}\right) \quad \text{from B.17} \tag{B.50}$$

$$= \frac{\left(\mathrm{Cov}(x_i, y_i) + \bar{x}\bar{y}\right)^2}{\left(\mathrm{Var}(x_i) + \bar{x}^2\right)\left(\mathrm{Var}(y_i) + \bar{y}^2\right)} \tag{B.51}$$

Now, as before, suppose that x was observed with error so that $w = x + u$ was observed rather than x where $u \sim N(0, \sigma_u^2)$ is independent of both x and y (this is the classical additive measurement error model). SS_{tot} is unchanged, but $SS_{reg(w)} = \sum_{i=1}^{n} \hat{y}_i^2 = \hat{\beta}_w^* \sum_{i=1}^{n} (x_i + u_i)^2$.
Therefore:

$$R_w^{2*} = \frac{\left(\mathrm{Cov}(x_i, y_i) + \bar{x}\bar{y}\right)^2}{\left(\mathrm{Var}(x_i) + \sigma_u^2 + \bar{x}^2\right)\left(\mathrm{Var}(y_i) + \bar{y}^2\right)} \tag{B.52}$$

$$= \left(\frac{\mathrm{Var}(x_i) + \bar{x}^2}{\mathrm{Var}(x_i) + \sigma_u^2 + \bar{x}^2}\right) \frac{\left(\mathrm{Cov}(x_i, y_i) + \bar{x}\bar{y}\right)^2}{\left(\mathrm{Var}(x_i) + \bar{x}^2\right)\left(\mathrm{Var}(y_i) + \bar{y}^2\right)} \tag{B.53}$$

$$= \lambda_r^* R_x^{2*} \tag{B.54}$$

The proportion of explained variation is therefore biased by the same ratio as the estimate of the slope parameter.

Appendix C

Glossary

ART *Antiretroviral Treatment* describes the class of drugs used to treat HIV infection.

Biomarker A biomarker is a marker that is objectively measured and is used as an indicator of disease or disease progression.

CFU A quantitative bacteriological culture method of a sputum sample yields a number of *Colony Forming Units*—which is a measure of the bacillary burden of the sample.

Consumption. The name given to tuberculosis before the era of effective treatment for TB as the disease was said to draw the life out of a sufferer or consume them.

Continuation Phase. The last four months of a six month regimen. It follows the *intensive phase* as a continuation of only two (isoniazid and rifampicin) of the four drugs given for the first two months.

CRF *Case Report Forms* are the paper or electronic forms on which data is captured in a clinical trial and are considered as the primary source documents.

Culture A microbiological culture is a method of multiplying mycobacteria (in the case of TB) by letting them reproduce in predetermined culture media under controlled laboratory conditions. Cultures are used to diagnose TB and quantify the number of *colony forming units* (CFUs) in the sample. They can also be used to distinguish between *Mycobacteria tuberculosis* and other *non-tuberculous mycobacteria* and to determine the drug resistance pattern.

Cure   A TB patient is classified as a *cure* if they show a favourable response to treatment and do not show evidence of relapse after the end of treatment.

DOTS   The *Directly Observed Treatment Strategy* is multi-faceted treatment strategy recommended by the WHO to improve case detection, adherence and treatment outcomes.

Double Data Entry   *Double Data Entry (DDE)* is the process of entering data from a clinical trial onto the computer on two separate occasions using two different data entry clerks. The two different versions of the same data are then compared and any discrepancies identified and checked against the original forms. Double data entry is used to reduce human error in data entry.

Fair Outcome   A *Fair Outcome* to treatment is defined in this thesis as a favourable outcome at the end of treatment in addition to no evidence of relapse during follow-up.

FDA   The *US Food and Drug Administration* are the body in the USA responsible for the regulation of new drugs and medical treatments.

FPF   The *False Positive Fraction* (equal to the *Specificity* subtracted from 1) is the proportion of those with a fair outcome that also have a negative marker value, that is the proportion of those with a fair outcome that are correctly identified as such.

Global Alliance for TB Drug Development   The *Global Alliance for TB Drug Development* is an international partnership of private and public bodies tasked with supporting the different phases of pre-clinical and clinical drug development.

Global Plan   This is the document produced by the *Stop TB Partnership* of the WHO outlining international strategies and objectives designed to tackle the global TB epidemic.

Heavily Positive Culture   A heavily positive culture is defined as a positive culture with the growth of 20 or more colonies.

Intensive Phase.   The first two months of a six month regimen for treating tuberculosis. Consists of four drugs (isoniazid, rifampicin, pyrazinamide and ethambutol) give daily or sometimes thrice- weekly. The intensive phase is followed by the *continuation phase*.

INTERTB   The *International Consortium for Trials of Chemotherapeutic Agents in Tuberculosis* is an international consortium created to evaluate the clinical and bacteriological outcomes of chemotherapeutic agents for the treatment of tuberculosis.

IPD   *Individual Patient Data* includes data from individual study participants and is used in contrast to study summary data.

IUATLD.   The *International Union Against Tuberculosis and Lung Disease*, a worldwide organisation for promoting lung health.

Latent TB   The bacilli that cause TB in an individual can lie dormant with the individual suffering no ill effects of the infection. The bacilli may subsequently activate leading to active TB disease requiring treatment.

MDR-TB.   *Multi-Drug Resistant Tuberculosis*, disease that is resistant to the two most potent anti-TB drugs isoniazid and rifampicin.

MRC, BMRC.   The *British Medical Research Council*, responsible for carrying out many of the early clinical trials demonstrating the efficacy of short-course chemotherapy and determining the most effective combination of drugs.

NPV   The *Negative Predictive Value* is the proportion of those with a negative marker value that will go on to have a fair outcome to treatment.

Phthisis.   The Greek word for tuberculosis.

Point of Dichotomy   The point of dichotomy corresponds to the value at which the continuous marker is dichotomised. Let X be the continuous marker and $\tilde{X}$ be the dichotomised binary marker and c is the point of dichotomy where $\tilde{X} = 0$ if $X < c$ and $\tilde{X} = 1$ if $X \geq c$.

Poor Outcome,   A *Poor Outcome* to treatment is defined in this thesis as either failure at the end of treatment or relapse after successful treatment. Deaths due to respiratory causes or TB-related deaths occurring at the end of treatment or during follow-up are also classified as poor outcomes.

PPV   The *Positive Predictive Value* is the proportion of those with a positive marker value that will go on to have a poor outcome to treatment.

Prognostic Marker   A prognostic marker is one that is predictive of a patients' disease outcome or *prognosis*.

PTB Pulmonary tuberculosis.

RCT A *randomised controlled-trial* is a study designed to compare one or more experimental interventions with a control intervention where the allocation of interventions to patients follows no pattern.

Regimen It has been shown that treatment of TB with a single drug (*monotherapy*) in ineffective as it leads to drug resistance. TB is therefore treatment with a combination of up to four drugs taken together to protect against the development of drug resistant. This combination of drugs is known as a *regimen*.

Recurrence, Relapse, Reinfection. Recurrence of disease is a subsequent episode of TB after initial favourable response to treatment. Such recurrences can be separated into *relapse* corresponding to endogenous reactivation of disease and *exogenous reinfection* caused by reinfection with a new strain. See section 3.4.1 on page 73 for discussion.

ROC Curve The *Receiver Operating Characteristic curve* is the set of all possible values of the True Positive Fraction (TPF) and False Positive Fraction (FPF), $ROC = \{(FPF(c), TPF(c)); c \in (-\infty, +\infty)\}$, where c is a point of dichotomy of the marker X.

Short-course chemotherapy. Treatment for tuberculosis for only six months or less, to distinguish such treatment from longer courses lasting eighteen months or more that were replaced by short-course chemotherapy following trials showing superiority in the 1970s and 1980s.

Smear A microbiological smear is a method of identifying acid fast bacilli in a sputum sample. Staining methods are used to highlight the mycobacteria for identification by a trained technician using a microscope.

Stop TB Partnership The *Stop TB Partnership* is a group within the WHO tasked with promoting and supporting the movement to tackle the TB epidemic.

Surrogate Endpoint, Surrogate Marker A marker used in a clinical trial to substitute for the true endpoint which is not observed. The marker is usually measured earlier than the true endpoint and must fully capture the treatment effect on the true endpoint to be a valid surrogate (see section 2.2 for a fuller definition).

TB-HIV A patient described as having *TB-HIV* will have the co-infection of HIV and TB.

TB Tuberculosis.

TBTC The *Tuberculosis Trials Consortium* is a consortium of individuals and research organisations involved in conducting clinical trials in TB. The team responsible for coordinating the consortium is placed within the Division of TB Elimination (DBTE) in the US Center for Disease Control and Prevention (CDC).

TPF The *True Positive Fraction* (equal to the *Sensitivity*) is the proportion of those with a poor outcome that also have a positive marker value, that is the proportion of those with a poor outcome that are correctly identified as such.

WHO The *World Health Organisation*, the body appointed by the United Nations responsible for global health.

XDR-TB *eXtensively-Drug Resistant Tuberculosis*, disease that is resistant to any fluoroquinolone, and at least one of three injectable second-line drugs (capreomycin, kanamycin, and amikacin) as well as isoniazid and rifampicin.

Appendix D

Notation List

It has been intended that notation used throughout this thesis is consistent. Below is an explanation of the main elements used.

- $i = 1, \ldots, N$ is the index corresponding to *trial*, where there are N trials.

- $j = 1, \ldots, m_i$ is the index corresponding to *treatment regimen* within trial i, or the *treatment comparison* of treatment j with the control regimen (identified by $j = 0$). There are m_i treatment comparisons in trial i (corresponding to $m_i + 1$ treatment regimens).

- $k = 1, \ldots, n_i$ is the index corresponding to *individual* within trial i, where there are n_i individuals in trial i.

- n_{ij} is the total number of individuals in trial i allocated to treatment regimen j. $\sum_{j=0}^{m_i} n_{ij} = n_i$

- Z_{ijk} is the *treatment regimen indicator variable* denoting individual k in trial i. $Z_{ijk} = 1$ if individual k is included in treatment comparison j (given either treatment regimen j or the control regimen) and $Z_{ijk} = 0$ otherwise.

- T_{ik} is the value of the *true endpoint* for individual k in trial i.

- S_{ik} is the value of the *surrogate endpoint* for individual k in trial i.

- $\text{logit}(p) = \log_e(p)\text{-}\log_e(1 - p)$ is the logistic function. All logarithms referred to in this thesis are base e.

- $\hat{x}$ is an estimate of a parameter x.

- $\hat{x}_{RC}$ is the *regression calibration* estimate of parameter x.

- $\hat{x}_{SIMEX}$ is the estimate of parameter x resulting from the *SIMEX algorithm*.

267

- λ_r is the attenuating factor known as the *Reliability Ratio*.

- G_i^2 is the contribution of the ith term to the log likelihood ratio test statistic, denoted by G^2.

- α_{ij} and β_{ij} are the treatment effects on the surrogate endpoint and true endpoint respectively for treatment comparison j in trial i, expressed as log odds ratios for binary true and surrogate endpoints, T_{ik} and S_{ik}.

- δ and κ are the intercept and slope parameters for the regression of β_{ij} on α_{ij}.

- R^2 is the proportion of variation in the independent variable that is explained by the dependent variables.

Bibliography

C. Abe, S. Hosojima, Y. Fukasawa, Y. Kazumi, M. Takahashi, K. Hirano, and T. Mori. Comparison of MB-CHECK, BACTEC, and egg-based media for recovery of mycobacteria. *J Clin Microbiol*, 30:878–881, 1992.

V. R. Aber and A. J. Nunn. Short term chemotherapy of tuberculosis. factors affecting relapse following short term chemotherapy. *Bull Int Union Tuberc*, 53:276–80, 1978.

V. R. Aber, B. W. Allen, D. A. Mitchison, P. Ayuma, E. A. Edwards, and A. B. Keyes. Quality-control in tuberculosis bacteriology 1. laboratory studies on isolated positive cultures and the efficiency of direct smear examination. *Tubercle*, 61:123–133, 1980.

J. C. Abrahantes, G. Molenberghs, T. Burzykowski, Z. Shkedy, A. A. Abad, and D. Renard. Choice of units of analysis and modeling strategies in multilevel hierarchical models. *Comput Stat Data An*, 47:537–563, 2004.

K. L. Adams, P. T. Steele, M. J. Bogan, N. M. Sadler, S. I. Martin, A. N. Martin, and M. Frank. Reagentless detection of Mycobacteria tuberculosis H37Ra in respiratory effluents in minutes. *Anal Chem*, 80:5350–5357, 2008.

B. Afghani, J. M. Lieberman, M. B. Duke, and H. R. Stutman. Comparison of quantitative polymerase chain reaction, acid fast bacilli smear, and culture results in patients receiving therapy for pulmonary tuberculosis. *Diagn Microbiol Infect Dis*, 29:73–79, 1997.

D. Agranoff, D. Fernandez-Reyes, M. C. Papadopoulos, S. A. Rojas, M. Herbster, A. Loosemore, E. Tarelli, J. Sheldon, A. Schwenk, R. Pollok, C. F. J. Rayner, and S. Krishna. Identification of diagnostic markers for tuberculosis by proteomic fingerprinting of serum. *Lancet*, 368:1012–1021, 2006.

M. S. Al-Moamary, W. Black, E. Bessuille, R. K. Elwood, and S. Vedal. The significance of the persistent presence of acid-fast bacilli in sputum smears in pulmonary tuberculosis. *Chest*, 116:726–731, 1999.

J. M. Albert, J. P. Ioannidis, P. Reichelderfer, B. Conway, R. W. Coombs, L. Crane, R. Demasi, D. O. Dixon, P. Flandre, M. D. Hughes, L. A. Kalish,

K. Larntz, D. Lin, I. C. Marschner, A. Muñoz, J. Murray, J. Neaton, C. Pettinelli, W. Rida, J. M. Taylor, and S. L. Welles. Statistical issues for HIV surrogate endpoints: point/counterpoint. An NIAID workshop. *Stat Med*, 17:2435–2462, 1998.

R. K. Albert, M. Iseman, J. A. Sbarbaro, A. Stage, and D. J. Pierson. Monitoring patients with tuberculosis for failure during and after treatment. *Am Rev Respir Dis*, 114:1051–1060, 1976.

F. Alcaide, M. A. Benitez, J. M. Escriba, and R. Martin. Evaluation of the BACTEC MGIT 960 and the MB/BacT systems for recovery of mycobacteria from clinical specimens and for species identification by DNA AccuProbe. *J Clin Microbiol*, 38:398–401, 2000.

Algerian working group/British Medical Research Council. Controlled clinical trial comparing a 6-month and a 12-month regimen in the treatment of pulmonary tuberculosis in the Algerian Sahara. Algerian working group/British Medical Research Council cooperative study. *Am Rev Respir Dis*, 129:921–928, 1984.

B. Alisjahbana and R. van Crevel. Improved diagnosis of tuberculosis by better sputum quality. *Lancet*, 369:1908–1909, 2007.

B. W. Allen and D. A. Mitchison. Counts of viable tubercle bacilli in sputum related to smear and culture gradings. *Med Lab Sci*, 49:94–98, 1992.

A. Alonso and G. Molenberghs. Surrogate marker evaluation from an information theory perspective. *Biometrics*, 63:180–186, 2007.

A. Alonso, H. Geys, G. Molenberghs, M. G. Kenward, and T. Vangeneugden. Validation of surrogate markers in multiple randomized clinical trials with repeated measurements. *Biometrical J*, 45:931–945, 2003.

A. Alonso, H. Geys, G. Molenberghs, M. G. Kenward, and T. Vangeneugden. Validation of surrogate markers in multiple randomized clinical trials with repeated measurements: Canonical correlation approach. *Biometrics*, 60: 845–853, 2004a.

