Modelling departure from randomised treatment in randomised controlled trials with survival outcomes

Thesis submitted in accordance
with the requirements of the University of Liverpool
for the degree of Doctor in Philosophy
by

Susanna Dodd

December 2014

To the memory of my father, Jeremy Hinds.

I hope you would have been proud of me, dad.

Abstract

Randomised controlled trials are considered the gold standard study design, as random treatment assignment provides balance in prognosis between treatment arms and protects against selection bias. When trials are subject to departures from randomised treatment, however, simple but naïve statistical methods that purport to estimate treatment efficacy, such as per protocol or as treated analyses, fail to respect this randomisation balance and typically introduce selection bias. This bias occurs because departure from randomised treatment is often clinically indicated, resulting in systematic differences between patients who do and do not adhere to their assigned intervention.

There exist more appropriate statistical methods to adjust for departure from randomised treatment but, as demonstrated by a review of published trials, these are rarely employed, primarily due to their complexity and unfamiliarity. The focus of this research has been to explore, explain, demonstrate and compare the use of causal methodologies in the analysis of trials, in order to increase the accessibility and comprehensibility by non-specialist analysts of the available, but somewhat technical, statistical methods to adjust for treatment deviations.

An overview of such methods is presented, intended as an aid to researchers new to the field of causal inference, with an emphasis on practical considerations necessary to ensure appropriate implementation of techniques, and complemented by a number of guidance tools summarising the necessary clinical and statistical considerations when carrying out such analyses. Practical demonstrations of causal analysis techniques are then presented, with existing methods extended and adapted to allow for complexities arising from the trial scenarios.

A particular application from epilepsy demonstrates the impact of various statistical factors when adjusting for skewed time-varying confounders and different reasons for

treatment changes on a complicated time to event outcome, including choice of model (pooled logistic regression versus Cox models for inverse probability of censoring weighting methods, compared with a rank-preserving structural failure time model), time interval (for creating panel data for time-varying confounders and outcome), confidence interval estimation method (standard versus bootstrapped) and the considerations regarding use of spline variables to estimate underlying risk in pooled logistic regression.

In this example, the structural failure time model is severely limited by its restriction on the types of treatment changes that can be adjusted for; as such, the majority of treatment changes are necessarily censored, introducing bias similar to that in a per protocol analysis. With inverse probability weighting adjustment, as more treatment changes and confounders are accounted for, treatment effects are observed to move further away from the null. Generally, Cox models seemed to be more susceptible to changes in modelling factors (confidence interval estimation, time interval and confounder adjustment) and displayed greater fluctuations in treatment effect than corresponding pooled logistic regression models. This apparent greater stability of logistic regression, even when subject to severe overfitting, represents a major advantage over Cox modelling in this context, countering the inherent complications relating to the fitting of spline variables.

This novel application of complex methods in a complicated trial scenario provides a useful example for discussion of typical analysis issues and limitations, as it addresses challenges that are likely to be common in trials featuring problems with nonadherence. Recommendations are provided for analysts when considering which of these analysis methods should be applied in a given trial setting.

Acknowledgements

For financial support I am grateful to the National Institute of Health Research who gave me this opportunity by funding this project.

I would like to thank my supervisors, Paula for helping me to always keep sight of the bigger picture, Ian for help with methodological insights and technical support, Ruwanthi for daily encouragement while we shared an office and Marta for so often giving freely of her time to assist me.

I am also indebted to numerous other colleagues in my department, who supported me on a daily basis through the past five years; in particular thanks go to Girvan Burnside for help with thesis format, Laura Bonnett for background information on SANAD and Kath and Karen for administrative support and frequent encouraging conversations!

I am extremely thankful for the unwavering support of my brothers, who encouraged me on numerous occasions not to give up when things looked bleak, and to my three beautiful girls who always saw past the stressed maternal figure and constantly believed I would reach the status of Dr Dodd!

Finally, I would like to thank my dear mother, who bore the brunt of my hard times, reminding me that

"Success is going from failure to failure…. without losing enthusiasm" (Winston Churchill)

Publications

Work included in this thesis has been published in the following peer-reviewed journal article.

Chapter 6

Dodd S, White I, Williamson P. (2011) Departure from treatment protocol in published randomised controlled trials: a review. Trials; 12 (Suppl 1): A129.

The following publication is also referred to in the thesis:

Robson V, **Dodd S**, Thomas S. (2009) Standardized antibacterial honey (Medihoney) with standard therapy in wound care: randomised controlled trial. Journal of Advanced Nursing; 65(3): 565-575.

Table of contents

Abbreviations

AED	Anti-epileptic drug
AFTM	Accelerated failure time model
ARC	AIDS-related complex
AT	As treated
BMJ	British Medical Journal
CACE	Complier adjusted causal estimation
CBZ	Carbamazepine
cDAG	Causal directed acyclic graph
CDP	Coronary Drug Project
CI	Confidence interval
ER	Exclusion restriction
IPW	Inverse probability weighted
IPTW	Inverse probability of treatment weighted
IPCW	Inverse probability of censoring weighted
ISC	Inadequate seizure control
ITT	Intention to treat
JAMA	Journal of the American Medical Association
LOCF	Last observation carried forward
LTFU	Loss to follow up
LTG	Lamotrigine
MEMS	Medication event monitoring system
NEJM	New England Journal of Medicine
NUC	No unmeasured confounders
OS	Overall survival
PD	Pharmacodynamic
PFS	Progression-free survival
PK	Pharmacokinetic
PO	Potential outcome
PP	Per protocol
PS	Principal stratification
RBEE	Randomisation-based efficacy estimate
RCT	Randomised controlled trial
RPSFTM	Rank preserving structural failure time model
SANAD	Standard and new antiepileptic drugs
SMM	Structural mean model
SNMM	Structural nested mean model
SUTVA	Stable unit treatment value assumption

TTE	Time to event
TTP	Time to progression
T12mR	Time to 12 month remission
TFS	Time to first seizure
UAE	Unacceptable adverse events
VPS	Sodium valproate
WD	Weight-determining (model)
WHO	World Health Organisation
WO	Weighted outcome (model)

1. Introduction

The focus of this thesis is an investigation into the issue of nonadherence to prescribed intervention in randomised trials, and a demonstration and comparison of statistical methods that may be used to appropriately adjust for such nonadherence. In particular, the following research questions will be explored:

- What is the extent, nature, impact of and potential solutions for nonadherence in general practice and in randomised controlled trials?
- What methods are available to adjust for such nonadherence in randomised trials and are these methods applied?
- What are the assumptions made, limitations of and conclusions drawn from application of these methods in trials demonstrating various deviations from the treatment protocol?

Nonadherence to medical advice has been recognised as a widespread problem since the 1970s but its impact on clinical trial analyses and results is less well recognised. Failure to appropriately allow for changes to randomised treatment in trials may lead to biased results and misleading conclusions.

Analysis methods must reflect the clinical research question of interest. When estimating the effectiveness of treatment policy (in other words, the effect of being *assigned* treatment) treatment changes can be ignored in the analysis, under the assumption that such treatment changes mirror what will happen in practice outside of the trial setting. Ignoring such treatment changes will, however, distort estimation of the efficacy (or biological effect) of treatment (in other words, the effect of treatment *received*).

Randomised controlled trials (RCTs) are considered the gold standard study design, as random treatment assignment provides balance between treatment arms with respect to measured and unmeasured covariates that may influence prognosis. Simple but naïve statistical methods that purport to estimate treatment efficacy, such as 'per protocol' (PP) or 'as treated' (AT) approaches, fail to respect this balance produced by randomisation and typically lead to selection bias. This occurs because departure from prescribed or randomised treatment is often clinically indicated, resulting in systematic differences between those subgroups of patients who do and do not adhere to their assigned intervention.

There exist more appropriate methods to adjust for departure from randomised treatment but these are rarely employed by trial statisticians, in part due to the fact that they are not widely known. This thesis will investigate their use in a number of expository trials.

In Chapter 2, the problem of nonadherence in general medical practice is described, followed by the implications of nonadherence for the design, analysis and interpretation of RCTs in Chapter 3.

In Chapter 4, a number of expository RCTs are described in terms of the form and extent of nonadherence to randomised treatment that occurred and the resulting impact on trial analysis. Some of these trials will be used to illustrate the application of particular statistical methods to adjust for such nonadherence in this and subsequent chapters.