A. Alonso, G. Molenberghs, T. Burzykowski, D. Renard, H. Geys, Z. Shkedy, F. Tibaldi, J. C. Abrahantes, and M. Buyse. Prentice's approach and the meta-analytic paradigm: a reflection on the role of statistics in the evaluation of surrogate endpoints. *Biometrics*, 60:724–8, 2004b.

A. Alonso, G. Molenberghs, H. Geys, M. Buyse, and T. Vangeneugden. A unifying approach for surrogate marker validation based on prentice's criteria. *Stat Med*, 25:205–221, 2006.

L. Apers, J. Mutsvangwa, J. Magwenzi, N. Chigara, A. Butterworth, P. Mason, and P. V. der Stuyft. A comparison of direct microscopy, the concentration method and the Mycobacteria Growth Indicator Tube for the examination of sputum for acid-fast bacilli. *Int J Tuberc Lung Dis*, 7:376–381, 2003.

F. Ardito, M. Sanguinetti, L. Sechi, B. Posteraro, L. Masucci, G. Fadda, and S. Zanetti. Comparison of the mycobacteria growth indicator tube with radiometric and solid culture for isolation of mycobacteria from clinical specimens and susceptibility testing of Mycobacterium tuberculosis. *New Microbiol*, 23:151–158, 2000.

S. M. Arend, S. F. T. Thijsen, E. M. S. Leyten, J. J. M. Bouwman, W. P. J. Franken, B. F. P. J. Koster, F. G. J. Cobelens, A.-J. van Houte, and A. W. J. Bossink. Comparison of Two Interferon-gamma Assays and Tuberculin Skin Test for Tracing Tuberculosis Contacts. *Am J Respir Crit Care Med*, 175:618–627, 2007.

S. G. Baker. A simple meta-analytic approach for using a binary surrogate endpoint to predict the effect of intervention on true endpoint. *Biostatistics*, 7:58–70, 2006a.

S. G. Baker. Surrogate endpoints: wishful thinking or reality? *J Natl Cancer Inst*, 98:502–503, 2006b.

S. G. Baker and B. S. Kramer. A perfect correlate does not a surrogate make. *BMC Med Res Methodol*, 3, 2003.

R. Balasubramanian, S. Sivasubramanian, V. K. Vijayan, R. Ramachandran, M. S. Jawahar, C. N. Paramasivan, N. Selvakumar, and P. R. Somasundaram. Five year results of a 3-month and two 5-month regimens for the treatment of sputum-positive pulmonary tuberculosis in south India. *Tubercle*, 71:253–258, 1990.

R. D. Barker and F. J. Millard. Two excellent management tools for national tuberculosis programmes. *Int J Tuberc Lung Dis*, 3:454–455, 1999.

P. F. Barnes. Diagnosing latent tuberculosis infection: the 100-year upgrade. *Am J Respir Crit Care Med*, 163:807–808, 2001.

P. F. Barnes. Diagnosing latent tuberculosis infection: turning glitter to gold. *Am J Respir Crit Care Med*, 170:5–6, 2004.

P. F. Barnes, L. S. Chan, and S. F. Wong. The course of fever during treatment of pulmonary tuberculosis. *Tubercle*, 68:255–260, 1987.

P. F. Barnes, J. M. Leedom, L. S. Chan, S. F. Wong, J. Shah, L. A. Vachon, G. D. Overturf, and R. L. Modlin. Predictors of short-term prognosis in patients with pulmonary tuberculosis. *J Infect Dis*, 158:366–371, 1988.

C. B. Begg and D. H. Y. Leung. On the use of surrogate end points in random-ized trials. *J R Stat Soc Ser A Stat Soc*, 163:15–24, 2000a.

C. B. Begg and D. H. Y. Leung. Comments on the paper by Begg and Leung. *J R Stat Soc Ser A Stat Soc*, 163:24–28, 2000b.

M. A. Behr, S. A. Warren, H. Salamon, P. C. Hopewell, A. P. de Leon, C. L. Daley, and P. M. Small. Transmission of Mycobacterium tuberculosis from patients smear-negative for acid-fast bacilli. *Lancet*, 353:444–449, 1999.

D. Benator, M. Bhattacharya, L. Bozeman, W. Burman, A. Cantazaro, R. Chaisson, F. Gordin, C. R. Horsburgh, J. Horton, A. Khan, C. Lahart, B. Metchock, C. Pachucki, L. Stanton, A. Vernon, M. E. Villarino, Y. C. Wang, M. Weiner, and S. Weis. Rifapentine and isoniazid once a week versus rifampicin and isoniazid twice a week for treatment of drug-susceptible pulmonary tuberculosis in HIV-negative patients: a randomised clinical trial. *Lancet*, 360: 528–534, 2002.

V. W. Berger. Does the prentice criterion validate surrogate endpoints? *Stat Med*, 23:1571–8, 2004.

Biomarker Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther*, 69: 89–95, 2001.

J. A. G. Blanco, I. S. Toste, M. L. Fernández, R. G. Morales, R. F. Alvarez, G. R. Cuadrado, A. M. Gonzálvez, and I. J. G. Martín. Tobacco smoking and sputum smear conversion in pulmonary tuberculosis. *Medicina clinica*, 128:565–568, 2007.

H. M. Blumberg, M. K. Leonard, and R. M. Jasmer. Update on the treatment of tuberculosis and latent tuberculosis infection. *JAMA*, 293:2776–2784, 2005.

B. A. Blumenstein. Medical Research data. *Control Clin Trials*, 16:453–455, 1995.

J. P. Boissel, J. P. Collet, P. Moleur, and M. Haugh. Surrogate endpoints: a basis for a rational approach. *Eur J Clin Pharmacol*, 43:235–44, 1992.

F. J. H. Botha, F. A. Sirgel, D. P. Parkin, B. W. vandeWal, P. R. Donald, and D. A. Mitchison. Early bactericidal activity of ethambutol, pyrazinamide and the fixed combination of isoniazid, rifampicin and pyrazinamide (rifater) in patients with pulmonary tuberculosis. *S. Afr. Med. J.*, 86:155–158, 1996.

R. A. M. Breen, G. A. D. Hardy, F. M. R. Perrin, S. Lear, S. Kinloch, C. J. Smith, I. Cropley, G. Janossy, and M. C. I. Lipman. Rapid diagnosis of smear-negative tuberculosis using immunology and microbiology with induced sputum in HIV-infected and uninfected individuals. *PLoS ONE*, 2, 2007.

R. Brindle, J. Odhiambo, and D. Mitchison. Serial counts of Mycobacterium tuberculosis in sputum as surrogate markers of the sterilising activity of rifampicin and pyrazinamide in treating pulmonary tuberculosis. *BMC Pulm Med*, 1, 2001.

R. J. Brindle, P. P. Nunn, W. Githui, B. W. Allen, S. Gathua, and P. Waiyaki. Quantitative bacillary response to treatment in HIV-associated pulmonary tuberculosis. *Am Rev Respir Dis*, 147:958–961, 1993.

British Medical Research Council. Long-term chemotherapy in the treatment of chronic pulmonary tuberculosis with cavitation. *Tubercle*, 43:201–267, 1962.

British Thoracic Society. A controlled trial of 6 months' chemotherapy in pulmonary tuberculosis. Final report: results during the 36 months after the end of chemotherapy and beyond. *Br J Dis Chest*, 78:330–336, 1984.

H. C. Bucher, G. H. Guyatt, D. J. Cook, A. Holbrook, and F. A. McAlister. Users' Guides to the Medical Literature: XIX. Applying Clinical Trial Results; A. How to Use an Article Measuring the Effect of an Intervention on Surrogate End Points. *JAMA*, 282:771–778, 1999.

W. Burman, D. Benator, A. Vernon, A. Khan, B. Jones, C. Silva, C. Lahart, S. Weis, B. King, B. Mangura, M. Weiner, W. El-Sadr, and T. T. Consortium. Acquired rifamycin resistance with twice-weekly treatment of HIV-related tuberculosis. *Am J Respir Crit Care Med*, 173:350–356, 2006a.

W. Burman, D. McNeeley, L. H. Moulton, M. Spigelman, and A. Vernon. Advancing the science in clinical trials for new TB drugs. *Int J Tuberc Lung Dis*, 12:111–112, 2008.

W. J. Burman. The hunt for the elusive surrogate marker of sterilizing activity in tuberculosis treatment. *Am J Respir Crit Care Med*, 167:1299–301, 2003.

W. J. Burman, S. Goldberg, J. L. Johnson, G. Muzanye, M. Eagle, A. W. Mosher, S. Choudhri, C. L. Daley, S. S. Munsiff, Z. Zhao, A. Vernon, and R. E. Chaisson. Moxifloxacin versus ethambutol in the first 2 months of treatment for pulmonary tuberculosis. *Am J Respir Crit Care Med*, 174:331–338, 2006b.

A. Burton, D. G. Altman, P. Royston, and R. L. Holder. The design of simulation studies in medical statistics. *Stat Med*, 25:4279–4292, 2006.

T. Burzykowski. Surrogate endpoints: wishful thinking or reality? *Stat Methods Med Res*, 17:463–466, 2008.

T. Burzykowski and G. Molenberghs. Validation of surrogate end points in multiple randomized clinical trials with failure time end points. *J R Stat Soc Ser C Appl Stat*, 50:405–422, 2001.

T. Burzykowski, G. Molenberghs, and M. Buyse. The validation of surrogate end points by using data from randomized clinical trials: a case-study in advanced colorectal cancer. *J R Stat Soc Ser A Stat Soc*, 167:103–124, 2004.

T. Burzykowski, G. Molenberghs, and M. E. Buyse. *The Evaluation of Surrogate Endpoints*. Statistics for Biology and Health. Springer, New York, 2005.

M. Buyse and G. Molenberghs. Criteria for the validation of surrogate end-points in randomized experiments. *Biometrics*, 54:1014–29, 1998.

M. Buyse, G. Molenberghs, T. Burzykowski, D. Renard, and H. Geys. The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics*, 1:49–67, 2000a.

M. Buyse, G. Molenberghs, T. Burzykowski, D. Renard, and H. Geys. Statistical validation of surrogate endpoints: problems and proposals. *Drug Inf J*, 34:447–454, 2000b.

M. Buyse, P. Thirion, R. W. Carlson, T. Burzykowski, G. Molenberghs, and P. Piedbois. Relation between tumour response to first-line chemotherapy and survival in advanced colorectal cancer: a meta-analysis. *Lancet*, 356: 373–378, 2000c.

P. W. Bycott and J. M. Taylor. An evaluation of a measure of the proportion of the treatment effect explained by a surrogate marker. *Control Clin Trials*, 19: 555–68, 1998.

J. A. Caminero. Treatment of multidrug-resistant tuberculosis: evidence and controversies. *Int J Tuberc Lung Dis*, 10:829–837, 2006.

R. Camp, R. Jefferys, T. Swan, and J. Syed. *What's in the Pipeline: New HIV Drugs, Vaccines, Microbicides, HCV and TB Therapies in Clinical Trials.* Treatment Action Group, New York, NY, USA, 2006.

J. Carpenter and J. Bithell. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Stat Med*, 19:1141–1164, 2000.

D. Carroll, R.J.; Ruppert and L. A. Stefanski. *Measurement error in nonlinear models.* Chapman & Hall, London, 1995.

N. Carroll, A. H. P. Uys, K. Lawrence, F. S. C. Pheiffer, K. Duncan, N. Beyers, and P. van Helden. Prediction of delayed treatment response in pulmonary tuberculosis: Use of time to positivity values of BACTEC cultures. *Tuberculosis*, 88:624–630, 2008.

M. Casenghi and T. von Schoen-Angerer. Development Of New Drugs For TB Chemotherapy: Analysis Of The Current Drug Pipeline. Médecins Sans Frontières, Geneva, Switzerland, 2006.

K. G. Castro and D. E. Snider. The good news and the bad news about multidrug-resistant tuberculosis. *Clin Infect Dis*, 21:1265–1266, 1995.

R. E. Chaisson. Tuberculosis chemotherapy - still a double-edged sword. *Am J Respir Crit Care Med*, 167:1461–1462, 2003.

S. L. Chan, W. W. Yew, W. K. Ma, D. J. Girling, V. R. Aber, D. Felmingham, B. W. Allen, and D. A. Mitchison. The early bactericidal activity of rifabutin measured by sputum viable counts in Hong-Kong patients with pulmonary tuberculosis. *Tuber Lung Dis*, 73:33–38, 1992.

K. C. Chang, C. C. Leung, W. W. Yew, S. C. Ho, and C. M. Tam. A nested case-control study on treatment-related risk factors for early relapse of tuberculosis. *Am J Respir Crit Care Med*, 170:1124–1130, 2004.

K. C. Chang, C. C. Leung, W. W. Yew, S. L. Chan, and C. M. Tam. Dosing schedules of 6-month regimens and relapse for pulmonary tuberculosis. *Am J Respir Crit Care Med*, 174:1153–1158, 2006.

E. Check. After decades of drought, new drug possibilities flood TB pipeline. *Nat Med*, 13:266, 2007.

C. Chen, H. Wang, and S. M. Snapinn. Proportion of treatment effect (pte) explained by a surrogate marker. *Stat Med*, 22:3449–59, 2003.

H. Chen, Z. Geng, and J. Z. Jia. Criteria for surrogate end points. *J R Stat Soc Ser B Stat Meth*, 69:919–932, 2007.

T. T. Chen, R. M. Simon, E. L. Korn, S. J. Anderson, A. S. Lindblad, H. S. Wieand, H. O. Douglass, B. Fisher, J. M. Hamilton, and M. A. Friedman. Investigation of disease-free survival as a surrogate endpoint for survival in cancer clinical trials. *Commun Stat Theor Meth*, 27:1363–1378, 1998.

C.-L. Cheng and J. W. V. Ness. On estimating linear relationships when both variables are subject to errors. *J R Stat Soc Ser B Stat Meth*, 56:167–183, 1994.

C.-Y. Chiang and L. W. Riley. Exogenous reinfection in tuberculosis. *Lancet Infect Dis*, 5:629–636, 2005.

N. Chierakul, A. Chaiprasert, N. Tingtoy, W. Arjratanakul, and S. N. Pattanakitsakul. Can serial qualitative polymerase chain reaction monitoring predict outcome of pulmonary tuberculosis treatment? *Respirology*, 6:305–309, 2001.

P. Chirac and E. Torreele. Global framework on essential health r&d. *Lancet*, 367:1560–1561, 2006.

S. Choi, S. W. Lagakos, R. T. Schooley, and P. A. Volberding. CD4+ lymphocytes are an incomplete surrogate marker for clinical progression in persons with asymptomatic HIV infection taking zidovudine. *Ann Intern Med*, 118:674–680, 1993.

D. Christie and E. Tansey, editors. *Short-Course Chemotherapy for Tuberculosis: The transcript of a witness seminar held by the Wellcome Trust Centre for the History of Medicine at UCL, London, 3rd Febuary 2004*, volume 24. Wellcome Trust Centre for the History of Medicine at UCL, 2005.

I. Ciglenecki, J. R. Glynn, A. Mwinga, B. Ngwira, A. Zumla, P. E. M. Fine, and A. Nunn. Population differences in death rates in HIV-positive patients with tuberculosis. *Int J Tuberc Lung Dis*, 11:1121–1128, 2007.

J. M. Cliff, I. N. J. Andrade, R. Mistry, C. L. Clayton, M. G. Lennon, A. P. Lewis, K. Duncan, P. T. Lukey, and H. M. Dockrell. Differential gene expression identifies novel markers of CD4+ and CD8+ T cell activation following stimulation by Mycobacterium tuberculosis. *J Immunol*, 173:485–493, 2004.

A. Cochrane. 1931–1971: A critical review, with particular reference to the medical profession. In G. Teeling-Smith and N. Wells, editors, *Medicines for the Year 2000*, pages 1–11. Office of Health Economics, London, 1979.