In Chapter 5, an overview is provided of the statistical methods that may be used to adjust for nonadherence when analysing RCT data. The issue of confounding and selection bias, introduced with use of naïve but common methods of analysis such as PP or AT analyses, is discussed. An overview of causal methods is provided, introducing the concept of potential outcomes, and two methods appropriate for analysis of survival data (the rank preserving structural failure time model (RPSFTM) and inverse probability of censoring weights (IPCW)) are described in detail. In the spirit of the aims of the funding body (NIHR) supporting this studentship, this overview of statistical methods is provided as an aid to researchers who are new to the field of causal inference, and has thus been presented with an emphasis on practical considerations necessary to ensure appropriate implementation of techniques, with a particular focus on interpretation of methods, rather than technical detail.

In Chapter 6, a review of recently published RCTs is presented, summarising the degree of nonadherence reported in RCTs and the extent to which trialists attempt to adjust for this nonadherence in their analysis.

In line with the remit of this PhD, Chapter 7 provides a clear description of the necessary practical and statistical considerations that are involved when implementing causal methodology, in order to clarify the process and enable other non-specialists to more easily carry out their own causal research. This chapter presents recommendations and guidelines to trial analysts in order to simplify and facilitate wider use of available causal estimation techniques, thus addressing potential barriers to the wider use of such methods by the trial community. In particular, a number of pictograms are provided to clarify the appropriate methods to use and statistical issues that must be considered, according to the particular trial scenario.

The next three chapters demonstrate how to translate a clinical question of interest into an appropriate causal analysis procedure, particularly for trials with survival outcomes (which typically feature deviations from treatment protocol). These practical applications demonstrate how clinical considerations and data constraints must appropriately inform the statistical choices, not just as part of the modelling exercise but also in preparation of data. These issues are demonstrated in trials of contrasting complexity featuring different compliance issues.

In Chapter 8, the RPSFTM is used to adjust for departure from randomised treatment in three trials (the honey trial, SANAD arm A and SANAD arm B); the applicability, appropriateness and limitations of applying this model are discussed for each trial setting.

In Chapter 9, the IPCW model is demonstrated in the most complex of these trial settings (SANAD arm B). The IPCW methodology relies on the assumption of no unmeasured confounders (NUC), necessitating a variable selection process to determine which of the many measured covariates should be adjusted for in this model.

Considerations for this selection process include choice of threshold for inclusion of covariates, forward versus backward selection procedures and reliability of selection.

The work in these chapters aims to investigate the impact of various statistical factors on treatment effect estimates when adjusting for treatment changes, including choice of model (survival versus pooled logistic regression (PLR) for IPCW, and IPCW versus RPSFTM), time interval (for defining time-varying covariates and outcome variables: monthly, fortnightly or weekly intervals), confidence interval (CI) estimation method (standard versus bootstrapped CIs), adjustment for time-varying confounding variables (seizure count, AE count and dose) and reason for treatment changes (to allow for the competing reasons of inadequate seizure control, unacceptable adverse events and personal choice).

In Chapter 10, the IPCW and RPSFTM are compared in the context of the SANAD arm B, using the standard set of covariates determined by investigations described in Chapter 9, highlighting the challenges and limitations associated with each method.

Chapter 11 concludes with a summary of the main findings of the thesis, ideas for further development of methods and recommendations for trialists and statisticians when analysing and reporting trial nonadherence.

2. Nonadherence in general clinical practice

2.1.Introduction

The aim of this thesis chapter is to describe the degree to which nonadherence is present and its impact in general medical practice. This will provide an important grounding for the purposes of this thesis, given that it is necessary to understand the issues relating to nonadherence in the context of routine healthcare before moving on to considering its impact in clinical trials. This chapter therefore summarises a review of the clinical research literature on adherence to prescribed intervention, including a description of the extent, nature and causes of nonadherence to prescribed treatment in general practice.

2.2.History of compliance research

It has long been recognised that patients will often fail to follow medical advice, with the earliest record of nonadherence dating back to Hippocrates in 400 BC who is believed to have advised "keep watch also on the faults of the patients, which often make them lie about the taking of things prescribed". The first record in modern medicine was by Robert Koch in 1882, who complained that non-adherent patients were 'vicious consumptives, careless and/or irresponsible' (1).

Research into the area of nonadherence began in the early 1970s when, in 1972, David Sackett realised that poor response to treatment in hypertensive patients was likely to be due to non-compliance. The McMaster Workshop/Symposium on compliance with therapeutic regimens was held in 1974, and 'patient compliance' was included as an official Medical Subject Heading in the US National Library of Medicine in 1975. Sackett and Haynes' landmark book "Compliance with therapeutic regimens' was published in 1976 (2), providing the first, and most commonly cited, definition of medical compliance: "the extent to which a patient's behaviours (in terms of taking medication, following diets, or executing lifestyle changes) coincide with health care providers' recommendations for health and medical advice" (3).

Over the next twenty years, more than 10,000 research papers were published investigating the issue of nonadherence (4). This extensive research has demonstrated that poor compliance with recommended medical intervention is a common problem, widespread across all areas of medicine, irrespective of disease type or severity (5, 6).

2.3.Types of nonadherence

Specific manifestations of nonadherence vary according to the type and duration of prescribed intervention or medical advice.

Adherence to prescribed drug treatment (referred to in this thesis as 'medication adherence') may be categorised into three phases relating to the start, execution and end of the treatment period, defined respectively by Vrijens (3) as initiation, implementation and persistence. Typical patterns of medication nonadherence include taking too much or too little (even none) of the prescribed course of medication, or taking medication at different times or intervals, or under contraindicated conditions, to those directed (7).

Failing to initiate therapy (or 'primary non-compliance') occurs, for example, when patients receive a prescription but fail to have it made up at a pharmacy. During the implementation period, patients may not follow the correct treatment schedule by taking treatment at incorrect doses, times or intervals; the most common form of such nonadherence is delaying or omitting doses, referred to as 'partial nonadherence' (4, 8). Patients may also repeatedly stop and restart treatment, taking 'drug holidays' (stopping treatment for at least three days before restarting) or may improve compliance around the time of follow up appointments (known as 'white-coat compliance') (9). Premature cessation of prescribed treatment occurs, for example, when patients fail to obtain repeat prescriptions or simply decide to stop taking their treatment.

Nonadherence may also result when patients endeavour to seek out, receive and persevere with other forms of treatment, especially those involving lengthy and complex processes. General forms of nonadherence to medical intervention carrying consequences from a population-wide, as well as a personal, perspective include delay in

seeking appropriate health care, refusing to participate in screening programmes and failure to attend medical follow up appointments (4, 9).

2.4.Prevalence

Compliance is typically higher with curative or short-term symptom-relieving, rather than preventative, treatments (10). As such, nonadherence to prescribed medication tends to be a greater problem with chronic rather than acute diseases, especially when the disease is asymptomatic or when nonadherence seems to present no obvious, immediate or drastic consequences. The likelihood of nonadherence in chronic conditions is also affected by the long term nature of treatment, along with the need for continual monitoring or follow up, all of which can present challenges to adherence. Haynes (11) gives the example of a hypertensive patient who must adhere not only to long-term treatment regimens but also to advice on lifestyle factors (weight, diet, exercise, stress) as well as attending numerous investigative and monitoring appointments, all potentially for the rest of their lives. Haynes (6) estimates that up to half of patients drop out of medical care completely within the first year of starting a treatment regime for chronic conditions, and of those who persist with treatment, one third do not take sufficient medication to achieve therapeutic benefit.

As such, estimated adherence rates with long-term interventions for chronic diseases are typically lower (40-50%) than for acute illness or short-term interventions (70-80%) (12). Even lower compliance is observed for behavioural or lifestyle prescriptions (20-30%), and only 10% of patients are believed to comply with advice on smoking or weight loss in the long term. Although approximately half of patients with chronic disease are believed to demonstrate some form of nonadherence, this estimate varies more widely in certain disease states; for example, non-compliance is typically low with

prescriptions of oral contraceptives (8%) but may be up to 70% for long term treatment of arthritis (6).

In the past 60 years, disease burden has shifted from acute to chronic disease as modern medicine has advanced and life expectancy in developed countries has increased. Noncommunicable diseases, mental health disorders, HIV/AIDS and tuberculosis made up more than half of all global illness in 2001 and are expected to exceed two thirds of all illness by 2020 (5). This shift in disease profiles, along with the high rates of nonadherence associated with chronic disease, has led to increasingly significant worldwide consequences of nonadherence, both clinical and economic; the impact of these consequences has been recognised for decades, with non-compliance having been designated "America's other drug problem" (13).

2.5.Consequences

"Drugs don't work in patients who don't take them."(14)

In chronic conditions (such as diabetes, AIDS and asthma), patient's disease may worsen as a direct consequence of inadequate compliance, with clinical consequences in terms of both decreased efficacy and increased toxicity. Medication nonadherence may result in more extreme relapses than when treatment is taken as recommended, and may decrease patients' chances of responding to treatment in the long term. Patients may also risk adverse effects, rebound or recurrent first-dose effects (for example, when therapy is restarted after a drug holiday), accidents (for example, if ignoring advice regarding lifestyle changes necessary when taking medication) and treatment dependence (5, 15).

Further repercussions on patient health may occur if undisclosed non-compliance leads physicians to make inappropriate decisions regarding treatment (16). For example,

white-coat compliance may cause clinical confusion, as short-term responses to treatment (for example, blood pressure) measured at clinic visits may be normal, reflecting the patient's improved compliance to treatment prior to appointments, but longer term outcomes may be poor due to general nonadherence otherwise (17). If clinicians are unaware of patients' non-compliance, they may deem it necessary to (inappropriately) prescribe higher doses or more invasive or stronger, more toxic medication (18).

Nonadherence in infectious disease carries both direct and indirect consequences. For example, underdosing of anti-infective drugs (caused by delayed or omitted doses, premature discontinuation of treatment or drug holidays in particular) can lead, not only to poor outcome on the part of individuals involved, but also to consequences affecting the wider population, such as drug-resistance caused by excessive microorganism replication in the presence of inadequate or intermittent doses (19, 20).

Nonadherence to appropriate medication prescription impacts not only the health of individual patients but also the health care system and society as a whole (21). **Economic consequences** result not only because of wasted medication but also because of increased health care demands to deal with subsequent unnecessary health deterioration (21), manifested as disease exacerbation, crisis or relapse (5). The cost to the NHS of unused medication has been estimated at £100 million per year (21), while the overall cost of medication nonadherence in US is estimated at $100 billion per year (4). The burden on health care systems increases as a result of more frequent hospital admissions, increased length of hospital stay and increased expenditure on health care. One tenth of US hospital admissions are reported to be due to inappropriate use of medication (while between one and two thirds of all *medication-related* admissions are caused by poor medication adherence (14)), and nearly one quarter of admissions to

nursing homes are said to be caused by patients' inability to manage their own treatment (4).

The clinical and economic consequences of nonadherence vary according to the severity of the patient's condition and their comorbidities. Failing to correctly follow treatment for a life-threatening disease, for example, incurs greater costs than nonadherence to treatment for a mild condition (22). Indirect costs result from patients' reduced quality of life (5) and loss of productivity (18). Additional costs may result if apparent poor response to treatment (caused by undisclosed non-compliance) causes the clinician to prescribe an alternative more costly or complex treatment (18).

2.6.Measurement of adherence

Measurement of patient adherence has occurred since the time of Hippocrates, who noted whether or not patients had taken various prescribed potions when recording their effectiveness. Numerous **direct** and **indirect** adherence measurement methods exist, each featuring their own particular advantages and disadvantages.

Direct measurement methods generally provide more accurate compliance information than indirect methods but they suffer from a number of disadvantages which preclude their use in long term compliance assessment, most notably being labour intensive or invasive. Although theoretically the most accurate measurement technique available, **direct observation** of therapy (observing a patient taking their prescribed drug) still provides opportunity for patients to hide medication in their mouth, rather than swallowing it, and is feasible only for monitoring single-dose or intermittent treatment of hospitalised patients rather than long term self-administered treatment (9, 14).

Biochemical measurement in body fluids (blood, saliva or urine) of drug levels, metabolites or biological markers added to medication provides an alternative direct

method to assess medication adherence, but this too suffers from major disadvantages. Such measurements are subject to pharmacokinetic variation between patients (being affected, for example, by diet and rates of absorption and excretion) and thus may be misleading or highly variable, they may be unacceptably invasive or prohibitively expensive and not all medications can be readily measured in bodily fluids. Drugs level in body fluids may be distorted by timings of doses taken and in particular by white-coat compliance, as this improved compliance prior to a follow up visit will overestimate typical levels of compliance (1, 14).

Indirect measurement measures tend to be more practicable, convenient and affordable but at the expense of accuracy.

Patients are typically nervous about disclosing any deviation from their prescribed regimen and thus are likely to overstate their compliance when self-reporting (10, 20). Thus various commonly used **patient self-reporting methods** (such as interviews, treatment diaries or questionnaires) are all prone to intentional overestimation of adherence by the patient, as well as recall error. Diaries filled in concurrently may help to overcome the problem of inaccurate recall but are likewise easily distorted by patients; furthermore patients may resent the constant reminder of their disease that these diaries present and thus fail to complete them regularly. Given the reluctance on the part of patients to disclose adherence problems when faced with direct questioning or open discussions with their treating doctor, it is unsurprising that doctors have been shown to be poor at detecting nonadherence in their patients, typically doing so no better than chance (6, 23).

Pill counts (or weighing) of returned pill bottles is another easy, inexpensive method of assessing compliance to longterm medication and has been the most common method of adherence assessment used to date in clinical research. This method provides an

indication of proportion of drug taken but captures no information on the timing of doses taken; furthermore it has been demonstrated to be entirely unreliable when compared with more objective methods, unsurprising given the opportunity for patients to simply discard or hoard any left-over medication (14, 23).

Patients' **prescription refill dates** may be obtained from supplying pharmacies' records to inform on the regularity with which patients obtain prescribed medication, but such data only provide information on the supply, rather than intake of medication, and will be accurate only if a patient takes all of the dispensed medication and utilises a single pharmacy whose records are complete and up-to-date.

A relatively new technology, first developed in 1977 and taking the form of a microprocessor incorporated into electronic monitoring devices fitted to treatment dispensers, has provided the most reliable indirect medication adherence data to date (17). Such **medication event monitoring systems** (MEMS, where "medication event" is defined as the set of actions needed in order to take a dose of medication) store time and date data when the relevant treatment dispenser (for example, medicine bottle, blister pack, inhalator or eye drop dispenser) is triggered.

The detailed dose timing information obtained from MEMS has been used to test the reliability of other methods, demonstrating, for example, the unreliability of pill counts and other patient self-reporting methods. By recording the exact time and date of opening or activation of drug dispensers, MEMS data have provided valuable insight into the patterns of patients' drug-taking, unearthing various typical manifestations of nonadherence such as drug holidays and white-coat compliance (17). Such data can also be input into pharmacokinetic/pharmacodynamic (PK/PD) models, as the dosing records allow prediction of the time courses of drug plasma concentrations and drug actions (15).

In providing a complete dosing history and thus allowing detailed comparison between prescribed and realised drug regimens, MEMS is considered to be the most accurate and useful single approach available for measuring medication adherence in clinic trials or adherence research (17). These devices are, however, prohibitively expensive and cumbersome for routine practice (5, 6). Furthermore, the account of treatment taking provided by such devices, although detailed, nevertheless only provides an indirect account of treatment timing rather than a direct measure of treatment ingestion. There is therefore the potential for patients to mislead (if they regularly activate the treatment dispenser at the correct dose times but then fail to actually take the intervention); although such behaviour would require substantial effort on the part of the patient and is therefore unlikely to be sustained over long periods (15), there is some evidence indicating such behaviour (in a cancer setting with self-administered chemotherapy) (16).

Thus, even the most reliable methods for measuring patient's self-administered treatment provide only an estimate of actual intake (5). The resulting lack of a gold standard measurement technique for measuring compliance has been cited as a major barrier to compliance research (4, 22).

2.7.Causes

Despite these difficulties in reliably measuring compliance, recognition of its impact and extent has prompted a huge amount of research into its causes. Traditionally **patient factors** have been considered the primary cause of non-compliance (14) but research has demonstrated that demographics (such as age, sex and social class) do not strongly predict compliance (9). Age impacts on a patient's practical ability to comply, with elderly patients more like to suffer problems with memory, vision, mobility and

cognition; however elderly patients may also be more concerned and focused on their health than younger individuals so, with appropriate help, may be more likely to comply to prescribed regimens. Heterogeneous findings regarding educational levels have been reported; some studies have demonstrated higher compliance among those with higher education levels (because of a greater understanding of the disease and its consequences) while others suggest that a lower educational level will enhance compliance (because of a greater trust in clinical advice). Similarly, contradictory results have been reported for gender, ethnicity and marital status (18). Inconsistent results regarding demographic factors may be due to underlying correlations with variables relating to psychological influences or a patient's cultural and socioeconomic background (18).

Social and economic factors also play a role, along with the practical limitations presented by daily life. Employment status may impact on practical limitations on time due to work commitments or may provide insufficient funds to meet the cost long term treatment. The quality of health care provision from **health care systems** impacts on compliance, for example in terms of affordable medication, adequately trained and supported health care workers, continuity of care from the same provider, sufficiently long appointment times and convenient access to appropriate health care facilities (5, 18).