W. A. Colburn. Optimizing the use of biomarkers, surrogate endpoints, and clinical endpoints for more efficient drug development. *J Clin Pharmacol*, 40: 1419–1427, 2000.

S. R. Cole, H. Chu, and S. Greenland. Multiple-imputation for measurement-error correction. *Int J Epidemiol*, 35:1074–1081, 2006.

D. Collett. *Modelling Survival Analysis Data in Medical Research*. Chapman & Hall/CRC, Boca Raton, FL, 2nd edition, 2003.

C. H. Collins and J. M. Grange. Mycobacterial disease–old problems, new solutions. *Soc Appl Bacteriol Symp Ser*, 25:vii–viii, 1996.

D. L. Combs, R. J. O'Brien, and L. J. Geiter. Usphs tuberculosis short-course chemotherapy trial 21: effectiveness, toxicity, and acceptability. the report of final results. *Ann Intern Med*, 112:397–406, 1990.

M. B. Conde, A. Efron, C. Loredo, G. R. M. D. Souza, N. P. Graça, M. C. Cezar, M. Ram, M. A. Chaudhary, W. R. Bishai, A. L. Kritski, and R. E. Chaisson. Moxifloxacin versus ethambutol in the initial treatment of tuberculosis: a double-blind, randomised, controlled phase ii trial. *The Lancet*, 373:1183–1189, 2009.

L. E. Connolly, P. H. Edelstein, and L. Ramakrishnan. Why is long-term therapy required to cure tuberculosis? *PLoS Med*, 4, 2007.

J. R. Cook and L. A. Stefanski. Simulation-extrapolation estimation in parametric measurement error models. *J Am Stat Assoc*, 89:1314–1328, 1994.

P. P. Cook, R. A. Maldonado, C. T. Yarnell, and D. Holbert. Safety and completion rate of short-course therapy for treatment of latent tuberculosis infection. *Clin Infect Dis*, 43:271–275, 2006.

G. S. Cooke, S. J. Campbell, S. Bennett, C. Lienhardt, K. P. W. J. McAdam, G. Sirugo, O. Sow, P. Gustafson, F. Mwangulu, P. van Helden, P. Fine, E. G. Hoal, and A. V. S. Hill. Mapping of a novel susceptibility locus suggests a role for MC3R and CTSZ in human tuberculosis. *Am J Respir Crit Care Med*, 178:203–207, 2008.

E. L. Corbett, C. J. Watt, N. Walker, D. Maher, B. G. Williams, M. C. Raviglione, and C. Dye. The growing burden of tuberculosis: global trends and interactions with the HIV epidemic. *Arch Intern Med*, 163:1009–1021, 2003.

M. K. Cowles. Bayesian estimation of the proportion of treatment effect captured by a surrogate marker. *Stat Med*, 21:811–34, 2002.

D. R. Cox. A remark on censoring and surrogate response variables. *J R Stat Soc Ser B Meth*, 45:391–393, 1983.

H. S. Cox, M. Morrow, and P. W. Deutschmann. Long term efficacy of DOTS regimens for tuberculosis: systematic review. *BMJ*, 336:484–487, 2008.

U. G. Dafni and A. A. Tsiatis. Evaluating surrogate markers of clinical outcome when measured with error. *Biometrics*, 54:1445–62, 1998.

J. R. Dale. Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics*, 42:909–17, 1986.

M. Danhof, G. Alvan, S. G. Dahl, J. Kuhlmann, and G. Paintaud. Mechanism-based pharmacokinetic-pharmacodynamic modeling - a new classification of biomarkers. *Pharm Res*, 22:1432–1437, 2005.

V. S. Daniel and T. M. Daniel. Old Testament biblical references to tuberculosis. *Clin Infect Dis*, 29:1557–1558, 1999.

M. J. Daniels and M. D. Hughes. Meta-analysis for the evaluation of potential surrogate markers. *Stat Med*, 16:1965–82, 1997.

S. Das, S. L. Chan, B. W. Allen, D. A. Mitchison, and D. B. Lowrie. Application of dna fingerprinting with is986 to sequential mycobacterial isolates obtained from pulmonary tuberculosis patients in Hong Kong before, during and after short-course chemotherapy. *Tubercle and Lung Disease*, 74:47–51, 1993.

G. R. Davies, R. Brindle, S. H. Khoo, and L. J. Aarons. Use of nonlinear mixed-effects analysis for improved precision of early pharmacodynamic measures in tuberculosis treatment. *Antimicrob Agents Chemother (Bethesda)*, 50:3154–3156, 2006a.

G. R. Davies, S. H. Khoo, and L. J. Aarons. Optimal sampling strategies for early pharmacodynamic measures in tuberculosis. *J Antimicrob Chemother*, 58:594–600, 2006b.

G. R. Davies, P. P. J. Phillips, and A. J. Nunn. Biomarkers and surrogate end points in clinical trials of tuberculosis treatment. *J Infect Dis*, 196:648–649, 2007.

A. L. Davis. *Tuberculosis: A Comprehensive International Approach*, chapter A historical perspective on tuberculosis and its control. Marcel Dekker, New York, 2nd edition, 2000.

277

N. E. Day and S. W. Duffy. Trial design based on surrogate end points–application to comparison of different breast screening frequencies. *J R Stat Soc Ser A Stat Soc*, 159:49–60, 1996.

N. E. Day and S. W. Duffy. Comments on the paper by Begg and Leung. *J R Stat Soc Ser A Stat Soc*, 163:24–28, 2000.

S. Day, P. Fayers, and D. Harvey. Double data entry: what value, what price? *Control Clin Trials*, 19:15–24, 1998.

G. P. de Bruyn G. Mycobacterium vaccae immunotherapy for treating tuberculosis. *Cochrane DB Syst Rev*, 1, 2003.

A. de Francisco. Drug development for neglected diseases. *Lancet*, 360:1102, 2002.

V. De Gruttola and T. Fleming. Viral load and response to treatment of HIV (letter). *N Engl J Med*, 334:1672–1673, 1996.

V. De Gruttola and X. M. Tu. Modelling progression of CD4-lymphocyte count and its relationship to survival time. *Biometrics*, 50:1003–14, 1994.

V. De Gruttola, T. Fleming, D. Y. Lin, and R. Coombs. Perspective: validating surrogate markers–are we being naive? *J Infect Dis*, 175:237–46, 1997.

Delta Coordinating Committee and Virology Group. An evaluation of HIV RNA and CD4 cell count as surrogates for clinical outcome. *AIDS*, 13:565–73, 1999.

L. E. Desjardin, M. D. Perkins, K. Wolski, S. Haun, L. Teixeira, Y. Chen, J. L. Johnson, J. J. Ellner, R. Dietze, J. Bates, M. D. Cave, and K. D. Eisenach. Measurement of sputum Mycobacterium tuberculosis messenger rna as a surrogate for response to chemotherapy. *Am J Respir Crit Care Med*, 160: 203–210, 1999.

N. D. D'Esopo. Clinical trials in pulmonary tuberculosis. *Am Rev Respir Dis*, 125:85–93, 1982.

S. Devadatta, S. Radhakrishna, W. Fox, D. A. Mitchison, S. Rajagopalan, S. Sivasubramanian, and H. Stott. Comparative value of sputum smear examination and culture examination in assessing the progress of tuberculous patients receiving chemotherapy. *Bull World Health Organ*, 34:573–587, 1966.

A. H. Diacon, R. F. Patientia, A. Venter, P. D. van Helden, P. J. Smith, H. McIlleron, J. S. Maritz, and P. R. Donald. Early Bactericidal Activity of High-Dose Rifampin in Patients with Pulmonary Tuberculosis Evidenced by Positive Sputum Smears. *Antimicrob Agents Chemother (Bethesda)*, 51:2994–2996, 2007.

E. M. Dinnett, M. M. Mungall, J. A. Kent, E. S. Ronald, K. E. McIntyre, E. Anderson, A. Gaw, and PROSPER Study Group. Unblinding of trial participants to their treatment allocation: lessons from the Prospective Study of Pravastatin in the Elderly at Risk (PROSPER). *Clin Trials*, 2:254–259, 2005.

O. Diraa, K. Fdany, M. Boudouma, N. Elmdaghri, and M. Benbachir. Assessment of the Mycobacteria Growth Indicator Tube for the bacteriological diagnosis of tuberculosis. *Int J Tuberc Lung Dis*, 7:1010–1012, 2003.

S. Ditlevsen and N. Keiding. A comment on: statistical evaluation of biomarkers as surrogate endpoints: a literature review by C. J. Weir and R. J. Walley. *Stat Med*, 26:1415–1416, 2007.

S. Ditlevsen, U. Christensen, J. Lynch, M. T. Damsgaard, and N. Keiding. The mediation proportion: a structural equation approach for estimating the proportion of exposure effect on outcome explained by an intermediate variable. *Epidemiology*, 16:114–120, 2005.

K. A. Do. Biostatistical approaches for modeling longitudinal and event time data. *Clin Cancer Res*, 8:2473–4, 2002.

R. Doll. Controlled trials: the 1948 watershed. *BMJ*, 317:1217–1220, 1998.

P. R. Donald. The early bactericidal activity of anti-tuberculosis agents. *Int J Tuberc Lung Dis*, 10:591–591, 2006.

P. R. Donald, F. A. Sirgel, F. J. Botha, H. I. Seifart, D. P. Parkin, M. L. Vandenplas, B. W. vandeWal, J. S. Maritz, and D. A. Mitchison. The early bactericidal activity of isoniazid related to its dose size in pulmonary tuberculosis. *Am J Respir Crit Care Med*, 156:895–900, 1997.

P. R. Donald, F. A. Sirgel, T. P. Kanyok, L. H. Danziger, A. Venter, F. J. Botha, D. P. Parkin, H. I. Seifart, B. W. Van de Wal, J. S. Martiz, and D. A. Mitchison. Early bactericidal activity of paromomycin (aminosidine) in patients with smear-positive pulmonary tuberculosis. *Antimicrob Agents Chemother (Bethesda)*, 44:3285–3287, 2000.

C. J. Dore and A. J. Nunn. Bactericidal activity of antituberculosis drugs. *Am J Respir Crit Care Med*, 167:663, 2003.

M. Drum and P. McCullagh. Regression models for discrete longitudinal responses : Comment. *Stat Sci*, 8:300–301, 1993.

L. C. du Toit, V. Pillay, and M. P. Danckwerts. Tuberculosis chemotherapy: current drug delivery approaches. *Respir Res*, 7, 2006.

R. Dubos and J. Dubos. *The White Plague: Tuberculosis, Man, and Society*. Rutgers University Press, New Brunswick, New Jersey, 3rd edition edition, 1996.

W. DuMouchel. Hierarchical bayes linear models for meta-analysis. Technical report, National Institute of Statistical Sciences, 1994.

C. Dye, S. Scheele, P. Dolin, V. Pathania, and M. C. Raviglione. Consensus statement. Global burden of tuberculosis: estimated incidence, prevalence, and mortality by country. WHO Global Surveillance and Monitoring Project. *JAMA*, 282:677–686, 1999.

East African/British Medical Research Council. Controlled clinical trial of short-course (6-month) regimens of chemotherapy for treatment of pulmonary tuberculosis. *Lancet*, 1:1079–1085, 1972.

East African/British Medical Research Council. Controlled clinical trial of four short-course (6-month) regimens of chemotherapy for treatment of pulmonary tuberculosis. Second report. *Lancet*, 1:1331–1338, 1973a.

East African/British Medical Research Council. Isoniazid with thiacetazone (thioacetazone) in the treatment of pulmonary tuberculosis in East Africa. Third Report of Fifth Investigation. A co-operative study in East Africian hospitals, clinics and laboratories with the collaboration of the East African and British Medical Research Council. *Tubercle*, 54:169–179, 1973b.

East African/British Medical Research Council. Controlled clinical trial of four short-course (6-month) regimens of chemotherapy for treatment of pulmonary tuberculosis. *Lancet*, 2:1100–1106, 1974a.

East African/British Medical Research Council. Controlled clinical trial of four short-course (6-month) regimens of chemotherapy for treatment of pulmonary tuberculosis. Third report. *Lancet*, 2:237–240, 1974b.

East African/British Medical Research Council. Controlled clinical trial of four 6-month regimens of chemotherapy for pulmonary tuberculosis. Second report. Second East African/British Medical Research Council Study. *Am Rev Respir Dis*, 114:471–475, 1976.

East African/British Medical Research Council. Results at 5 years of a controlled comparison of a 6-month and a standard 18-month regimen of chemotherapy for pulmonary tuberculosis. *Am Rev Respir Dis*, 116:3–8, 1977.

East African/British Medical Research Council. Controlled clinical trial of five short-course (4-month) chemotherapy regimens in pulmonary tuberculosis. first report of 4th study. East African and British Medical Research Council. *Lancet*, 2:334–338, 1978a.

East African/British Medical Research Council. Controlled clinical trial of four short-course regimens of chemotherapy for two durations in the treatment of pulmonary tuberculosis: first report: Third East African/British Medical Research Councils study. *Am Rev Respir Dis*, 118:39–48, 1978b.

East African/British Medical Research Council. Controlled clinical trial of five short-course (4-month) chemotherapy regimens in pulmonary tuberculosis. First report of 4th study. *Lancet*, 2:334–338, 1978c.

East African/British Medical Research Council. Controlled clinical trial of four short-course regimens of chemotherapy for two durations in the treatment of pulmonary tuberculosis. Second report. Third East African/British Medical Research Council Study. *Tubercle*, 61:59–69, 1980.

East African/British Medical Research Council. Controlled clinical trial of five short-course (4-month) chemotherapy regimens in pulmonary tuberculosis. second report of the 4th study. East African/British Medical Research Councils study. *Am Rev Respir Dis*, 123:165–70, 1981a.

East African/British Medical Research Council. Controlled clinical trial of five short-course (4-month) chemotherapy regimens in pulmonary tuberculosis. Second report of the 4th study. East African/British Medical Research Councils Study. *Am Rev Respir Dis*, 123:165–170, 1981b.

East and Central African/British Medical Research Council. Controlled clinical trial of 4 short-couse regimens of chemotherapy (three 6-month and one 8-month) for pulmonary tuberculosis. *Tubercle*, 64:153–166, 1983.

East and Central African/British Medical Research Council. Controlled clinical trial of 4 short-course regimens of chemotherapy (three 6-month and one 8-month) for pulmonary tuberculosis: final report. East and Central African/British Medical Research Council Fifth Collaborative Study. *Tubercle*, 67:5–15, 1986.

G. A. Ellard, D. R. Ellard, B. W. Allen, D. J. Girling, A. J. Nunn, S. K. Teo, T. H. Tan, H. K. Ng, and S. L. Chan. The bioavailability of isoniazid, rifampin, and pyrazinamide in two commercially available combined formulations designed for use in the short-course treatment of tuberculosis. *Am Rev Respir Dis*, 133:1076–1080, 1986.

S. S. Ellenberg and J. M. Hamilton. Surrogate endpoints in clinical trials - cancer. *Stat Med*, 8:405–413, 1989.

D. Enarson, H. Rieder, T. Arnadottir, and A. Trebucq. Management of tuberculosis: a guide for low income countries. International Union Against Tuberculosis and Lung Disease, Paris, 2000.

D. A. Enarson. Controlling tuberculosis–is it really feasible? *Tuber Lung Dis*, 80:57–59, 2000.

M. D. Epstein, N. W. Schluger, A. L. Davidow, S. Bonk, W. N. Rom, and B. Hanna. Time to detection of Mycobacterium tuberculosis in sputum culture correlates with outcome in patients receiving treatment for pulmonary tuberculosis. *Chest*, 113:379–386, 1998.

L. S. Erdreich and E. T. Lee. Use of relative operating characteristic analysis in epidemiology. a method for dealing with subjective judgement. *Am J Epidemiol*, 114:649–662, 1981.