Social support may be required to overcome practical problems with compliance, or may help to reinforce positive attitudes to treatment and thus enhance patients' motivation (18). On the other hand, patients who fail to comply with medication may in fact relish the increased care and attention received from health care professionals, friends and family as a result of the clinical consequences of their non-compliance, which will serve only to encourage their non-compliance further.

Treatment factors which impact on adherence include the convenience, duration, cost and complications involved. Dosing frequency and number of prescribed treatments both affect compliance, though greater problems are caused by the number of daily doses required rather than the number of different prescribed treatments (9). Invasive or complicated treatments will discourage compliance, and side effects of treatment may negatively impact a patient's motivation and trust in the treatment or prescribing clinician (18).

Disease factors, such as severity of symptoms, duration of disease and rate of disease progression, also impact on adherence. Long term illness typically leads to a greater probability of compromised compliance, although newly diagnosed chronic disease patients may demonstrate lower compliance rates than those who have had a disease for many years, perhaps because they feel no sense of urgency or have not yet experienced the adverse effects of living with the disease (18). Noticeable improvement of symptoms with treatment, or worsening of symptoms when treatment is discontinued, encourage compliance, while comorbidities may hinder compliance, especially depression, poor cognition and alcohol or drug abuse (5). Patients are naturally less likely to adhere when taking long-term preventative medication for asymptomatic conditions presenting no immediate risk than for conditions with life-threatening or serious consequences. Premature cessation of prescribed treatment, for example, occurs most often in asymptomatic disease when risk is not immediately obvious (for example, in hypertension) (19).

Patient knowledge regarding disease and treatment is known to influence compliance rates (18). Patients need sufficient information on the impact of the prescribed treatment on the course of disease and the consequences of suboptimal treatment compliance and should be provided with written information to complement verbal

instructions. The quality of this information will impact on their likelihood of adherence, as patients may forget, not understand or simply ignore inadequate instructions.

Indeed non-compliance is not always intentional; it may simply be due to ignorance or lack of understanding on the part of the patient. A patient may occasionally forget to take treatment (indeed, forgetfulness has been reported as the leading reason for poor compliance with long-term medication) (14, 24, 25), or may be more severely limited when following a prescribed treatment regimen, either by their mental or physical capability or because of inadequate financial or health care resources (21).

Patient beliefs and their experience of their prescribed medication and medication as a whole is a very important determinant of compliance, and patients will balance their beliefs of the benefit and harms of treatment when deciding on how to take their medication. The Health Belief Model has been applied to the study of compliance, such that patients are believed to comply more readily if they believe they actually have the disease in question, that they will benefit from following the prescribed course of therapy or will otherwise be susceptible to consequences of the disease course (9). Patients make decisions regarding treatment compliance by, consciously or unconsciously, weighing up the health benefits gained from complying with treatment against the practical implications of compliance on their daily lives, taking into account cost, inconvenience or side effects (4).

Psychological factors, such as embarrassment, anxiety or anger regarding disease or medicine-taking, fear of side effects, interactions with other treatments or developing addiction to long-term treatment, or poor motivation to persist with treatment or behavioural changes, will impact on a patient's compliance status (26). Those who are not concerned by their ill-health will be less likely to comply, especially if the act of

taking regular medication presents physical or practical problems (1). Although disease severity has been shown to correlate with compliance, evidence suggests that a patient's perception of their health status may in fact be more influential on compliance than actual disease severity (18). **Negative beliefs about medication in general**, for example that they are unnatural or harmful, or patients' suspicions regarding chemicals or science, will impact on adherence. Patients may be in denial regarding their diagnosis or may wish to avoid the constant reminder of their illness presented by taking regular medication (10, 27). Nonadherence may be a coping mechanism, if patients are otherwise unable to express their fear, anger or frustration with their situation (12).

The quality of **communication** between the patient and their clinician is cited as a key factor in determining compliance (12, 28) though it is difficult to measure, quantify and therefore adequately study the impact of precise factors relating to the **patient-doctor relationship** (9). It is acknowledged that clinicians often fail to inform patients of crucial information when initially prescribing treatments (29). Insufficient consultation time, or poor rapport, with prescribing clinicians may leave patients with unanswered questions (for example, regarding diagnosis, absence of symptoms, expected adverse events, and expected delay from start, or duration, of treatment effect) (9).

Prescribing practice varies amongst clinicians; thus, when confronted with conflicting dosing advice from different clinicians, patients may be inclined to instead rely more heavily on their own judgement. Blackwell (27) sums it up nicely: "…the physician will be expected to prescribe with only approximate accuracy, and the patient will be expected to comply with only modest fidelity. Thus mankind has been able to survive bleeding, cupping, leeches, mustard plasters, turpentine stupes, and Panalba." As such, patients with chronic diseases often become "experts of their own disease" and, regardless of clinical advice, will adjust their treatment intake (e.g. dosing levels or

timing) according to perceived therapeutic or side effects in order to balance perceived benefit and harms, or to determine their own personal minimum effective dose of treatment (9).

Thus extensive research into the causes of non-compliance has revealed numerous associated factors related to the patient, disease, treatment, health care provider and health care system (4, 5). However, despite suggestions of causal influence, none have demonstrated consistent or fully predictive association with compliance (9), due in part to the difficulties in accurately measuring and quantifying the complex interaction between these factors.

2.8.Terminology

"One of the most striking reasons for the lack of progress in compliance research is the absence of a crucial factor: the patient's perspective." (9).

In spite of extensive research into its causes in the 1970s and 1980s, nonadherence continued to be a problem, with no apparent consistent predictors or solutions to the problem. As Becker commented "Patient non-compliance has become the best documented but least understood health related behaviour" (30). This research demonstrated not only the pervasive problem of discrepancy between prescribed and received interventions, but also of the moral limitations of the original term "compliance" (9). This traditional term has negative connotations, implying a paternalistic relationship between a doctor and their patient, whereby patients should submit to their doctor's advice, and failure to do so is considered disobedience (9). Indeed, the Cambridge Dictionary defines compliance as "the act of obeying an order, rule, or request". Furthermore, use of the term "non-compliance" has been cited as a potential tool by health care providers to label and discredit their patients (31).

Justified use of the term 'compliance' implies certain underlying assumptions which may not be realised in practice, namely that the clinician correctly diagnoses the patient's condition and prescribes the most appropriate individually tailored treatment with clear instructions to ensure adherence. Although medication is usually prescribed by clinicians in good faith that it will provide the best help available to the patient, a clinician is unable to guarantee to a patient that the benefit of the prescribed treatment will indeed outweigh its harm; "compliance" with a prescribed treatment may therefore not be in a patient's best interest. Rather than assuming a purely passive role, patients (especially those with long-term illness) may need to evaluate and make decisions regarding their treatment intake, in order to maximise benefits and minimise harms. Donovan (4) raises the point that recommended doses are based on average patient responses, but given the great variation between patients' metabolism rates, it may therefore be necessary for individual patients to experiment with dosage and timings of treatment in order to find their optimal dose requirements. Thus when patients alter their drug intake levels, this may in fact be a "sensible and rationale response".

Therefore, with increased recognition of these multifactorial causes of non-compliance, including not only patient factors but also those related to the disease and drug itself, the treatment providers and health care system, there was a move away from using the term "compliance". Indeed, researchers have long recognised the need to work with a "problem-solving approach" to encourage patients in developing strategies in working towards their own self-management of long-term condition and treatment (27), and that patients should be encouraged to assert their own needs, preferences and expectations as regards their treatment (12). Even Sackett's landmark book on compliance (2) referred to the need for a "tailored consensual regimen" with consideration of practical problems encountered by the patient and the need to prevent any notion of blame when

assessing compliance behaviour (32), but this focus has only recently permeated into compliance research.

Research in the 1990s therefore shifted from an assumption of an authoritative relationship between doctor and patient, introducing a new regard for the need to understand a patient's motivations and attitudes towards treatment, and the importance of the patient as ultimate decision maker. Thus, in an attempt to move away from the blame associated with the term 'compliance', the term 'adherence' was introduced in the WHO document (5), defined as 'the extent to which a person's behaviour – taking medication, following a diet, and/or executing lifestyle changes, corresponds with agreed recommendations from a health care provider". It was hoped that this term would better portray the notion of cooperation and partnership between patient and prescriber, the main difference with compliance being the need for the patient to agree to the prescribed course of therapy rather than simply passively following orders (9, 14). However, over time the term "adherence" became almost synonymous with "compliance" because of the persisting underlying implication of obligation on the part of the patient to follow the assigned treatment plan (with the Cambridge dictionary defining adherence as "behaving exactly according to rules, beliefs").