Extreme Tuberculosis (2006, September 14). New York Times, p. A26.

A. R. Feinstein. Misguided efforts and future challenges for research on "diagnostic tests". *J Epidemiol Community Health*, 56:330–332, 2002.

E. C. Fieller. The biological standardization of insulin. *Supp J R Stat Soc*, 7: 1–64, 1940.

P. E. Fine and P. M. Small. Exogenous reinfection in tuberculosis. *N Engl J Med*, 341:1226–1227, 1999.

A. P. Fitzgerald, V. G. DeGruttola, and F. Vaida. Modelling HIV viral rebound using non-linear mixed effects models. *Stat Med*, 21:2093–2108, 2002.

G. M. Fitzmaurice, N. M. Laird, and A. G. Rotnitzky. Regression models for discrete longitudinal responses. *Stat Sci*, 8:284–299, 1993.

P. Flandre and J. O'Quigley. A two-stage procedure for survival studies with surrogate endpoints. *Biometrics*, 51:969–76, 1995.

P. Flandre and Y. Saidi. Estimating the proportion of treatment effect explained by a surrogate marker. *Stat Med*, 18:107–9, 1999.

T. Fleming, V. DeGruttola, and D. DeMets. Surrogate endpoints. In C. T. Armitage P, editor, *Encyclopaedia of Biostatistics*, volume 6. Wiley, New York, 1998.

T. R. Fleming. Surrogate markers in AIDS and cancer trials. *Stat Med*, 13: 1423–1435, 1994.

T. R. Fleming. Surrogate Endpoints And FDA's Accelerated Approval Process. *Health Aff*, 24:67–78, 2005.

T. R. Fleming and D. L. DeMets. Surrogate end points in clinical trials: Are we being misled? *Ann Intern Med*, 125:605–613, 1996.

T. R. Fleming, R. L. Prentice, M. S. Pepe, and D. Glidden. Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and AIDS research. *Stat Med*, 13:955–68, 1994.

T. R. Fleming, K. Sharples, J. McCall, A. Moore, A. Rodgers, and R. Stewart. Maintaining confidentiality of interim data to enhance trial integrity and credibility. *Clin Trials*, 5:157–167, 2008.

J. Fortún, P. Martín-Dávila, A. Molina, E. Navas, J. M. Hermida, J. Cobo, E. Gómez-Mampaso, and S. Moreno. Sputum conversion among patients with pulmonary tuberculosis: are there implications for removal of respiratory isolation? *J Antimicrob Chemother*, 59:794–798, 2007.

P. B. Fourie, J. J. Ellner, and J. L. Johnson. Whither Mycobacterium vaccae–encore. *Lancet*, 360:1032–1033, 2002.

W. Fox. The current status of short course chemotherapy with particular reference to regimens and mechanisms (author's translation). *Tubercle*, 60:177–190, 1979a.

W. Fox. Chemotherapy of pulmonary tuberculosis: A review. *Chest*, 76:785–796, 1979b.

W. Fox. Whither short-course chemotherapy? *Br J Dis Chest*, 75:331–357, 1981.

W. Fox, G. A. Ellard, and D. A. Mitchison. Studies on the treatment of tuberculosis undertaken by the British Medical Research Council Tuberculosis Units, 1946-1986, with relevant subsequent publications. *Int J Tuberc Lung Dis*, 3:S231–S279, 1999.

C. E. Frangakis and D. B. Rubin. Principal stratification in causal inference. *Biometrics*, 58:21–9, 2002.

G. E. Fraser and D. O. Stram. Regression Calibration in Studies with Correlated Variables Measured with Error. *Am J Epidemiol*, 154:836–844, 2001.

L. S. Freedman. Confidence intervals and statistical power of the 'validation' ratio for surrogate or intermediate endpoints. *J Stat Plan Inference*, 96:143–153, 2001.

L. S. Freedman, B. I. Graubard, and A. Schatzkin. Statistical validation of intermediate endpoints for chronic diseases. *Stat Med*, 11:167–78, 1992.

L. S. Freedman, V. Fainberg, V. Kipnis, D. Midthune, and R. J. Carroll. A new method for dealing with measurement error in explanatory variables of regression models. *Biometrics*, 60:172–181, 2004.

N. Freemantle and M. Calvert. Composite and surrogate outcomes in randomised controlled trials. *BMJ*, 334:756–757, 2007.

M. Freire and G. Roscigno. Joining forces to develop weapons against TB: together we must. *Bull World Health Organ*, 80:429, 2002.

T. R. Frieden, T. R. Sterling, S. S. Munsiff, C. J. Watt, and C. Dye. Tuberculosis. *Lancet*, 362:887–899, 2003.

L. Friedman and S. Yusuf. Surrogate endpoints in clinical-trials. *Control Clin Trials*, 6:222–222, 1985.

B. V. Frosini. Causality and causal models: A conceptual perspective. *Int Stat Rev*, 74:305–334, 2006.

E. A. Gaensler. The surgery for pulmonary tuberculosis. *Am Rev Respir Dis*, 125:73–84, 1982.

M. H. Gail, R. Pfeiffer, H. C. Van Houwelingen, and R. J. Carroll. On meta-analytic assessment of surrogate outcomes. *Biostatistics*, 1:231–46, 2000.

S. Galbraith, I. C. Marschner, and J. Simes. Missing data methods for the assessment of surrogate outcomes and treatment mechanisms in clinical trial substudies. *Stat Med*, 25:415–431, 2006.

X. F. Gao, L. Wang, G. J. Liu, J. Wen, X. Sun, Y. Xie, and Y. P. Li. Rifampicin plus pyrazinamide versus isoniazid for treating latent tuberculosis infection: a meta-analysis. *Int J Tuberc Lung Dis*, 10:1080–1090, 2006.

L. J. Geiter, R. J. O'Brien, D. L. Combs, and D. E. Snider. United states public health service tuberculosis therapy trial 21: preliminary results of an evaluation of a combination tablet of isoniazid, rifampin and pyrazinamide. *Tubercle*, 68:41–46, 1987.

C. Genest and J. MacKay. The joy of copulas: Bivariate distributions with uniform marginals. *Am Stat*, 40:280–283, 1986.

C. Genest and L.-P. Rivest. Statistical inference procedures for bivariate archimedean copulas. *J Am Stat Assoc*, 88:1034–1043, 1993.

D. Gibson, A. J. Harvey, V. Everett, and M. K. Parmar. Is double data entry necessary? The CHART trials. CHART Steering Committee. Continuous, Hyperfractionated, Accelerated Radiotherapy. *Control Clin Trials*, 15:482–488, 1994.

P. B. Gilbert, H. J. Ribaudo, L. Greenberg, G. Yu, R. J. Bosch, C. Tierney, and D. R. Kuritzkes. Considerations in choosing a primary endpoint that measures durability of virological suppression in an antiretroviral trial. *AIDS*, 14:1961–1972, 2000.

P. B. Gilbert, V. G. DeGruttola, M. G. Hudgens, S. G. Self, S. M. Hammer, and L. Corey. What constitutes efficacy for a human immunodeficiency virus vaccine that ameliorates viremia: issues involving surrogate end points in phase 3 trials. *J Infect Dis*, 188:179–193, 2003.

W. Gilks, S. Richardson, and D. Spiegelhalter. *Markov chain Monte Carlo in practice*. Chapman and Hall, Boca Raton, Fla., 1998.

S. H. Gillespie and B. M. Charalambous. A novel method for evaluating the antimicrobial activity of tuberculosis treatment regimens. *Int J Tuberc Lung Dis*, 7:684–689, 2003.

S. H. Gillespie, R. D. Gosling, and B. M. Charalambous. A reiterative method for calculating the early bactericidal activity of antituberculosis drugs. *Am J Respir Crit Care Med*, 166:31–35, 2002.

M. Güler, E. Unsal, B. Dursun, O. Aydln, and N. Capan. Factors influencing sputum smear and culture conversion time among patients with new case pulmonary tuberculosis. *Int J Clin Pract*, 61:231–235, 2007.

Global Alliance for TB Drug Development. Tuberculosis. Scientific blueprint for tuberculosis drug development. *Tuberculosis (Edinb)*, 81:1–52, 2001.

Global Alliance for TB Drug Development. Pathway to Patients: Charting the Dynamics of the Global TB Drug Market, 2007.

Global Alliance for TB Drug Development. Handbook of anti-tuberculosis agents. Edinburgh, Scotland, 2008.

J. R. Glynn, M. D. Yates, A. C. Crampin, B. M. Ngwira, F. D. Mwaungulu, G. F. Black, S. D. Chaguluka, D. T. Mwafulirwa, S. Floyd, C. Murphy, F. A. Drobniewski, and P. E. M. Fine. DNA fingerprint changes in tuberculosis: reinfection, evolution, or laboratory error? *J Infect Dis*, 190:1158–1166, 2004.

P. Godfrey-Faussett. District-randomized phased implementation: strengthening the evidence base for cotrimoxazole for HIV-positive tuberculosis patients. *AIDS*, 17:1079–1081, 2003.

R. D. Gosling, L. Heifets, and S. H. Gillespie. A multicentre comparison of a novel surrogate marker for determining the specific potency of anti-tuberculosis drugs. *J Antimicrob Chemother*, 52:473–476, 2003.

E. Green, G. Yothers, and D. J. Sargent. Surrogate endpoint validation: statistical elegance versus clinical relevance. *Stat Methods Med Res*, 17:477–486, 2008.

V. D. Gruttola, L. A. Beckett, R. W. Coombs, J. M. Arduino, H. H. Balfour, S. Rasheed, F. B. Hollinger, M. A. Fischl, and P. Volberding. Serum p24 antigen level as an intermediate end point in clinical trials of zidovudine in people infected with human immunodeficiency virus type 1. Aids Clinical Trials Group Virology Laboratories. *J Infect Dis*, 169:713–721, 1994.

V. G. D. Gruttola, P. Clax, D. L. DeMets, G. J. Downing, S. S. Ellenberg, L. Friedman, M. H. Gail, R. Prentice, J. Wittes, and S. L. Zeger. Considerations in the evaluation of surrogate endpoints in clinical trials. summary of a National Institutes of Health workshop. *Control Clin Trials*, 22:485–502, 2001.

T. Gumbo, A. Louie, W. Liu, P. G. Ambrose, S. M. Bhavnani, D. Brown, and G. L. Drusano. Isoniazid's bactericidal activity ceases because of the emergence of resistance, not depletion of Mycobacterium tuberculosis in the log phase of growth. *J Infect Dis*, 195:194–201, 2007.

R. Hafner, J. A. Cohn, D. J. Wright, N. E. Dunlap, M. J. Egorin, M. E. Enama, K. Muth, C. A. Peloquin, N. Mor, L. B. Heifets, N. Dunlap, P. Phillips, R. Campo, P. James, M. Sension, M. Bourie, M. Witt, S. Kruger, D. Mushatt, and D. Greenspan. Early bactericidal activity of isoniazid in pulmonary tuberculosis - optimization of methodology. *Am J Respir Crit Care Med*, 156: 918–923, 1997.

G. E. Hagle, T. M.and Mitchell. Goodness-of-fit measures for probit and logit. *Am J Pol Sci*, 36:762–784, 1992.

J. A. Hanley, A. Negassa, M. D. d. Edwardes, and J. E. Forrester. Statistical analysis of correlated data using generalized estimating equations: An orientation. *Am J Epidemiol*, 157:364–375, 2003.

B. A. Hanna, A. Ebrahimzadeh, L. B. Elliott, M. A. Morgan, S. M. Novak, S. Rusch-Gerdes, M. Acio, D. F. Dunbar, T. M. Holmes, C. H. Rexer, C. Savthyakumar, and A. M. Vannier. Multicenter evaluation of the BACTEC MGIT 960 system for recovery of mycobacteria. *J Clin Microbiol*, 37:748–752, 1999.

J. W. Hardin and J. Hilbe. *Generalized estimating equations*. Chapman & Hall/CRC, Boca Raton, FL, 2003.

A. Harries and C. Dye. Tuberculosis. *Ann Trop Med Parasitol*, 100:415–431, 2006.

P. D. Hart. Chemotherapy of tuberculosis–research during the past 100 years. *Br Med J*, 2:805–855, 1946.

N. Hasegawa, T. Miura, A. Ishizaka, K. Yamaguchi, and K. Ishii. Detection of mycobacteria in patients with pulmonary tuberculosis undergoing chemotherapy using MGIT and egg-based solid medium culture systems. *Int J Tuberc Lung Dis*, 6:447–453, 2002.

T. J. Hellyer, L. E. DesJardin, G. L. Hehman, M. D. Cave, and K. D. Eisenach. Quantitative Analysis of mRNA as a Marker for Viability of Mycobacterium tuberculosis. *J Clin Microbiol*, 37:290–295, 1999.

R. Henderson, P. Diggle, and A. Dobson. Identification and efficacy of longitudinal markers for survival. *Biostatistics*, 3:33–50, 2002.

I. Hershkovitz, H. D. Donoghue, D. E. Minnikin, G. S. Besra, O. Y.-C. Lee, A. M. Gernaey, E. Galili, V. Eshed, C. L. Greenblatt, E. Lemma, G. K. Bar-Gal, and M. Spigelman. Detection and molecular characterization of 9000-year-old Mycobacterium tuberculosis from a neolithic settlement in the eastern mediterranean. *PLoS ONE*, 3, 2008.

J. Herson. The use of surrogate endpoints in clinical-trials (an introduction to a series of 4 papers). *Stat Med*, 8:403–404, 1989.

A. B. Hill. *Principles of medical statistics*. Lancet, London, 1937.

A. B. Hill. Suspended judgment. memories of the British streptomycin trial in tuberculosis. the first randomized clinical trial. *Control Clin Trials*, 11:77–79, 1990.

P. C. Hill, R. H. Brookes, A. Fox, K. Fielding, D. J. Jeffries, D. Jackson-Sillah, M. D. Lugos, P. K. Owiafe, S. A. Donkor, A. S. Hammond, J. K. Otu, T. Corrah, R. A. Adegbola, and K. P. W. J. McAdam. Large-scale evaluation of enzyme-linked immunospot assay and skin test for diagnosis of Mycobacterium tuberculosis infection against a gradient of exposure in The Gambia. *Clin Infect Dis*, 38:966–973, 2004.

P. C. Hill, D. J. Jackson-Sillah, A. Fox, R. H. Brookes, B. C. de Jong, M. D. Lugos, I. M. Adetifa, S. A. Donkor, A. M. Aiken, S. R. Howie, T. Corrah, K. P. McAdam, and R. A. Adegbola. Incidence of tuberculosis and the predictive value of elispot and mantoux tests in gambian case contacts. *PLoS ONE*, 3, 2008.

A. Hillis and D. Seigel. Surrogate endpoints in clinical trials - ophthalmologic disorders. *Stat Med*, 8:427–430, 1989.

H. Hinshaw and W. Feldman. Streptomycin in treatment of pulmonary tuberculosis: a preliminary report. *Proc Staff Meet Mayo Clin*, 20:313–318, 1945.

HIV Surrogate Marker Collaborative Group. Human immunodeficiency virus type 1 RNA level and CD4 count as prognostic markers and surrogate end points: a meta-analysis. *AIDS Res Hum Retroviruses*, 16:1123–33, 2000.

R. S. Hogg, B. Yip, K. J. Chan, E. Wood, K. J. Craib, M. V. O'Shaughnessy, and J. S. Montaner. Rates of disease progression by baseline CD4 cell count and viral load after initiating triple-drug therapy. *JAMA*, 286:2568–2577, 2001.

T. H. Holtz, M. Sternberg, S. Kammerer, K. F. Laserson, V. Riekstina, E. Zarovska, V. Skripconoka, C. D. Wells, and V. Leimane. Time to sputum culture conversion in multidrug-resistant tuberculosis: predictors and relationship to treatment outcome. *Ann Intern Med*, 144:650–659, 2006.

Hong Kong Chest Service/British Medical Research Council. Controlled trial of 6- and 9-month regimens of daily and intermittent streptomycin plus isoniazid plus pyrazinamide for pulmonary tuberculosis in Hong Kong. *Tubercle*, 56:81–96, 1975.

Hong Kong Chest Service/British Medical Research Council. Controlled trial of 6-month and 8-month regimens in the treatment of pulmonary tuberculosis: the results up to 24 months. *Tubercle*, 60:201–210, 1979.

Hong Kong Chest Service/British Medical Research Council. Controlled trial of four thrice-weekly regimens and a daily regimen all given for 6 months for pulmonary tuberculosis. *Lancet*, 1:171–174, 1981.

Hong Kong Chest Service/British Medical Research Council. Controlled trial of 4 three-times-weekly regimens and a daily regimen all given for 6 months for pulmonary tuberculosis. Second report: the results up to 24 months. *Tubercle*, 63:89–98, 1982.

Hong Kong Chest Service/British Medical Research Council. Five-year follow-up of a controlled trial of five 6-month regimens of chemotherapy for pulmonary tuberculosis. *Am Rev Respir Dis*, 136:1339–1342, 1987.

Hong Kong Chest Service/British Medical Research Council. Controlled trial of 2, 4, and 6 months of pyrazinamide in 6-month, three-times-weekly regimens for smear-positive pulmonary tuberculosis, including an assessment of a combined preparation of isoniazid, rifampin, and pyrazinamide. Results at 30 months. *Am Rev Respir Dis*, 143:700–706, 1991a.

Hong Kong Chest Service/British Medical Research Council. A controlled clinical comparison of 6 and 8 months of antituberculosis chemotherapy in the treatment of patients with silicotuberculosis in Hong Kong. Hong Kong Chest Service/tuberculosis Research Centre, Madras/British Medical Research Council. *Am Rev Respir Dis*, 143:262–267, 1991b.

N. J. Horton and S. R. Lipsitz. Review of software to fit generalized estimating equation regression models. *Am Stat*, 53:160–169, 1999.

J. P. Hughes. Mixed effects models with censored data with application to HIV rna levels. *Biometrics*, 55:625–9, 1999.

M. D. Hughes. Evaluating surrogate endpoints. *Control Clin Trials*, 23:703–7, 2002.

M. D. Hughes. The evaluation of surrogate endpoints in practice: Experience in HIV. In G. Molenberghs, M. E. Buyse, and T. Burzykowski, editors, *The Evaluation of Surrogate Endpoints*, pages 295–321. Springer, New York, 2005.

M. D. Hughes, V. DeGruttola, and S. L. Welles. Evaluating surrogate markers. *J Acquir Immune Defic Syndr Hum Retrovirol*, 10:S1–8, 1995.

M. D. Hughes, M. J. Daniels, M. A. Fischl, S. Kim, and R. T. Schooley. CD4 cell count as a surrogate endpoint in HIV clinical trials: a meta-analysis of studies of the AIDS clinical trials group. *AIDS*, 12:1823–32, 1998.

International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals For Human Use. Structure and content of clinical study reports, 1995.

International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals For Human Use. Statistical principles for clinical trials, 1998.

J. P. Ioannidis, J. C. Cappelleri, and J. Lau. Viral load and response to treatment of HIV. *N Engl J Med*, 334:1671–1673, 1996.

M. D. Iseman. An unholy trinity–three negative sputum smears and release from tuberculosis isolation. *Clin Infect Dis*, 25:671–672, 1997.

M. D. Iseman. Tuberculosis therapy: past, present and future. *Eur Respir J Suppl*, 36:87s–94s, 2002.

M. D. Iseman and L. B. Heifets. Rapid detection of tuberculosis and drug-resistant tuberculosis. *N Engl J Med*, 355:1606–1608, 2006.

M. Jacobsen, D. Repsilber, A. Gutschmidt, A. Neher, K. Feldmann, H. J. Mollenkopf, A. Ziegler, and S. H. E. Kaufmann. Candidate biomarkers for discrimination between infection and disease caused by Mycobacterium tuberculosis. *J Mol Med*, 85:613–621, 2007.

R. M. Jasmer, L. Bozeman, K. Schwartzman, M. D. Cave, J. J. Saukkonen, B. Metchock, A. Khan, W. J. Burman, and The Tuberculosis Trials Consortium. Recurrent Tuberculosis in the United States and Canada: Relapse or Reinfection? *Am J Respir Crit Care Med*, 170:1360–1366, 2004.

C. Y. Jeon and M. B. Murray. Diabetes mellitus increases the risk of active tuberculosis: A systematic review of 13 observational studies. *PLoS Med*, 5, 2008.

A. Jindani, V. R. Aber, E. A. Edwards, and D. A. Mitchison. The early bactericidal activity of drugs in patients with pulmonary tuberculosis. *Am Rev Respir Dis*, 121:939–49, 1980.

A. Jindani, C. J. Dore, and D. A. Mitchison. Bactericidal and sterilizing activities of antituberculosis drugs during the first 14 days. *Am J Respir Crit Care Med*, 167:1348–54, 2003.

A. Jindani, A. J. Nunn, and D. A. Enarson. Two 8-month regimens of chemotherapy for treatment of newly diagnosed pulmonary tuberculosis: international multicentre randomised trial. *Lancet*, 364:1244–1251, 2004.

J. L. Johnson, D. J. Hadad, W. H. Boom, C. L. Daley, C. A. Peloquin, K. D. Eisenach, D. D. Jankus, S. M. Debanne, E. D. Charlebois, E. Maciel, M. Palaci, and R. Dietze. Early and extended early bactericidal activity of levofloxacin, gatifloxacin and moxifloxacin in pulmonary tuberculosis. *Int J Tuberc Lung Dis*, 10:605–612, 2006.

R. Joshi, A. L. Reingold, D. Menzies, and M. Pai. Tuberculosis among healthcare workers in low- and middle-income countries: a systematic review. *PLoS Med*, 3, 2006.

M. Kato-Maeda and P. M. Small. Topic in review: How molecular epidemiology has changed what we know about tuberculosis. *West J Med*, 172: 256–259, 2000.

S. H. E. Kaufmann and S. K. Parida. Changing funding patterns in tuberculosis. *Nat Med*, 13:299–303, 2007.

V. P. Keane, N. de Klerk, T. Krieng, G. Hammond, and A. W. Musk. Risk factors for the development of non-response to first-line treatment for tuberculosis in southern vietnam. *Int J Epidemiol*, 26:1115–1120, 1997.

E. Keeler, M. D. Perkins, P. Small, C. Hanson, S. Reed, J. Cunningham, J. E. Aledort, L. Hillborne, M. E. Rafael, F. Girosi, and C. Dye. Reducing the global burden of tuberculosis: the contribution of improved diagnostics. *Nature*, 444:49–57, 2006.

N. Kennedy, R. Fox, G. M. Kisyombe, A. O. Saruni, L. O. Uiso, A. R. Ramsay, F. I. Ngowi, and S. H. Gillespie. Early bactericidal and sterilizing activities of ciprofloxacin in pulmonary tuberculosis. *Am Rev Respir Dis*, 148:1547–1551, 1993.

N. Kennedy, S. H. Gillespie, A. O. S. Saruni, G. Kisyombe, R. McNerney, F. I. Ngowi, and S. Wilson. Polymerase chain-reaction for assessing treatment response in patients with pulmonary tuberculosis. *J Infect Dis*, 170:713–716, 1994.

N. Kennedy, L. Berger, J. Curram, R. Fox, J. Gutmann, G. M. Kisyombe, F. I. Ngowi, A. R. Ramsay, A. O. Saruni, N. Sam, G. Tillotson, L. O. Uiso, M. Yates, and S. H. Gillespie. Randomized controlled trial of a drug regimen that includes ciprofloxacin for the treatment of pulmonary tuberculosis. *Clin Infect Dis*, 22:827–833, 1996.

J. T. Kent. Information gain and a general measure of correlation. *Biometrika*, 70:163–173, 1983.

M. M. Kent and S. Yin. Controlling infectious diseases. *Popul Bull*, 61:3–9, 2006.

Kenyan/Zambian/British Medical Research Council. Controlled clinical trial of levamisole in short-course chemotherapy for pulmonary tuberculosis. A Kenyan/Zambian/British Medical Research Council Collaborative Study. *Am Rev Respir Dis*, 140:990–995, 1989.

A. Khan, T. R. Sterling, R. Reves, A. Vernon, and C. R. Horsburgh. Lack of weight gain and relapse risk in a large tuberculosis treatment trial. *Am J Respir Crit Care Med*, 174:344–348, 2006.

M. S. Khan, O. Dar, C. Sismanidis, K. Shah, and P. Godfrey-Faussett. Improvement of tuberculosis case detection and reduction of discrepancies between men and women by simple sputum-submission instructions: a pragmatic randomised controlled trial. *Lancet*, 369:1955–1960, 2007.

S. S. Kiblawi, S. J. Jay, R. B. Stonehill, and J. Norton. Fever response of patients on therapy for pulmonary tuberculosis. *Am Rev Respir Dis*, 123:20–4, 1981.

L. B. Klebanov and A. Y. Yakovlev. A new approach to testing for sufficient follow-up in cure-rate analysis. *J Stat Plan Inference*, 137:3557–3569, 2007.

N. Konomi, E. Lebwohl, K. Mowbray, I. Tattersall, and D. Zhang. Detection of mycobacterial DNA in Andean mummies. *J Clin Microbiol*, 40:4738–4740, 2002.

E. L. Korn and R. Simon. Measures of explained variation for survival data. *Stat Med*, 9:487–503, 1990.

E. L. Korn, P. S. Albert, and L. M. McShane. Assessing surrogates as trial endpoints using mixed models. *Stat Med*, 24:163–182, 2005.

B. Kreis, S. Pretet, J. Birenbaum, P. Guibout, J. J. Hazeman, E. Orin, S. Perdrizet, and J. Weil. Two three-month treatment regimens for pulmonary tuberculosis. *Bull Int Union Tuberc*, 51:71–75, 1976.

S. W. Lagakos and D. F. Hoth. Surrogate markers in AIDS: where are we? where are we going? *Ann Intern Med*, 116:599–601, 1992.

N. M. Laird. Missing data in longitudinal studies. *Stat Med*, 7:305–15, 1988.

N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38:963–974, 1982.

M.-L. Lambert, E. Hasker, A. V. Deun, D. Roberfroid, M. Boelaert, and P. V. der Stuyft. Recurrence in tuberculosis: relapse or reinfection? *Lancet Infect Dis*, 3:282–287, 2003.

H. R. M. Landis. Disease of the lungs. In G. W. Norris and H. R. M. Landis, editors, *Disease of the Chest and the Principles of Physical Diagnosis*. W. B. Saunders, Philadelphia, 1920.

M. Lassere, K. Johnson, M. Hughes, D. Altman, M. Buyse, S. Galbraith, and G. Wells. Simulation studies of surrogate endpoint validation using single trial and multitrial statistical approaches. *J Rheumatol*, 34:616–619, 2007a.

M. N. Lassere, K. R. Johnson, M. Boers, P. Tugwell, P. Brooks, L. Simon, V. Strand, P. G. Conaghan, M. Ostergaard, W. P. Maksymowych, R. Landewe, B. Bresnihan, P.-P. Tak, R. Wakefield, P. Mease, C. O. Bingham, M. Hughes, D. Altman, M. Buyse, S. Galbraith, and G. Wells. Definitions and validation criteria for biomarkers and surrogate endpoints: development and testing of a quantitative hierarchical levels of evidence schema. *J Rheumatol*, 34:607–615, 2007b.

S. Lawn, N. Bangani, M. Vogt, L.-G. Bekker, M. Badri, M. Ntobongwana, H. Dockrell, R. Wilkinson, and R. Wood. Utility of interferon-gamma ELISPOT assay responses in highly tuberculosis-exposed patients with advanced HIV infection in South Africa. *BMC Infect Dis*, 7, 2007.

J. J. Lee, J. Suo, C. B. Lin, J. D. Wang, T. Y. Lin, and Y. C. Tsai. Comparative evaluation of the BACTEC MGIT 960 system with solid medium for isolation of mycobacteria. *Int J Tuberc Lung Dis*, 7:569–574, 2003.

L. J. Lesko and A. J. Atkinson. Use of biomarkers and surrogate endpoints in drug development and regulatory decision making: criteria, validation, strategies. *Annu Rev Pharmacol Toxicol*, 41:347–366, 2001.

D. H.-Y. Leung. Statistical methods for clinical studies in the presence of surrogate end points. *J R Stat Soc Ser A Stat Soc*, 164:485–503, 2001.

G. Levée, P. Glaziou, B. Gicquel, and S. Chanteau. Follow-up of tuberculosis patients undergoing standard anti-tuberculosis chemotherapy by using a polymerase chain reaction. *Res Microbiol*, 145:5–8, 1994.

X. Li, Y. Zhang, X. Shen, G. Shen, X. Gui, B. Sun, J. Mei, K. Deriemer, P. M. Small, and Q. Gao. Transmission of Drug-Resistant Tuberculosis among Treated Patients in Shanghai, China. *J Infect Dis*, 195:864–869, 2007.

Z. Li and M. P. Meredith. Exploring the relationship between surrogates and clinical outcomes: analysis of individual patient data vs. meta-regression on group-level summary statistics. *J Biopharm Stat*, 13:777–92, 2003.

Z. Li, M. P. Meredith, and M. S. Hoseyni. A method to assess the proportion of treatment effect explained by a surrogate endpoint. *Stat Med*, 20:3175–88, 2001.

C. Lienhardt, K. Manneh, V. Bouchier, G. Lahai, P. J. Milligan, and K. P. McAdam. Factors determining the outcome of treatment of adult smear-positive tuberculosis cases in the gambia. *Int J Tuberc Lung Dis*, 2:712–718, 1998.

C. Lienhardt, S. Cook, V. Yorke-Edwards, M. Burgos, G. Anyo, S. J. Kim, A. Jindani, D. A. Enarson, and A. Nunn. Investigation of the safety and efficacy of a 4-FDC for the treatment of tuberculosis (study c): methods and preliminary results of the 12-month patient follow-up. *Int J Tuberc Lung Dis*, 12: S46–S47, 2008.

D. Y. Lin, T. R. Fleming, and V. De Gruttola. Estimating the proportion of treatment effect explained by a surrogate marker. *Stat Med*, 16:1515–27, 1997.

H. Lin, C. E. McCulloch, and S. T. Mayne. Maximum likelihood estimation in the joint analysis of time-to-event and multiple longitudinal variables. *Stat Med*, 21:2369–2382, 2002.

M. J. Lindstrom and D. M. Bates. Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *J Am Stat Assoc*, 83: 1014–1022, 1988.

M. J. Lindstrom and D. M. Bates. Nonlinear mixed effects models for repeated measures data. *Biometrics*, 46:673–687, 1990.

J. S. Long. *Regression Models for Categorical and Limited Dependent Variables.* Sage Publications, Thousand Oaks, 1997.

J. S. Long and J. Freese. *Regression models for categorical variables using stata.* Stata Press, College Station, TX, 2nd edition, 2006.

R. Long, M. Scalcini, J. Manfreda, M. Jean-Baptiste, and E. Hershfield. The impact of HIV on the usefulness of sputum smears for the diagnosis of tuberculosis. *Am J Public Health*, 81:1326–1328, 1991.

G. Maartens and R. J. Wilkinson. Tuberculosis. *Lancet*, 370:2030–2043, 2007.

D. P. MacKinnon, C. M. Lockwood, C. H. Brown, W. Wang, and J. M. Hoffman. The intermediate endpoint effect in logistic and probit regression. *Clin Trials*, 4:499–513, 2007.