Recognition of this flawed terminology led members of the working party of the Royal Pharmaceutical Society to propose a new framework to replace the traditional "compliance" model. The term "concordance" was introduced, describing a situation where the health care provider works with the patient to determine the optimal course of treatment for the patient, with consideration of their personal needs and values, thus aiming for "agreement and harmony" between patients and prescribers (9). Concordance (defined in the Cambridge dictionary as "the state of there being agreement or similarity between things") "signifies the practical and ethical goal of

treatment", founded on the patient as decision maker with encouragement from an empathetic treatment provider (9). Other terms that have been used synonymously with concordance include 'cooperation', 'agreement' and 'therapeutic alliance', all of which suggest a 'meeting of the minds' of patients and health care providers (3) with mutual respect between doctor and patient for their (clinical and personal) contribution towards the treatment decision. Key to the notion of concordance is acceptance that "increasing compliance with prescribing instructions is not as important as meeting patients' individual needs and priorities" (10).

Concordance in practice

The move towards concordance requires a shift from the traditional relationship between patient and clinician towards a more equal partnership with mutual respect for the respective personal and clinical expertise that each has to offer. Patients need to feel confident that disclosing their honest opinions about their difficulties with compliance or their treatment preferences will be received without judgement or reprimand. Concordance is particularly important when treatment is first prescribed or when changes are made to prescribed treatments (3) requiring prescribing clinicians to consider not only the effectiveness but also feasibility and acceptability of treatments (32).

However, barriers to achieving concordance in practice are substantial, with limited time in initial and follow up consultancy appointments, spent mostly in discussion regarding diagnosis or symptom control (32), patients' reticence to disclose their difficulties in following prescribed regimen and lack of perception on the part of clinicians to detect compliance problems.

Heath (33) feels that "patients need different information, not more of the same, and there is an urgent need for more honesty about the limitations of medicine and the uncertainties of medical knowledge". Aronson (34) agrees, stating that "clinicians' views about therapy can never be perfectly concordant with those of patients", as patients are unlikely to be able to make a sound judgement based on complex medical information (35). Aronson feels that the idea of concordance is well-intentioned but is "not relevant to the interaction between a clinician and a patient (which is not a negotiation)"; instead clinicians should focus on providing clear communication and appropriate patient support, with recognition that patients then have a right to choose whether to follow or deviate from their clinical advice.

2.9.Taxonomy

A systematic literature review of papers on adherence terminology published in the first half of 2009 demonstrated use of more than ten different terms to describe departure from prescribed treatments, including "adherence", "compliance", "persistence", "pharmionics", "therapeutic alliance", "persistency", "patient irregularity" and "pharmacoadherence" (3). Vrijens reported inconsistency in the definitions of terms and found that these definitions did not provide guidance on how to measure or report these quantities, for example when trying to succinctly but adequately summarise multiple dimensions (e.g. timing, dosage and frequency) of longitudinal time-series data on patients' dosing histories with single summary measures (36).

The need for a standard and unambiguous taxonomy has been recognised; indeed, Vrijens et al (3) identified the need for clarification not only of the terminology used to describe the study of adherence, but also of the fundamental structure of adherence in relation to both the process and measurement of adherence, in order to "facilitate a

smooth transition from conceptual to operational definitions". In their recent publication, they describe a foundation which clearly distinguishes between what they believe to be the three core elements of adherence research: 'adherence to medications', 'management of adherence' and 'adherence-related sciences'.

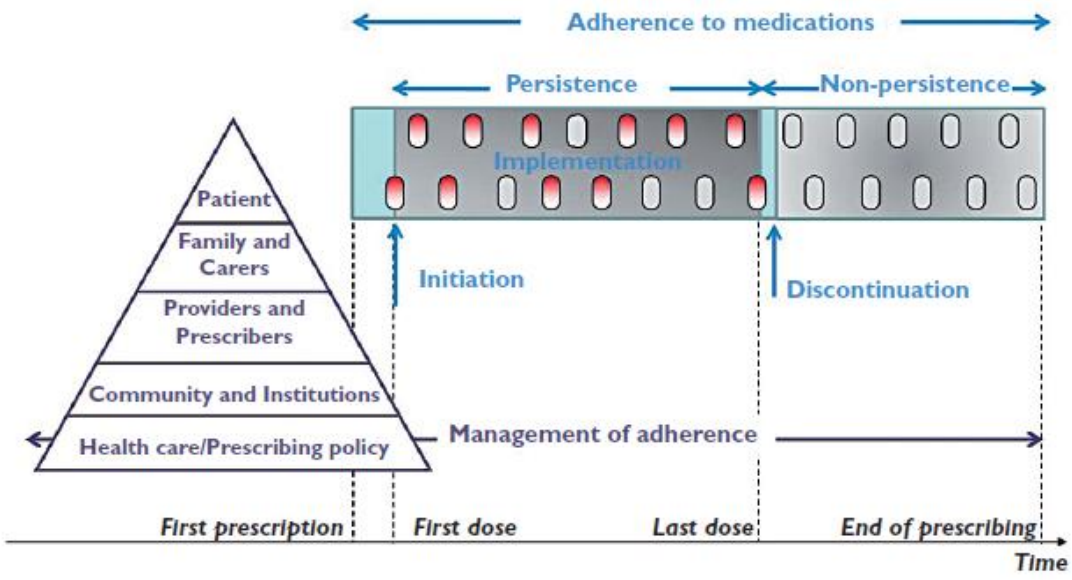


Figure 1 (from (3)): Illustration of the process of adherence to medication (light blue) and the process of management of adherence (dark blue)

The first component to this structure, 'adherence to medications', is defined as 'the process by which patients take their medications as prescribed' and consists of three stages, namely 'initiation', 'implementation' and 'discontinuation'. Initiation occurs at the point at which the patient takes their first dose of prescribed medication, and discontinuation occurs when the patient takes their final dose. The implementation period extends between these two time points, with 'persistence' being defined as the length of time between first and final doses. Therefore medication nonadherence may be manifested by a patient failing to initiate prescribed therapy, failing to follow the prescribed course of treatment over the implementation period or discontinuing treatment prematurely. This terminology framework has been used through this thesis, particularly in the review of published trials.

The second element of adherence research, 'management of adherence', relates to the process by which patients are monitored and supported in their adherence to prescribed medication by family, friends, patient support groups or health care providers and systems. The final component, 'adherence-related sciences', encompasses the many disciplines (including, for example, medicine, biostatistics, biopharmaceutics and health economics) involved in furthering understanding of the causes and impact of the nonadherence process.

One such relatively new adherence-related science is 'pharmionics', the study of how patients take medication that has been prescribed to them ('what the patient does with the drug') first introduced in 1987 to complement the two existing subdisciplines in the field of biopharmaceutics, namely pharmacokinetics ('what the patient's body does to the drug') and pharmacodynamics ('what the drug does to the patient's body').

2.10. Summary measures

Vrijens (36) discusses the limitations of terms such as 'adherence' and 'concordance' when measuring and analysing drug regimen compliance data, arguing that these "blanket" terms describe multidimensional processes but without physical dimensions and thus do not support quantitative analysis. Instead Vrijens proposes they be replaced by substituent measures, 'initiation', 'compliance' and 'persistence' in order to adequately summarise the differences between the actual and prescribed dosing regimens over the implementation period (the time over which the patient takes their prescribed treatment).

Given that initiation and discontinuation are events occurring at a particular time point, while implementation refers to a continuous time period, these three phases of the 'adherence to medication' process should be summarised separately and appropriately.

The definitions of typical summary measures used to describe these features of compliance are summarised in Table 1, along with a brief discussion of their potential limitations.

Table 1 Recommended measures to describe features of medication adherence

Feature	Definition	Typical summary measures	Limitations
Initiation	Time between receipt of prescription and first dose	Initiation rate • proportion of patients starting prescribed treatment	Potentially overly simplistic, providing no indication of the time between prescription receipt and starting treatment
Compliance	Degree of correspondence between patient's actual dosing history and prescribed dosing regimens	Within-patient summary statistics of this period; e.g. • proportion of prescribed drug taken • number of drug holidays • proportion of days when correct dosage is taken • proportion of doses taken on time	May be misleading in isolation (without further information on initiation or discontinuation) e.g. reporting average percentage of dose taken provides no information on the continuity of therapeutic action
		Therapeutic coverage • proportion of patients who are exposed to a (minimal clinically important) therapeutic dose for a minimum percentage of time	Typically difficult to justify choice of thresholds (for dose and time) required in order to define measure Requires detailed temporal dose data (e.g. from electronic monitoring device)
Persistence	Time between first dose and treatment discontinuation	Discontinuation rate • proportion of patients stopping treatment per unit time	Potentially overly simplistic, providing no indication of the length of time spent on treatment

Quantification of the implementation period is often particularly complicated, as it is necessary to compare two time-series: the prescribed drug dosing regimen versus the patient's actual drug dosing history (3). It is common to report within-patient summary statistics of this period, but such summary measures cannot fully describe the longitudinal implementation period and in isolation, without further information on initiation or discontinuation, may be misleading. Patients who take a similar proportion

of prescribed drug over the course of a trial, for example, may demonstrate very different profiles relating to initiation, discontinuation and implementation (3).