G. Magombedze, W. Garira, and E. Mwenje. Mathematical modeling of chemotherapy of human TB infection. *J Biol Syst*, 14:509–553, 2006.

R. A. Maller and X. Zhou. *Survival Analysis with Long-term Survivors.* Wiley, 1996.

B. Manns, W. F. Owen, W. C. Winkelmayer, P. J. Devereaux, and M. Tonelli. Surrogate markers in clinical studies: Problems solved or created? *Am J Kidney Dis*, 48:159–166, 2006.

J. Marks. Ending the routine guinea-pig test. *Tubercle*, 53:31–34, 1972.

J. N. Matthews, D. G. Altman, M. J. Campbell, and P. Royston. Analysis of serial measurements in medical research. *Br Med J*, 300:230–5, 1990.

Medical Research Council. Clinical trial of patulin in the common cold. *Lancet*, 2:373–375, 1944.

Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. *Br Med J*, 2:769–782, 1948.

Medical Research Council. Treatment of pulmonary tuberculosis with streptomycin and para-aminosalicylic acid. *Br Med J*, 2:1073–1085, 1950.

X. L. Meng. The EM algorithm and medical studies: a historical link. *Stat Methods Med Res*, 6:3–23, 1997.

D. Menzies. Notes from the field–TB control 'sound bites'. *Int J Tuberc Lung Dis*, 1:488–489, 1997.

G. Middlebrook, Z. Reggiardo, and W. D. Tigertt. Automatable radiometric detection of growth of Mycobacterium tuberculosis in selective media. *Am Rev Respir Dis*, 115:1066–1069, 1977.

R. Mistry, J. M. Cliff, C. L. Clayton, N. Beyers, Y. S. Mohamed, P. A. Wilson, H. M. Dockrell, D. M. Wallace, P. D. van Helden, K. Duncan, and P. T. Lukey. Gene-expression patterns in whole blood identify subjects at risk for recurrent tuberculosis. *J Infect Dis*, 195:357–365, 2007.

D. Mitchison and W. Sturm. The measurement of early bactericidal activity. In A. Malin and K. McAdam, editors, *Bailliere's Clinical Infectious Diseases: Mycobacterial Diseases Part II*, pages 185–206. Bailliere Tindall, London, 1997.

D. A. Mitchison. Treatment of tuberculosis. The Mitchell lecture 1979. *J R Coll Physicians Lond*, 14:91–99, 1980.

D. A. Mitchison. Mechanisms of the action of drugs in the short-course chemotherapy. *Bull Int Union Tuberc*, 60:36–40, 1985.

D. A. Mitchison. Infectivity of patients with pulmonary tuberculosis during chemotherapy. *Eur Respir J*, 3:385–386, 1990.

D. A. Mitchison. Understanding the chemotherapy of tuberculosis - current problems. *J Antimicrob Chemother*, 29:477–493, 1992.

D. A. Mitchison. Assessment of new sterilizing drugs for treating pulmonary tuberculosis by culture at 2 months. *Am Rev Respir Dis*, 147:1062–3, 1993.

D. A. Mitchison. Modern methods for assessing the drugs used in the chemotherapy of mycobacterial disease. *Soc Appl Bacteriol Symp Ser*, 25: 72S–80S, 1996.

D. A. Mitchison. Mechanisms of tuberculosis chemotherapy. *J Pharm Pharmacol*, 49:31–36, 1997.

D. A. Mitchison. Role of individual drugs in the chemotherapy of tuberculosis. *Int J Tuberc Lung Dis*, 4:796–806, 2000.

D. A. Mitchison. A reiterative method for calculating bactericidal activity. *Am J Respir Crit Care Med*, 167:663, 2003.

D. A. Mitchison. The diagnosis and therapy of tuberculosis during the past 100 years. *Am J Respir Crit Care Med*, 171:699–706, 2005.

D. A. Mitchison. Clinical development of anti-tuberculosis drugs. *J Antimicrob Chemother*, 58:494–495, 2006.

D. A. Mitchison and J. M. Dickinson. Short term chemotherapy of tuberculosis. Bactericidal mechanisms in short term chemotherapy. *Bull Int Union Tuberc*, 53:270–275, 1978.

D. A. Mitchison and A. J. Nunn. Influence of initial-drug resistance on the response to short-course chemotherapy of pulmonary tuberculosis. *Am Rev Respir Dis*, 133:423–430, 1986.

D. A. Mitchison, A. B. Keyes, E. A. Edwards, P. Ayuma, S. P. Byfield, and A. J. Nunn. Quality-control in tuberculosis bacteriology 2. the origin of isolated positive cultures from the sputum of patients in 4 studies of short course chemotherapy in Africa. *Tubercle*, 61:135–144, 1980.

D. A. Mitchison, A. Jindani, G. R. Davies, and F. Sirgel. Isoniazid activity is terminated by bacterial persistence. *J Infect Dis*, 195:1871–1872, 2007.

R. S. Mitchison, B. W. Allen, and D. A. Mitchison. Letter: False-positive acid-fast smears. *Lancet*, 2:281, 1975.

C. D. Mitnick, K. G. Castro, M. Harrington, L. V. Sacks, and W. Burman. Randomized trials to optimize treatment of multidrug-resistant tuberculosis. *PLoS Med*, 4, 2007.

G. Molenberghs, H. Geys, and M. Buyse. Evaluation of surrogate endpoints in randomized experiments with mixed discrete and continuous outcomes. *Stat Med*, 20:3023–38, 2001.

G. Molenberghs, M. Buyse, H. Geys, D. Renard, T. Burzykowski, and A. Alonso. Statistical challenges in the evaluation of surrogate endpoints in randomized trials. *Control Clin Trials*, 23:607–25, 2002.

G. Molenberghs, T. Burzykowski, A. Alonso, and M. Buyse. A perspective on surrogate endpoints in controlled clinical trials. *Stat Methods Med Res*, 13: 177–206, 2004.

G. Molenberghs, T. Burzykowski, A. Alonso, P. Assam, A. Tilahun, and M. Buyse. The meta-analytic framework for the evaluation of surrogate endpoints in clinical trials. *J Stat Plan Inference*, 138:432–449, 2008.

Molenberghs, Geert and Lesaffre, Emmanuel. Marginal modeling of correlated ordinal data using a multivariate plackett distribution. *J Am Stat Assoc*, 89:633–644, 1994.

P. Moleur and J. P. Boissel. Definition of a surrogate end-point. *Control Clin Trials*, 8:304–304, 1987.

T. Moll. TB on the back burner, losing curable status. *Int J Tuberc Lung Dis*, 11: 355, 2007.

D. A. Moore, C. A. Evans, R. H. Gilman, L. Caviedes, J. Coronel, A. Vivar, E. Sanchez, Y. Pinedo, J. C. Saravia, C. Salazar, R. Oberhelman, M.-G. Hollm-Delgado, D. LaChira, A. R. Escombe, and J. S. Friedland. Microscopic-Observation Drug-Susceptibility Assay for the Diagnosis of TB. *N Engl J Med*, 355:1539–1550, 2006.

J. Moore-Gillon. Multidrug-resistant tuberculosis: this is the cost. *Ann N Y Acad Sci*, 953:233–240, 2001.

D. M. Morens. At the deathbed of consumptive art. *Emerg Infect Dis*, 8:1353–1358, 2002.

B. J. T. Morgan. *Analysis of Quantal Response Data*. Chapman & Hall, London, 1992.

C. A. Morris and B. W. Barton. Is guinea pig inoculation ever justified for the diagnosis of tuberculosis? *J Clin Pathol*, 36:719–720, 1983.

A. M. Morsy, H. H. Zaher, M. H. Hassan, and A. Shouman. Predictors of treatment failure among tuberculosis patients under DOTS strategy in Egypt. *East Mediterr Health J*, 9:689–701, 2003.

S. A. Munro, S. A. Lewin, H. J. Smith, M. E. Engel, A. Fretheim, and J. Volmink. Patient adherence to tuberculosis treatment: a systematic review of qualitative research. *PLoS Med*, 4, 2007.

H. Murad and L. S. Freedman. Estimating and testing interactions in linear regression models when explanatory variables are subject to classical measurement error. *Stat Med*, 26:4293–4310, 2007.

J. F. Murray. The white plague: down and out, or up and coming? J. Burns Amberson lecture. *Am Rev Respir Dis*, 140:1788–1795, 1989.

J. F. Murray. A century of tuberculosis. *Am J Respir Crit Care Med*, 169:1181–1186, 2004.

P. Mwaba, M. Maboshe, C. Chintu, B. Squire, S. Nyirenda, R. Sunkutu, and A. Zumla. The relentless spread of tuberculosis in zambia-trends over the past 37 years (1964-2000). *S Afr Med J*, 93:149–152, 2003.

R. E. Nettles, D. Mazo, K. Alwood, R. Gachuhi, G. Maltas, K. Wendel, W. Cronin, N. Hooper, W. Bishai, and T. R. Sterling. Risk factors for relapse and acquired rifamycin resistance after directly observed tuberculosis treatment: a comparison by HIV serostatus and rifamycin use. *Clin Infect Dis*, 38:731–6, 2004.

E. Nuermberger, I. Rosenthal, S. Tyagi, K. N. Williams, D. Almeida, C. A. Peloquin, W. R. Bishai, and J. H. Grosset. Combination chemotherapy with the nitroimidazopyran PA-824 and first-line drugs in a murine model of tuberculosis. *Antimicrob Agents Chemother (Bethesda)*, 50:2621–2625, 2006.

A. J. Nunn, P. P. Phillips, and S. H. Gillespie. Design issues in pivotal drug trials for drug sensitive tuberculosis (TB). *Tuberculosis*, 88:S85–S92, 2008.

P. Nunn, J. Porter, W. Githui, and J. Odhiambo. Treating tuberculosis in HIV-positive Africans. *Lancet*, 338:1140–1141, 1991.

R. J. O'Brien. Studies of the early bactericidal activity of new drugs for tuberculosis - a help or a hindrance to antituberculosis drug development? *Am J Respir Crit Care Med*, 166:3–4, 2002.

R. J. O'Brien and P. P. Nunn. The need for new drugs against tuberculosis. Obstacles, opportunities, and next steps. *Am J Respir Crit Care Med*, 163: 1055–1058, 2001.

W. A. O'Brien, P. M. Hartigan, D. Martin, J. Esinhart, A. Hill, S. Benoit, M. Rubin, M. S. Simberkoff, and J. D. Hamilton. Changes in plasma HIV-1 RNA and CD4+ lymphocyte counts and the risk of progression to AIDS. Veterans Affairs Cooperative Study Group on AIDS. *N Engl J Med*, 334:426–431, 1996.

P. C. Onyebujoh, J. B. Levin, F. B. Fourie, V. Garhiram, L. C. Tembe, N. P. Phili, T. C. P. Mthiyane, T. Moniwa, G. Bayer, I. M. Ramajoe, T. A. B. Mncwabe, L. G. M. Mallisar, T. N. M. Saul, J. B. Levin, T. H. F. G. Jackson, S. Suparsad, P. E. M. Fine, D. J. S. Pendlebury, E. Fine, I. Houghton, J. Clyde, H. P. Vos, N. Padayatchi, A. Pala, A. Ramjee, M. Ramjee, J. Ramdeen, I. H. Masters, G. Osbourne, K. Naidu, S. Bamba, B. Mazur, R. Czarnocki, K. Landers, G. Ndlovu, N. Maphumulo, V. Garhiram, A. W. Sturm, J. Moodley, C. Pillay, L. Roux, R. Moodley, A. Sarawan, T. Jali, F. Manickam, A. Smith, Gopaul, T. Durosanmi, R. Moonsammy, P. Wyld, J. McCallum, C. Fulton, K. Bisset, S. Henderson, D. Stewart, O. Nticinka, D. Watson, N. Tuckwell, D. Kennard, J. L. Stanford, G. A. Rook, J. M. Grange, A. A. Zumla, E. Bateman, P. Hopwell, J. Darbyshire, L. Geiter, A. Nunn, and K. Weyer. Immunotherapy with Mycobacterium vaccae in patients with newly diagnosed pulmonary tuberculosis: a randomised controlled trial. *Lancet*, 354:116–119, 1999.

J. O'Quigley and P. Flandre. Quantification of the prentice criteria for surrogate endpoints. *Biometrics*, 62:297–300, 2006.

J. O'Quigley, R. Xu, and J. Stare. Explained randomness in proportional hazards models. *Stat Med*, 24:479–489, 2005.

A. M. C. Pachas, R. Blank, M. C. S. Fawzi, J. Bayona, M. C. Becerra, and C. D. Mitnick. Identifying early treatment failure on category i therapy for pulmonary tuberculosis in lima ciudad, peru. *Int J Tuberc Lung Dis*, 8:52–58, 2004.

M. Palaci, R. Dietze, D. J. Hadad, F. K. C. Ribeiro, R. L. Peres, S. A. Vinhas, E. L. N. Maciel, V. do Valle Dettoni, L. Horter, W. H. Boom, J. L. Johnson, and K. D. Eisenach. Cavitary Disease and Quantitative Sputum Bacillary Load in Cases of Pulmonary Tuberculosis. *J Clin Microbiol*, 45:4064–4066, 2007.

J.-C. Palomino, A. Martin, and F. Portaels. MODS assay for the diagnosis of TB. *N Engl J Med*, 356:188–189, 2007.

T. Park and S.-Y. Lee. A test of missing completely at random for longitudinal data with missing observations. *Stat Med*, 16:1859–1871, 1997.

C. Parry and P. D. Davies. The resurgence of tuberculosis. *Soc Appl Bacteriol Symp Ser*, 25:23S–26S, 1996.

C. Peloquin. What is the 'right' dose of rifampin? *Int J Tuberc Lung Dis*, 7:3–5, 2003.

M. S. Pepe. *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press, Oxford, 2003.

M. S. Pepe and G. L. Anderson. A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Commun Stat Simul Comp*, 23:939–951, 1994.

M. D. Perkins, G. Roscigno, and A. Zumla. Progress towards improved tuberculosis diagnostics for developing countries. *Lancet*, 367:942–943, 2006.

F. M. R. Perrin, M. C. I. Lipman, T. D. McHugh, and S. H. Gillespie. Biomarkers of treatment response in clinical trials of novel antituberculosis agents. *Lancet Infect Dis*, 7:481–490, 2007.

A. N. Pettitt. Censored observations, repeated measures and mixed effects models an approach using the EM algorithm and normal errors. *Biometrika*, 73:635–643, 1986.

C. Pheiffer, N. Carroll, N. Beyers, P. Donald, K. Duncan, P. Uys, and P. van Helden. Time to detection of Mycobacterium tuberculosis in BACTEC systems as a viable alternative to colony counting. *Int J Tuberc Lung Dis*, 12: 792–798, 2008.

P. Phillips and K. Fielding. The evaluation of culture conversion during treatment for tuberculosis as a surrogate for treatment failure. *Int J Tuberc Lung Dis*, 11:S161–S162, 2007.

A. E. Pitchenik and H. A. Rubinson. The radiographic appearance of tuberculosis in patients with the acquired immune deficiency syndrome (AIDS) and pre-AIDS. *Am Rev Respir Dis*, 131:393–396, 1985.

M. W. R. Pletz, A. D. Roux, A. Roth, K.-H. Neumann, H. Mauch, and H. Lode. Early bactericidal activity of moxifloxacin in treatment of pulmonary tuberculosis: a prospective, randomized study. *Antimicrob Agents Chemother (Bethesda)*, 48:780–782, 2004.