In the absence of temporal dose data (from electronic monitoring devices), it is not possible to supplement these summary measures; if, however, longitudinal dose data are available, a rather more informative summary measure which takes into account some important properties of the drug is 'therapeutic coverage' (defined as the proportion of patients who are exposed to a (minimal clinically important) therapeutic dose for a minimum percentage of time), but justification for the chosen thresholds of these properties may be difficult (37).

It is not possible or appropriate to define a universal definition of compliance across all clinical areas, given the disparate requirements in different clinical settings. The assessment of compliance is necessarily disease- and treatment-specific; as such, the most appropriate specific definition of non-compliance will depend on clinical features (such as the severity, comorbidities and consequences of disease) and the therapeutic objective of therapy, as well as (in the case of drug therapy) the duration of action of the prescribed drug relative to pattern of non-compliance manifested (known as 'drug forgiveness') (17).

2.11. Solutions and initiatives

Various methods have been implemented in an attempt to improve adherence but with only limited success, primarily as they have generally focused only on patient-related factors rather than addressing the multiple causes of non-compliance (1, 5). The most effective methods have involved a combination of multiple adherence interventions and therefore have typically been expensive, complex and labour intensive (14). Generally,

however, there has been a marked lack of research into methods to improve adherence (38).

Thus, the prevalence and wide-reaching consequences of nonadherence, along with increasing recognition of the need to work with the patient to achieve concordance rather than unrealistically expecting the patient to follow whatever is prescribed for them, has led to a number of prominent initiatives seeking solutions to the problem.

The global WHO initiative "**Adherence to Long-term Therapies Project**" was launched in 2001. The resulting report, entitled "Adherence to Long-term Therapies: Evidence for Action", was aimed at health care professionals, managers and policy makers, and promotes strategies to improve worldwide adherence rates for chronic disease therapies (5).

The Expert Patient Task Force was set up in 1999 in recognition that patients and health care professionals each have their own area of knowledge and expertise with respect to treatment of chronic disease (the clinician providing knowledge on diagnosis, disease, prognosis, treatment options and probable outcomes, while the patient adds expertise on their illness experience, values, preferences and practical considerations), suggesting the value of them working together in developing self-management programs (or Expert Patient Programs) in order to increase patient's knowledge, confidence and motivation to manage life with a chronic condition.

The Task Force for Medicines Partnership was created in 2002 by the Department of Health to promote the principles of concordance in order to improve health outcomes and patient satisfaction in the NHS. Practical initiatives were developed to encourage medication adherence, including the introduction of the Motivation for Medicines Service (pharmacy based adherence support service to support patients newly

prescribed long term medication), Medication Reviews (annual meetings for patients taking long term medication with a pharmacist, nurse or doctor in order to discuss practical problems or ask medical questions, with the aim of resolving patient's problems with adherence and thus improving effectiveness of medication and reducing waste) and other patient adherence support programmes (such as the repeat prescription service and the introduction of telephone or online helplines for those prescribed certain medicines). Training packages were developed to enable health care professionals to develop concordant consultation skills in order to support patients with taking medication.

The **ABC** (Ascertaining Barriers for Compliance: policies for safe, effective and cost-effective use of medicines in Europe, http://abcproject.eu/index.php?page=project) project was launched in 2009 with the aim of improving patient compliance in Europe. Particular aims of the project were to standardise taxonomy of, and identify causes of non-compliance, research current compliance management techniques, assess the effectiveness of compliance enhancing interventions and develop policy recommendations to promote compliance within European health care.

The **ESPACOMP** (European Society for Patient Adherence, Compliance, and Persistence) association was established in 2009 to promote study into, and ultimately provide research information to patients, health care providers, policy makers and pharmaceutical researchers on, the reasons for, and clinical and economic consequences, of patients' nonadherence to prescribed medication.

NICE guidelines (Clinical Guidelines and Evidence Review for Medicines Adherence: involving patients in decisions about prescribed medicines and supporting adherence) were published in 2009 (21), providing recommendations on increasing patient involvement in decisions about medicines and supporting patients in medication

adherence: "Patient involvement and adherence are central to medicine-taking yet these areas are less well researched than medicines themselves." The guidelines recognise the potential for harm, as well as benefit, from prescribed medication and thus proposed initiatives aimed at providing patients with informed choice and supporting them in their role as primary decision maker.

2.12. Summary

This review of the clinical literature has highlighted that the problem of nonadherence with prescribed intervention is widespread and ubiquitous, affecting all areas general medical practice, from secondary prevention measures (detecting and treating disease that has not yet become symptomatic, for example screening programmes) to recommended behavioural interventions and prescribed medical treatments. The wide reaching clinical and economic consequences of nonadherence have prompted extensive research into its causes, extent and impact spanning more than four decades. However, despite these efforts, the prevalence of non-compliance to long term treatment has remained stable, with between one third and one half of patients demonstrating some form of non-compliance. Recognition that patients are increasingly more informed on their illness and medication, often in the case of chronic illness becoming experts in their own disease, and are more willing to take responsibility for their own health care, has led to a shift in focus from the traditional concept of compliance, which implies negative consequences if a patient does not follow the prescription of an (assumedly appropriate) intervention by a paternalistic authoritative doctor. This assumption makes no allowance for a patient's circumstances, beliefs, values or expert knowledge of their own disease (especially in the case of chronic illness), or for the fact that evidence for efficacy of prescribed treatments (e.g. from clinical trials) is based on average responses (rather than on individual's needs)

potentially from small, homogenous groups of patients (and thus may not be generalisable to general populations) or indeed for the possibility of incorrect or harmful treatments or diagnosis. The term "adherence" was therefore introduced, aiming to imply more of a shared decision-making relationship with the patient being considered an equal partner with the prescribing clinician, but the assumption that the patient should follow the determined course of treatment persisted still. As such the concept of "concordance" was founded, along with a number of initiatives providing guidance to health care providers and policy decision makers on how to improve patient's treatment experience by involving them in the decision making process.

However, for successful implementation of a concordant relationship, treatment providers must have the humility to accept the validity of each patient's view and make the effort to present a more accepting approach to patients, while patients must be convinced of the value of moving from passive acceptance (or unspoken rejection) of the authority of the clinical profession towards a more active decision-making role requiring openness and honesty. In addition, there exist other substantial barriers to achieving concordance in practice, in particular limited time in initial and follow up consultancy appointments and patients' reticence to disclose (and lack of perception on the part of clinicians to detect) compliance problems. Furthermore, ethical problems are introduced when patients' preferences do not match clinical opinion of the optimal course of treatment.

2.13. Conclusions

This literature review has highlighted that nonadherence is common across much of general clinical practice, most notably in the long term treatment of chronic disease, with multifactorial causes and substantial economic and clinical consequences. Inherent

difficulties in defining and measuring compliance have historically presented major barriers to research into the impact of compliance on patient outcomes. The introduction in recent years of a standard taxonomy to describe the multidimensional features of compliance has provided a clearer framework on which to base the measurement and reporting of compliance, and an increasing appreciation for the need to work with patients in achieving concordance in order to lessen the impact of nonadherence has led to a number of high profile initiatives. Nevertheless, despite these efforts, deviation from prescribed intervention continues to feature heavily in medical practice. The next chapter will describe how such issues translate into complications in the context of clinical trials.

3. Nonadherence in randomised controlled trials

"There could not be worse experimental animals on earth than human beings; they complain, they go on vacations, they take things they are not supposed to take, they lead incredibly complicated lives, and, sometimes, they do not take their medicine." (39)

3.1.Introduction

The literature review of the previous chapter revealed the widespread and continuing problem of nonadherence in general medical practice. In this chapter, focus shifts towards the issue of nonadherence to randomised treatment in trials, in particular how such nonadherence may be manifested in trials and its impact on design and analysis.

3.2.Trial design features

The randomised controlled trial (RCT) is considered the gold standard study design for clinical research, providing the strongest form of evidence available from a single study (40, 41). The strengths of the RCT lie in its design features, most notably randomisation, blinding and allocation concealment. These design features serve to maximise a trial's **internal validity** by preventing any biases or confounding, such that any differences between treatment arms in terms of outcome can be directly attributed to treatments (42).