S. J. Pocock. *Clinical Trials: A Practical Approach*. Wiley & Sons, Chicester, England, 1983.

R. L. Prentice. Surrogate endpoints in clinical trials: definition and operational criteria. *Stat Med*, 8:431–40, 1989.

R. L. Prentice. Comments on the paper by Begg and Leung. *J R Stat Soc Ser A Stat Soc*, 163:24–28, 2000.

R. L. Prentice. Discussion: Surrogate endpoint definition and evaluation. In G. Molenberghs, M. E. Buyse, and T. Burzykowski, editors, *The Evaluation of Surrogate Endpoints*, pages 341–348. Springer, New York, 2005.

R. L. Prentice and L. A. Mancl. Regression models for discrete longitudinal responses : Comment. *Stat Sci*, 8:302–304, 1993.

Prentice, R. L. and Gloeckler, L. A. Regression analysis of grouped survival data with application to breast cancer data. *Biometrics*, 34:57–67, 1978.

A. Pryseley, A. Tilahun, A. Alonso, and G. Molenberghs. Information-theory based surrogate marker evaluation from several randomized clinical trials with continuous true and binary surrogate endpoints. *Clin Trials*, 4:587–597, 2007.

Y. Qu and M. Case. Quantifying the indirect treatment effect via surrogate markers. *Stat Med*, 25:223–231, 2006.

Y. Qu and M. Case. Quantifying the effect of the surrogate marker by information gain. *Biometrics*, 63:958–960, 2007.

S. Rabe-Hesketh, A. Skrondal, and A. Pickles. Reliable estimation of generalized linear mixed models using adaptive quadrature. *Stata J*, 2:1–21, 2002.

H. Ramarokoto, H. Randriamiharisoa, A. Rakotoarisaonina, T. Rasolovavalona, V. Rasolofo, S. Chanteau, M. Ralamboson, B. Cauchoix, and D. Rakotondramarina. Bacteriological follow-up of tuberculosis treatment: a comparative study of smear microscopy and culture results at the second month of treatment. *Int J Tuberc Lung Dis*, 6:909–912, 2002.

M. P. Ravenel. The warfare against tuberculosis. *Proc Am Philos Soc*, 42:212–219, 1903.

M. Raviglione. XDR-TB: entering the post-antibiotic era? *Int J Tuberc Lung Dis*, 10:1185–1187, 2006.

M. C. Raviglione and A. Pio. Evolution of WHO policies for tuberculosis control, 1948-2001. *Lancet*, 359:775–780, 2002.

M. C. Raviglione, J. P. Narain, and A. Kochi. HIV-associated tuberculosis in developing countries: clinical features, diagnosis, and treatment. *Bull World Health Organ*, 70:515–526, 1992.

L. B. Reichman. Whither Mycobacterium vaccae? *Lancet*, 354:90, 1999.

D. Renard, H. Geys, G. Molenberghs, T. Burzykowski, and M. Buyse. Validation of surrogate endpoints in multiple randomized clinical trials with discrete outcomes. *Biometrical J*, 44:921–935, 2002.

H. L. Rieder. Sputum smear conversion during directly observed treatment for tuberculosis. *Tuber Lung Dis*, 77:124–129, 1996.

H. L. Rieder. Fourth-generation fluoroquinolones in tuberculosis. *The Lancet*, 373:1148–1149, 2009.

B. L. Riggs, S. F. Hodgson, W. M. O'Fallon, E. Y. Chao, H. W. Wahner, J. M. Muhs, S. L. Cedel, and L. J. Melton. Effect of fluoride treatment on the fracture rate in postmenopausal women with osteoporosis. *N Engl J Med*, 322:802–809, 1990.

R. D. Riley, K. R. Abrams, A. J. Sutton, P. C. Lambert, and J. R. Thompson. Bivariate random-effects meta-analysis and the estimation of between-study correlation. *BMC Med Res Methodol*, 7, 2007.

R. D. Riley, J. R. Thompson, and K. R. Abrams. An alternative model for bivariate random-effects meta-analysis when the within-study correlations are unknown. *Biostatistics*, 9:172–186, 2008.

S. R. Ritchie, A. C. Harrison, R. H. Vaughan, L. Calder, and A. J. Morris. New recommendations for duration of respiratory isolation based on time to detect Mycobacterium tuberculosis in liquid culture. *Eur Respir J*, 30: 501–507, 2007.

E. C. Rivers and R. L. Mancera. New anti-tuberculosis drugs with novel mechanisms of action. *Curr Med Chem*, 15:1956–1967, 2008.

G. A. W. Rook and R. Hernandez-Pando. The pathogenesis of tuberculosis. *Annu Rev Microbiol*, 50:259–284, 1996.

I. M. Rosenthal, M. Zhang, K. N. Williams, C. A. Peloquin, S. Tyagi, A. A. Vernon, W. R. Bishai, R. E. Chaisson, J. H. Grosset, and E. L. Nuermberger. Daily dosing of rifapentine cures tuberculosis in three months or less in the murine model. *PLoS Med*, 4, 2007.

B. M. Rothschild, L. D. Martin, G. Lev, H. Bercovier, G. K. Bar-Gal, C. Greenblatt, H. Donoghue, M. Spigelman, and D. Brittain. Mycobacterium tuberculosis complex DNA from an extinct bison dated 17,000 years before the present. *Clin Infect Dis*, 33:305–311, 2001.

E. Roy, D. B. Lowrie, and S. R. Jolles. Current strategies in TB immunotherapy. *Curr Mol Med*, 7:373–386, 2007.

P. Royston. The lognormal distribution as a model for survival time in cancer, with an emphasis on prognostic factors. *Stat Neerl*, 55:89–104, 2001.

P. Royston. Explained variation for survival models. *Stata J*, 6:83–96, 2006.

P. Royston and W. Sauerbrei. A new measure of prognostic separation in survival data. *Stat Med*, 23:723–748, 2004.

P. Royston, M. K. B. Parmar, and W. Qian. Novel designs for multi-arm clinical trials with survival outcomes with an application in ovarian cancer. *Stat Med*, 22:2239–2256, 2003.

P. Royston, D. G. Altman, and W. Sauerbrei. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med*, 25:127–141, 2006.

J. N. Ruskin. The cardiac arrhythmia suppression trial (CAST). *N Engl J Med*, 321:386–388, 1989.

R. Rustomjee, A. H. Diacon, J. Allen, A. Venter, C. Reddy, R. F. Patientia, T. C. P. Mthiyane, T. D. Marez, R. van Heeswijk, R. Kerstens, A. Koul, K. D. Beule, P. R. Donald, and D. F. McNeeley. Early bactericidal activity and pharmacokinetics of the diarylquinoline TMC207 in treatment of pulmonary tuberculosis. *Antimicrob Agents Chemother (Bethesda)*, 52:2831–2835, 2008a.

R. Rustomjee, C. Lienhardt, T. Kanyok, G. Davies, J. Levin, T. Mthiyane, C. Reddy, A. Sturm, F. Sirgel, J. Allen, D. Coleman, B. Fourie, D. Mitchison, and Gatifloxacin for TB OFLOTUB study team. A phase ii study of the sterilising activities of ofloxacin, gatifloxacin and moxifloxacin in pulmonary tuberculosis. *Int J Tuberc Lung Dis*, 12:128–138, 2008b.

F. M. Salaniponi, J. J. Christensen, F. Gausi, J. J. Kwanjana, and A. D. Harries. Sputum smear status at two months and subsequent treatment outcome in new patients with smear-positive pulmonary tuberculosis. *Int J Tuberc Lung Dis*, 3:1047–1048, 1999.

N. Salomon, D. C. Perlman, P. Friedmann, S. Buchstein, B. N. Kreiswirth, and D. Mildvan. Predictors and outcome of multidrug-resistant tuberculosis. *Clin Infect Dis*, 21:1245–1252, 1995.

T. Santha, R. Garg, T. R. Frieden, V. Chandrasekaran, R. Subramani, P. G. Gopi, N. Selvakumar, S. Ganapathy, N. Charles, J. Rajamma, and P. R. Narayanan. Risk factors associated with default, failure and death among tuberculosis patients treated in a DOTS programme in Tiruvallur District, South India, 2000. *Int J Tuberc Lung Dis*, 6:780–788, 2002.

T. Santha, F. Rehman, D. A. Mitchison, G. R. Sarma, A. M. Reetha, R. Prabhaker, and I. C. o. M. R. Tuberculosis Research Centre. Split-drug regimens for the treatment of patients with sputum smear-positive pulmonary tuberculosis-a unique approach. *Trop Med Int Health*, 9:551–558, 2004.

S. Sarkar and Y. Qu. Quantifying the treatment effect explained by markers in the presence of measurement error. *Stat Med*, 26:1955–1963, 2007.

M. Schemper and R. Henderson. Predictive accuracy and explained variation in Cox regression. *Biometrics*, 56:249–255, 2000.

M. Schemper and J. Stare. Explained variation in survival analysis. *Stat Med*, 15:1999–2012, 1996.

N. W. Schluger and W. N. Rom. Current approaches to the diagnosis of active pulmonary tuberculosis. *Am J Respir Crit Care Med*, 149:264–267, 1994.

H. J. Schünemann, A. D. Oxman, J. Brozek, P. Glasziou, R. Jaeschke, G. E. Vist, J. W. Williams, R. Kunz, J. Craig, V. M. Montori, P. Bossuyt, G. H. Guyatt, and G. R. A. D. E. W. Group. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ*, 336:1106–1110, 2008.

J. B. Selkon, S. Devadatta, K. G. Kulkarni, D. A. Mitchison, A. S. Narayana, C. N. Nair, and K. Ramachandran. The emergence of isoniazid-resistant cultures in patients with pulmonary tuberculosis during treatment with isoniazid alone or isoniazid plus PAS. *Bull World Health Organ*, 31:273–294, 1964.

S. E. Sharp, M. Lemes, S. G. Sierra, A. Poniecka, and R. J. Poppiti. Löwenstein-Jensen media. No longer necessary for mycobacterial isolation. *Am J Clin Pathol*, 113:770–773, 2000.

D. Shingadia and V. Novelli. Diagnosis and treatment of tuberculosis in children. *Lancet Infect Dis*, 3:624–632, 2003.

Singapore Tuberculosis Service/British Medical Research Council. Controlled trial of intermittent regimens of rifampin plus isoniazid for pulmonary tuberculosis in Singapore. The results up to 30 months. *Am Rev Respir Dis*, 116:807–820, 1977.

Singapore Tuberculosis Service/British Medical Research Council. Clinical trial of three 6-month regimens of chemotherapy given intermittently in the continuation phase in the treatment of pulmonary tuberculosis. Singapore Tuberculosis Service/British Medical Research Council. *Am Rev Respir Dis*, 132:374–378, 1985.

Singapore Tuberculosis Service/British Medical Research Council. Long-term follow-up of a clinical trial of six-month and four-month regimens of chemotherapy in the treatment of pulmonary tuberculosis. *Am Rev Respir Dis*, 133:779–783, 1986.

Singapore Tuberculosis Service/British Medical Research Council. Five-year follow-up of a clinical trial of three 6-month regimens of chemotherapy given intermittently in the continuation phase in the treatment of pulmonary tuberculosis. *Am Rev Respir Dis*, 137:1147–1150, 1988.

Singapore Tuberculosis Service/British Medical Research Council. Assessment of a daily combined preparation of isoniazid, rifampin, and pyrazinamide in a controlled trial of three 6-month regimens for smear-positive pulmonary tuberculosis. *Am Rev Respir Dis*, 143:707–712, 1991.

Singapore Tuberculosis Services/Brompton Hospital/British Medical Research Council. A controlled clinical trial of the role of thiacetazone-containing regimens in the treatment of pulmonary tuberculosis in Singapore. Singapore Tuberculosis Services-Brompton Hospital-British Medical Research Council Investigation. *Tubercle*, 52:88–116, 1971.

302

Singapore Tuberculosis Services/Brompton Hospital/British Medical Research Council. A controlled clinical trial of the role of thiacetazone-containing regimens in the treatment of pulmonary tuberculosis in Singapore: second report. *Tubercle*, 55:251–260, 1974.

R. Singla, N. Al-Sharif, M. O. Al-Sayegh, M. M. Osman, and M. A. Shaikh. Influence of anti-tuberculosis drug resistance on the treatment outcome of pulmonary tuberculosis patients receiving DOTS in Riyadh, Saudi Arabia. *Int J Tuberc Lung Dis*, 6:585–91, 2002.

R. Singla, M. M. Osman, N. Khan, N. Al-Sharif, M. O. Al-Sayegh, and M. A. Shaikh. Factors predicting persistent sputum smear positivity among pulmonary tuberculosis patients 2 months after treatment. *Int J Tuberc Lung Dis*, 7:58–64, 2003.

F. Sirgel, A. Venter, and D. Mitchison. Sources of variation in studies of the early bactericidal activity of antituberculosis drugs. *J Antimicrob Chemother*, 47:177–182, 2001.

F. A. Sirgel, F. J. H. Botha, D. P. Parkin, B. W. Vandewal, P. R. Donald, P. K. Clark, and D. A. Mitchison. The early bactericidal activity of rifabutin in patients with pulmonary tuberculosis measured by sputum viable counts - a new method of drug assessment. *J Antimicrob Chemother*, 32:867–875, 1993.

F. A. Sirgel, F. J. Botha, D. P. Parkin, B. W. vandeWal, R. Schall, P. R. Donald, and D. A. Mitchison. The early bactericidal activity of ciprofloxacin in patients with pulmonary tuberculosis. *Am J Respir Crit Care Med*, 156:901–905, 1997.

F. A. Sirgel, P. R. Donald, J. Odhiambo, W. Githui, K. C. Umapathy, C. N. Paramasivan, C. M. Tam, K. M. Kam, C. W. Lam, K. M. Sole, and D. A. Mitchison. A multicentre study of the early bactericidal activity of anti-tuberculosis drugs. *J Antimicrob Chemother*, 45:859–870, 2000.

F. A. Sirgel, P. B. Fourie, P. R. Donald, N. Padayatchi, R. Rustomjee, J. Levin, G. Roscigno, J. Norman, H. McIlleron, and D. A. Mitchison. The early bactericidal activities of rifampin and rifapentine in pulmonary tuberculosis. *Am J Respir Crit Care Med*, 172:128–135, 2005.

K. Slama, C.-Y. Chiang, D. Enarson, K. Hassmiller, A. Fanning, P. Gupta, and C. Ray. Tobacco and tuberculosis: a qualitative systematic review and meta-analysis. *Int J Tuberc Lung Dis*, 11:1049–1061, 2007.

P. Sonnenberg, J. Murray, S. Shearer, J. R. Glynn, B. Kambashi, and P. Godfrey-Faussett. Tuberculosis treatment failure and drug resistance-same strain or reinfection? *Trans R Soc Trop Med Hyg*, 94:603–607, 2000.

M. Spigelman and S. Gillespie. Tuberculosis drug development pipeline: progress and hope. *Lancet*, 367:945–947, 2006.

M. K. Spigelman. New tuberculosis therapeutics: a growing pipeline. *J Infect Dis*, 196:S28–S34, 2007.

R. Sposto. Cure model analysis in cancer: an application to data from the Children's Cancer Group. *Stat Med*, 21:293–312, 2002.

W. Stadler. Fuzzy thinking on biomarkers. *Urol Oncol*, 25:97–100, 2007.

J. L. Stanford, C. A. Stanford, J. M. Grange, N. N. Lan, and A. Etemadi. Does immunotherapy with heat-killed Mycobacterium vaccae, offer hope for the treatment of multi-drug-resistant pulmonary tuberculosis? *Respir Med*, 95: 444–447, 2001.

Stop TB Partnership. The Global Plan to Stop TB, 2006-2015, 2006.

M. Susser. *Causal Thinking in the Health Sciences*. Oxford University Press, New York, 1973.

M. C. Sutter. Assigning causation in disease: beyond koch's postulates. *Perspect Biol Med*, 39:581–92, 1996.

C. M. Tam, S. L. Chan, K. M. Kam, E. Sim, D. Staples, K. M. Sole, H. Al-Ghusein, and D. A. Mitchison. Rifapentine and isoniazid in the continuation phase of a 6-month regimen. interim report: no activity of isoniazid in the continuation phase. *Int J Tuberc Lung Dis*, 4:262–267, 2000.

C. M. Tam, S. L. Chan, K. M. Kam, R. L. Goodall, and D. A. Mitchison. Rifapentine and isoniazid in the continuation phase of a 6-month regimen. Final report at 5 years: prognostic value of various measures. *Int J Tuberc Lung Dis*, 6:3–10, 2002.

Tanzania/British Medical Research Council. Controlled clinical trial of two 6-month regimens of chemotherapy in the treatment of pulmonary tuberculosis. Tanzania/British Medical Research Council Study. *Am Rev Respir Dis*, 131:727–731, 1985.

Tanzania/British Medical Research Council. A controlled trial of a 4-weekly supplement of rifampicin, pyrazinamide and streptomycin in the continuation phase of a 7-month daily chemotherapy regimen for pulmonary tuberculosis. Tanzania/British Medical Research Council Collaborative Investigation. *S. Afr. Med. J.*, 86:960–965, 1996.

J. C. Tardif, T. Heinonen, D. Orloff, and P. Libby. Vascular biomarkers and surrogates in cardiovascular disease. *Circulation*, 113:2936–2942, 2006.

J. M. Taylor and Y. Wang. Surrogate markers and joint models for longitudinal and survival data. *Control Clin Trials*, 23:626–34, 2002.

J. M. G. Taylor, Y. Wang, and R. Thiebaut. Counterfactual links to the proportion of treatment effect explained by a surrogate marker. *Biometrics*, 61: 1102–1111, 2005.

E. E. Telzak, B. A. Fazal, C. L. Pollard, G. S. Turett, J. E. Justman, and S. Blum. Factors influencing time to sputum conversion among patients with smear-positive pulmonary tuberculosis. *Clin Infect Dis*, 25:666–670, 1997.

E. E. Telzak, B. A. Fazal, G. S. Turett, J. E. Justman, and S. Blum. Factors influencing time to sputum conversion among patients with smear-positive pulmonary tuberculosis. *Clin Infect Dis*, 26:775–776, 1998.

R. Temple. A regulatory authority's opinion about surrogate endpoints. In G. Nimmo, W.S.; Tucker, editor, *Clinical Measurement in Drug Evaluation*, pages 3–22. Wiley, New York, 1995.

R. Temple. Are surrogate markers adequate to assess cardiovascular disease drugs? *JAMA*, 282:790–795, 1999.

The Cardiac Arrhythmia Suppression Trial (CAST) Investigators. Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. *N Engl J Med*, 321: 406–412, 1989.

The Working Alliance for TB Drug Development. Working Alliance for TB Drug Development, Cape Town, South Africa, February 8th, 2000 Declaration. *Int J Tuberc Lung Dis*, 4:489–490, 2000.

A. Thomas, P. G. Gopi, T. Santha, V. Chandrasekaran, R. Subramani, N. Selvakumar, S. I. Eusuff, K. Sadacharam, and P. R. Narayanan. Predictors of relapse among pulmonary tuberculosis patients treated in a DOTS programme in South India. *Int J Tuberc Lung Dis*, 9:556–561, 2005.

V. O. Thomsen, A. Kok-Jensen, M. Buser, S. Philippi-Schulz, and H. J. Burkardt. Monitoring treatment of patients with pulmonary tuberculosis: can PCR be applied? *J Clin Microbiol*, 37:3601–3607, 1999.

F. Tibaldi, J. C. Abrahantes, G. Molenberghs, D. Renard, T. Burzykowski, M. Buyse, M. Parmar, T. Stijnen, and R. Wolfinger. Simplified hierarchical linear models for the evaluation of surrogate endpoints. *J Stat Comput Sim*, 73:643–658, 2003.

A. Tilahun, A. Pryseley, A. Alonso, and G. Molenberghs. Flexible surrogate marker evaluation from several randomized clinical trials with continuous endpoints, using R and SAS. *Comput Stat Data An*, 51:4152–4163, 2007.

A. Trébucq and H. L. Rieder. Two excellent management tools for national tuberculosis programmes: history of prior treatment and sputum status at two months. *Int J Tuberc Lung Dis*, 2:184–186, 1998.

P. Trouiller, P. Olliaro, E. Torreele, J. Orbinski, R. Laing, and N. Ford. Drug development for neglected diseases: a deficient market and a public-health policy failure. *Lancet*, 359:2188–2194, 2002.

305

S.-W. Um, S. W. Lee, S. Y. Kwon, H. I. Yoon, K. U. Park, J. Song, C.-T. Lee, and J.-H. Lee. Low serum concentrations of anti-tuberculosis drugs and determinants of their serum levels. *Int J Tuberc Lung Dis*, 11:972–978, 2007.

R. van Crevel, B. Alisjahbana, W. C. M. de Lange, F. Borst, H. Danusantoso, J. W. M. van der Meer, D. Burger, and R. H. H. Nelwan. Low plasma concentrations of rifampicin in tuberculosis patients in Indonesia. *Int J Tuberc Lung Dis*, 6:497–502, 2002.

H. C. van Houwelingen. The evaluation of surrogate endpoints. *Biometrics*, 62:948–949, 2006.

H. C. van Houwelingen, L. R. Arends, and T. Stijnen. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Stat Med*, 21: 589–624, 2002.

A. van Rie, R. Warren, M. Richardson, T. C. Victor, R. P. Gie, D. A. Enarson, N. Beyers, and P. D. van Helden. Exogenous reinfection as a cause of recurrent tuberculosis after curative treatment. *N Engl J Med*, 341:1174–1179, 1999.

A. van Rie, T. C. Victor, M. Richardson, R. Johnson, G. D. van der Spuy, E. J. Murray, N. Beyers, N. C. G. van Pittius, P. D. van Helden, and R. M. Warren. Reinfection and mixed infection cause changing Mycobacterium tuberculosis drug-resistance patterns. *Am J Respir Crit Care Med*, 172:636–642, 2005.

B. Verbeke and G. Molenberghs. *Linear Mixed Models for Longitudinal Data*. Springer Series in Statistics. Springer-Verlag, New York, 2000.

A. Vernon, W. Burman, D. Benator, A. Khan, and L. Bozeman. Acquired rifamycin monoresistance in patients with HIV-related tuberculosis treated with once-weekly rifapentine and isoniazid. tuberculosis trials consortium. *Lancet*, 353:1843–1847, 1999.

A. A. Vernon and M. F. Iademarco. In the treatment of tuberculosis, you get what you pay for... *Am J Respir Crit Care Med*, 170:1040–1042, 2004.

S. Verver, R. M. Warren, N. Beyers, M. Richardson, G. D. van der Spuy, M. W. Borgdorff, D. A. Enarson, M. A. Behr, and P. D. van Helden. Rate of reinfection tuberculosis after successful treatment is higher than rate of new tuberculosis. *Am J Respir Crit Care Med*, 171:1430–1435, 2005.

E. Villamor, F. Mugusi, W. Urassa, R. J. Bosch, E. Saathoff, K. Matsumoto, S. N. Meydani, and W. W. Fawzi. A trial of the effect of micronutrient supplementation on treatment outcome, t cell counts, morbidity, and mortality in adults with pulmonary tuberculosis. *J Infect Dis*, 197:1499–1505, 2008.

J. Volmink and P. Garner. Directly observed therapy for treating tuberculosis. *Cochrane DB Syst Rev*, 2, 2006.

E. F. Vonesh, T. Greene, and M. D. Schluchter. Shared parameter models for the joint analysis of longitudinal data and event times. *Stat Med*, 25:143–163, 2006.

J. A. Wagner, S. A. Williams, and C. J. Webster. Biomarkers and surrogate end points for fit-for-purpose development and regulatory evaluation of new drugs. *Clin Pharmacol Ther*, 81:104–107, 2007.

R. S. Wallis. Surrogate markers to assess new therapies for drug-resistant tuberculosis. *Expert Rev Anti Infect Ther*, 5:163–168, 2007.

R. S. Wallis and J. L. Johnson. The role of surrogate markers in the clinical evaluation of antituberculous chemotherapy. *Anti-Infect Ag Med Chem*, 4: 287–294, 2005.

R. S. Wallis, M. Perkins, M. Phillips, M. Joloba, B. Demchuk, A. Namale, J. L. Johnson, D. Williams, K. Wolski, L. Teixeira, R. Dietze, R. D. Mugerwa, K. Eisenach, and J. J. Ellner. Induction of the antigen 85 complex of My- cobacterium tuberculosis in sputum: a determinant of outcome in pul- monary tuberculosis treatment. *J Infect Dis*, 178:1115–1121, 1998.

R. S. Wallis, S. Patil, S. H. Cheon, K. Edmonds, M. Phillips, M. D. Perkins, M. Joloba, A. Namale, J. L. Johnson, L. Teixeira, R. Dietze, S. Siddiqi, R. D. Mugerwa, K. Eisenach, and J. J. Ellner. Drug tolerance in Mycobacterium tuberculosis. *Antimicrob Agents Chemother (Bethesda)*, 43:2600–6, 1999.

R. S. Wallis, M. D. Perkins, M. Phillips, M. Joloba, A. Namale, J. L. Johnson, C. C. Whalen, L. Teixeira, B. Demchuk, R. Dietze, R. D. Mugerwa, K. Eise- nach, and J. J. Ellner. Predicting the outcome of therapy for pulmonary tuberculosis. *Am J Respir Crit Care Med*, 161:1076–80, 2000.

R. S. Wallis, S. A. Vinhas, J. L. Johnson, F. C. Ribeiro, M. Palaci, R. L. Peres, R. T. Sa, R. Dietze, A. Chiunda, K. Eisenach, and J. J. Ellner. Whole blood bactericidal activity during treatment of pulmonary tuberculosis. *J Infect Dis*, 187:270–8, 2003.

S. D. Walter. Prognosis. In C. T. Armitage P, editor, *Encyclopaedia of Biostatistics*, volume 6, page 4254. Wiley, New York, 1998.

Y. Wang and J. M. Taylor. A measure of the proportion of treatment effect explained by a surrogate marker. *Biometrics*, 58:803–12, 2002.

R. A. Weinstein, W. E. Stamm, and R. L. Anderson. Early detection of false- positive acid-fast smears. An epidemiological approach. *Lancet*, 2:173–174, 1975.

C. J. Weir and R. J. Walley. Statistical evaluation of biomarkers as surrogate endpoints: a literature review. *Stat Med*, 25:183–203, 2005.

I. R. White. Commentary: dealing with measurement error: multiple imputation or regression calibration? *Int J Epidemiol*, 35:1081–1082, 2006.

D. Wilkinson. Sputum microscopy at 2 and 3 months. *Int J Tuberc Lung Dis*, 2: 862–863, 1998.

D. Wilkinson, S. Bechan, C. Connolly, E. Standing, and G. M. Short. Should we take a history of prior treatment, and check sputum status at 2-3 months when treating patients for tuberculosis? *Int J Tuberc Lung Dis*, 2:52–55, 1998.

S. A. Williams, D. E. Slavin, J. A. Wagner, and C. J. Webster. A cost-effectiveness approach to the qualification and acceptance of biomarkers. *Nat. Rev. Drug Discovery*, 5:897–902, 2006.

J. Williamson and K. Kim. A global odds ratio regression model for bivariate ordered categorical data from ophthalmologic studies. *Stat Med*, 15:1507–1518, 1996.

J. M. Williamson, K. Kim, and S. R. Lipsitz. Analyzing bivariate ordinal data using a global odds ratio. *J Am Stat Assoc*, 90:1432–1437, 1995.

F. A. G. Windmeijer. Goodness-of-fit measures in binary choice models. *Economet Rev*, 14:101–116, 1995.

J. Wittes, E. Lakatos, and J. Probstfield. Surrogate endpoints in clinical trials - cardiovascular diseases. *Stat Med*, 8:415–425, 1989.

World Health Organisation. Framework for effective tuberculosis control. Geneva, 1994.

World Health Organisation. An expanded DOTS framework for effective tuberculosis control. Geneva, 2002.

World Health Organisation. Adherence to long-term therapies: Evidence for action. Geneva, 2003.

World Health Organization. Treatement of tuberculosis: guidelines for national programmes. 3rd ed. Geneva, 2003.

World Health Organization. Guidelines for the programmatic management of drug-resistant tuberculosis. Geneva, 2006.

World Health Organization. Global tuberculosis control - surveillance, planning, financing. WHO Report 2008. Geneva, 2008.

A. Wright, G. Bai, L. Barrera, F. Boulahbal, N. Martin-Casabona, C. Gilpin, F. Drobniewsji, M. Havelkova, R. Lepe, R. Lumb, B. Metchock, F. Portaels, M. Rodrigues, S. Rusch-Gerdes, A. Van Deun, V. Vincent, V. Leimane, V. Riekstina, G. Skenders, T. Holtz, R. Pratt, K. Laserson, C. Wells, P. Cegielski, and N. S. Shah. Emergence of Mycobacterium tuberculosis with extensive resistance to second-line drugs. *JAMA*, 295:2349–2351, 2006.

L. Wu. A joint model for nonlinear mixed-effects models with censoring and covariates measured with error, with application to AIDS studies. *J Am Stat Assoc*, 97:955–964, 2002.

L. Wu. Simultaneous inference for longitudinal data with detection limits and covariates measured with errors, with application to AIDS studies. *Stat Med*, 23:1715–31, 2004.

J. Xu and S. L. Zeger. Joint analysis of longitudinal data comprising repeated measures and times to events. *J R Stat Soc Ser C Appl Stat*, 50:375–387, 2001a.

J. Xu and S. L. Zeger. The evaluation of multiple surrogate endpoints. *Biometrics*, 57:81–7, 2001b.

R. Xu. Measuring explained variation in linear mixed effects models. *Stat Med*, 22:3527–3541, 2003.

W. W. Yew and C. C. Leung. Prognostic significance of early weight gain in underweight patients with tuberculosis. *Am J Respir Crit Care Med*, 174: 236–237, 2006.

G. P. Youmans. *Tuberculosis*. W. B. Saunders, Philadelphia, 1979.

S. L. Zeger, K.-Y. Liang, and P. Heagerty. Regression models for discrete longitudinal responses : Comment. *Stat Sci*, 8:304–306, 1993.

Y. Zhang. The magic bullets and tuberculosis drug targets. *Annu Rev Pharmacol Toxicol*, 45:529–564, 2005.

F. Z. Zhao, M. H. Levy, and S. Wen. Sputum microscopy results at two and three months predict outcome of tuberculosis treatment. *Int J Tuberc Lung Dis*, 1:570–572, 1997.

M. Zierski, E. Bek, M. W. Long, and D. E. Snider. Short-course (6 month) co-operative tuberculosis study in Poland: results 18 months after completion of treatment. *Am Rev Respir Dis*, 122:879–889, 1980.

A. R. Zink, C. Sola, U. Reischl, W. Grabner, N. Rastogi, H. Wolf, and A. G. Nerlich. Characterization of Mycobacterium tuberculosis complex DNAs from Egyptian mummies by spoligotyping. *J Clin Microbiol*, 41:359–367, 2003.

A. Zumla. Tuberculosis–the tide can be turned, the battle can be won. *J R Soc Med*, 101:100–101, 2008.

A. Zumla and Z. Mullan. Turning the tide against tuberculosis. *Lancet*, 367: 877–878, 2006.

A. Zumla, R. Wallis, M. Doherty, N. Klein, S. Parida, O. Olesen, H. Lång, M. Vahedi, and P. Onyebujoh. Joint TDR/EC expert consultation on biomarkers in tuberculosis: Report of the joint TDR/EC expert consultation to evaluate the potential roles of biomarkers in the management of HIV-infected and HIV-uninfected patients with tuberculosis. Geneva, 2008.