Randomisation prevents the introduction of bias which may occur during the treatment allocation process if, consciously or unconsciously, recruiting clinicians or patients are able to influence choice of treatment, leading to systematic allocation of patients with generally poor or good prognoses to the different treatment groups (40). In order to protect against such biases, randomisation necessarily involves two processes: not only the use of a random allocation sequence (thus preventing any predictability of future treatment allocations given knowledge of previous allocations) but also the concealment of each treatment allocation until it is actually assigned to the trial participant (known as "allocation concealment").

Assuming sufficient numbers of recruited patients, randomisation will therefore provide baseline comparability between treatment groups, both in terms of known and unknown prognostic factors, ensuring fair group comparisons. Additionally, in providing an **'ignorable' assignment mechanism**, randomisation provides a valid basis for hypothesis testing (43, 44), allowing the use of probability theory to determine the likelihood that the observed differences between treatment groups have arisen purely by chance (41).

Allocation concealment may, depending on the trial, be extended beyond the point of randomisation, such that one or more parties (patients, clinicians, assessors or analysts) remain unaware of treatment allocations during the entire trial period, a design feature known as "**blinding**". Blinding prevents bias occurring when preconceived ideas or experience of treatments affect judgment or decisions made during the trial. For example, blinding may reduce the probability of co-intervention bias (when additional treatment is given to one treatment group only) or diagnostic-suspicion bias (when the analysis or interpretation of results is influenced by knowledge of treatment). Knowledge of treatment group can also influence decisions made during outcome assessment or statistical analysis, for example, on patient withdrawals, outcomes and time points to analyse and report. Thus, although patient and/or clinician blinding is not always practical, it may be possible to conceal treatment assignment from those evaluating and analysing outcomes in order to avoid measurement bias and ensure objectivity in statistical analyses (44-46).

Blinding may be achieved by use of a placebo, an inert substance which appears identical to the active treatment under study; in the case of active-controlled trials, treatment assignment may be masked using placebos for both treatments, providing a "double dummy". Use of a placebo is necessary to distinguish the placebo effect (caused by expectation of treatment) from the true effect of treatment (47).

These design features which serve to increase internal validity do so, however, at the expense of a trial's external validity. In aiming to ensure robust, valid conclusions, such constraints render the trial a somewhat artificial clinical environment, compromising the generalisability of its results to the wider population of interest. In particular, features (such as the nature of the patient-clinician relationship, and their individual preferences)

which typically influence a patient's response to treatment or likelihood of compliance may be altered by randomisation and blinding (48).

Thus, although generalisability (external validity) of trial results is generally based on the assumption that the sample is generalisable to the target population, this may not be the case. Trial participants may be typically more motivated and interested in the treatment than others with the same condition, and this greater medical awareness may mean they are more likely to comply with treatment instruction (37). More generally, by influencing their behaviour, trial conditions may also impact on a patient's likelihood of nonadherence, altering patterns and rates of nonadherence in the trial compared to those that would typically be expected in practice. In general, trial adherence rates are usually considered to be higher than those displayed by the general public, despite the fact that trial participants may face greater barriers to adherence than those being treated in a usual care setting. For example, the trial treatment protocol may be rather more involved or demanding, and may be accompanied by more intense follow-up or invasive assessments than would be experienced in general practice, all of which can hinder participation (11).

Conversely, the extra contact with health care professionals provided by the trial setting may, in itself, provide patients with incentive and encouragement to persevere. Blinding may also help to reduce problems with patient drop out, as knowledge of treatment allocation may cause disappointment and subsequent patient withdrawal, if they perceive their allocation is the inferior treatment. On the other hand, blinding, especially by use of placebo, may mean patients are less likely to persevere with treatment, given the possibility that they are receiving a dummy drug. Furthermore, patients' uncertainty about the efficacy of a drug being tested in the trial setting may mean that trial

participants are less likely to persevere with unpleasant side effects of treatment than those who are confident of the efficacy of their treatment.

Patients may fear embarrassment if they report side effects or withdraw from treatment that turns out to be a placebo; indeed, empirical evidence has shown that patients are more likely to report side effects or withdraw from treatment if the trial involves an active control treatment rather than placebo (49). Similarly, a clinician blinded to treatment allocation may proceed more cautiously (for example with dose escalation), perhaps monitoring the patient's response and symptoms more closely than usual (46).

Thus, unnatural trial conditions may affect a trial participant's behaviour, which is turn may distort their (perceived, reported or actual) trial outcomes, thus potentially limiting the reliability and generalisability of the trial's overall conclusions.

3.3.Aim of trial

The impact of such differences from real life presented by the trial setting will depend on the aim of the trial. Ultimately the purpose of any trial will be to provide an estimate of some aspect of the intervention's usefulness in achieving its therapeutic goal in the population being studied: the estimate of interest may be the treatment effectiveness (aiming to mirror how it will be used in practice) or efficacy (use under ideal circumstances as prescribed). The importance of a trial's external validity will therefore depend entirely on its research question, which may be broadly classified as either pragmatic or explanatory.

3.3.1. Efficacy trials

'Explanatory', 'efficacy' or 'experimental' trials aim to determine the biological or pharmacological efficacy of an intervention, in order words whether or not it produces

the expected result under ideal circumstances (48), with the primary purpose of furthering scientific knowledge (50). Efficacy, also referred to "proof of principle" (51) or "method-effectiveness" (52), measures the effect of the treatment actually administered.

Haynes (11) defines an efficacy trial as one whose aim is to determine whether a "treatment does more good than harm to those who take it", whereas an effectiveness trial assesses this in those to whom it is offered. As such, a trial which aims to demonstrate the explanatory effect of treatment requires highly controlled conditions; thus explanatory trials tend to be designed to prevent blurring of the treatment effect by extraneous factors.

In seeking to determine whether a treatment works in terms of biological processes, an explanatory trial will therefore typically restrict patient eligibility criteria to ensure recruitment of those thought most likely to respond and adhere to study treatments, potentially incorporating pre-randomisation screening using a placebo (thus assuming that non-compliance is a general characteristic of the person rather that the drug) to weed out non-compliant individuals (32, 44). Thus, such trials typically involve a homogeneous population without concomitant medical conditions who have demonstrated a degree of compliance prior to recruitment into the trial and are therefore likely to remain in the study (42). As such, patients recruited into an explanatory trial may not be wholly representative of the general population of interest.

3.3.2. Effectiveness trials

The purpose of a "pragmatic" or "effectiveness" trial, on the other hand, is to ascertain whether a treatment which has previously been shown to be efficacious actually works in real life conditions. Pragmatic trials are therefore said to measure the effectiveness of

an intervention and to inform choices between treatments (50). Effectiveness is also known as "proof of practice" (51) or "use-effectiveness" (52) because it measures the benefit observed under the usual conditions of use. The participants of a pragmatic trial will therefore tend to be more representative of the entire intended target population, and as such may demonstrate varying degrees of compliance and be taking concomitant medications for other medical conditions (42).

Effectiveness of treatment incorporates not only the true efficacy of treatment but also the compliance with the treatment, and thus regardless of the actual efficacy of treatment, effectiveness of treatment will decrease as compliance decreases (47). The pragmatic estimate inherently includes the effect of the acceptability to the patient, which is regarded as an inseparable part of the evaluation of treatment effectiveness (53).

Schwartz and Lellouch (54) were the first to make a "distinction" between explanatory and pragmatic trials. Schwartz explains that treatments are administered within a particular context (mode of administration, side effects (and the consequential treatments), diet, auxiliary care, associated treatment). An explanatory approach requires that these contextual factors are balanced across randomised groups in order that the only difference between groups is the exact treatment received. In a pragmatic trial, however, optimal levels of these factors are set separately for each treatment group in order to inform a choice between two treatment modalities, such that these contextual factors become part of the treatment package.

Although trials may be broadly classified as pragmatic or explanatory, efficacy and effectiveness are not discrete, independent quantities, but instead exist on a continuum, precluding any sensible dichotomisation (48). Rarely will a trial adopt an entirely pragmatic or explanatory approach; for example, even the most pragmatic trial design

will typically include collection of more outcome data than in usual practice (55). The degree to which a trial demonstrates evidence of efficacy or effectiveness should therefore be measured on a sliding scale; there are potential characteristics that help to distinguish pragmatic and explanatory elements of a trial and a number of tools exist for determining where on this continuum a trial lies.

Gartlehner et al (48) proposed the use of a simple, validated tool composed of seven criteria to distinguish explanatory and pragmatic trials, namely the population and care facilities, eligibility criteria, principal outcomes, treatment modality and study duration, adverse event assessment, sample size considerations and analysis population. Alternatively, the PRECIS (PRagmatic-Explanatory Continuum Indicator Summary) wheel is based on ten domains (outcomes, patient adherence, practitioner adherence, primary analysis, eligibility criteria, flexibility of experimental/control interventions, practitioner expertise with interventions, blinding) and is intended to aid researchers in designing trials in line with their purpose (55).

Trial design features therefore determine the balance between efficacy and effectiveness, ideally at a point which provides satisfactory internal and external validity relative to the trial aims. However, a satisfactory level of interval validity is required in order to achieve external validity (48); thus all trials require at least the basic elements of trial design, such as randomisation and allocation concealment, to ensure validity of its conclusions. As such, the setting of even the most pragmatic of trials will inevitably be artificial in some respects.

3.4.Manifestations of treatment nonadherence in trials

One such inevitable difference between a patient's clinical care in a trial and in practice is the requirement to follow the trial protocol, which contains specific details on patient

recruitment, treatment regimens, assessment, outcomes, follow up and analysis, providing a clear reference to permit accountability and reproducibility of the trial's design, conduct and analysis. Compared with a regular clinical setting, an added dimension to the problem of nonadherence arises in clinical trials because of the required adherence of both trial participants and treatment providers to the potentially stringent trial protocol.

Thus subjects may never receive their allocated treatment, may receive the intended treatment but comply only partially, or may change during the course of treatment from one treatment arm to another. As in general clinical care, premature termination of (or non-persistence with) treatment is a common problem in trials of long term patient-administered treatments, potentially exacerbated by overly demanding treatment regimens, treatment side effects, deterioration in patient's condition or lack of perceived benefit (56). Similarly, partial non-compliance occurs when patients deviate from randomised treatment protocol, for example forgetting to take occasional doses or taking drug holidays. Thus trial participant nonadherence may take the form of premature participant withdrawal from the study (commonly referred to as withdrawal of consent or participant discontinuation of the study) or discontinuation of treatment, either permanently (often referred to as withdrawal from treatment) or temporarily (treatment interruptions). Other forms of nonadherence include failing to follow treatment protocol as regards timing or dosage of randomised treatment, or failing to initiate allocated treatment at all.

Furthermore, in a trial setting, "adherence" comprises more than the patient simply following a prescribed treatment regimen or therapeutic intervention; it also includes cooperation on the part of treatment providers and assessors in following the procedures as specified in the treatment protocol. Clinician nonadherence typically

occurs when, based on their own experience and judgement (relating to side effects or efficacy of treatments) or in response to changes in the patient's state or other intervening complications, clinicians may make changes to a patient's prescription from those stipulated by the treatment protocol (57). For example, treatment providers may deviate from protocol when administering or prescribing treatment, by changing the type, dose or schedule of drug prescribed or failing to deliver treatment according to the procedure specified in the protocol (for example, during a surgical operation).

"Crossover" between randomised treatment arms, also known as "contamination" or "intrusion" (58), may occur for reasons entirely unrelated to a patient's prognosis or condition (for example, as a result of dispensing errors). More commonly, patient preference or interference on the part of the clinician, due to differential expectation of treatments, may result in immediate switches to the alternative treatment arm following randomisation. However, it may be necessary to permit switches to alternative treatment arm immediately following randomisation for practical reasons (for example, if the randomised intervention is not feasible). Treatment crossover to the alternative trial treatment, or non-trial treatments, during the course of the trial may also be permitted or encouraged in the treatment protocol for medical reasons, for example following disease progression, intolerable side effects or lack of efficacy. Indeed time to treatment withdrawal may well be a primary outcome of interest (53, 59).

Contamination may occur simply as a result of being recruited into a trial. For example, receipt of the extra attention or information about the aim of a trial may induce a certain therapeutic effect, especially in behavioural intervention trials; Simon (60) describes patients in the placebo arm of a depression trial who experienced benefits simply as a result of the extra contact with trial staff. Contamination may also result

when active treatment is readily available to those in the control group, for example, over the counter at a pharmacy (11).

The pattern and nature of departures from randomised treatment exhibited in a particular trial will depend on the form of randomised treatments being compared. Thus, participant compliance may be all-or-nothing (when treatment is given at a single time point or when treatment changes only occur prior to initiation of treatment, rather than during the course of treatment) or partial, fluctuating in a time-dependent manner (for example, when patients occasionally forget to take a dose of medication or take longer drug holidays). Levels of adherence may be constant or change quantitatively over the course of the trial. Treatment changes may take a variety of forms; treatment switches may take place in one treatment arm only (for example, when control patients are given the option switch to the experimental treatment on disease progression) or may be very complex (for example, when all patients are permitted to switch to the alternative trial treatment or external non-trial treatments).

Furthermore, in a trial setting, adherence is required, not only in relation to the provision of treatment but all other aspects of trial design, including the follow up assessments necessary to permit outcome determination and appropriate inclusion of the patient in the trial analysis. As such, in addition to previously discussed manifestations of treatment nonadherence, nonadherence to the follow up protocol in a trial setting may also occur if a patient fails to provide or receive outcome assessments. Missing outcome data may occur for technical or practical reasons or because of loss to follow up. Patient withdrawal from treatment often coincides with premature withdrawal from the trial, as outcome data are often collected at the time of treatment delivery (for example, when patients receive treatment and provide follow up information at the same clinic visit).

3.4.1. Deviation from follow up protocol: missing outcome data

Manifestation of nonadherence in a trial setting therefore generally takes the form of deviation from randomised treatment regimen, as defined in the treatment protocol (thus referred to herein as "deviation from treatment protocol") or absence of outcome data (or "deviation from follow up protocol").

Missing outcome data may be indicative of undetected nonadherence to treatment assignment in a trial, because failure to adhere to assigned treatment protocol is often associated with subsequent failure to provide follow up information. Therefore treatment protocol deviations may also be manifested as (or masked by) absence of outcome data, which is generally caused by withdrawal of patient consent, loss to follow up (LTFU) or missing assessments. Patients may become unavailable for follow up or withdraw their consent to taking part in the trial for reasons related or unrelated to their condition or treatment. For example, a patient's condition may improve (leading to a perception that they no longer require treatment) or worsen (such that they believe the treatment is detrimental) or they may simply move away for non-medical reasons.

Outcome data can be assumed to be **missing completely at random (MCAR)** if the reason for missingness is entirely unrelated to the patient's characteristics or outcome at the time of drop out, for example when a patient's medical notes go missing. In such a case, the missing outcomes can be considered as a random selection from all trial participants' outcomes and therefore their exclusion from analysis will not distort trial conclusions.

If, on the other hand, missingness is related to certain factors which have been observed and recorded prior to drop out (for example, if the patient's condition improves or deteriorates significantly), the outcome data are said to be **missing at**

random (MAR) for all such patients; the assumed dependence between the missing and observed outcome data in patients with these characteristics can be used to predict or appropriately account for these missing outcomes.

When missingness is related to some *unobserved* event occurring following an individual's final follow up (for example, a sudden unexpected disease relapse), data are said to be **missing not at random (MNAR)**. The probability of bias resulting when such data are excluded from analysis cannot be assessed or ruled out, because of a lack of information on the reason for missingness; as such, the missing outcome data cannot be modelled or predicted without risk of bias.

Access to information on the reasons for missing data is therefore vital in order to determine how best to approach the analysis of missing outcomes, because without explicit records on the reasons for missingness, it is not possible for analysts to test whether data are MAR or MCAR. However, as previously discussed, given that outcome determination in a trial setting often occurs at the same time as receipt of treatment, discontinuation of treatment typically coincides with loss to follow up; this is particularly problematic given that treatment discontinuation is typically related to a patient's clinical condition (60).

Missing outcome data may therefore cause bias, even in the case of an analysis carried out according to randomisation (an intention to treat, ITT) analysis (40). A typical **complete case (or "respondent-based") analysis** which excludes patients with missing outcomes, or censors at the point of termination of follow up, will be unbiased only if the reason for missing outcome is unrelated (entirely or within prognostic factor group) to the risk of the outcome at the point of loss to follow up (in other words, MCAR or MAR respectively), such that the average prognosis of those without follow up information does not differ between randomised groups (60). This is typically

unlikely to be the case, as loss to follow up is often related to a patient's state; furthermore, it is usually difficult to verify whether or not it is true, given that missing outcome determination typically coincides with complete loss to follow up relating to all other patient factors. As such, the strict definition of an ITT analysis requires complete follow up on all patients (61).

Thus any exclusions, whether due to poor compliance or lack of missing data will potentially put into jeopardy the balance provided by randomisation: even if it is possible to demonstrate that groups remain balanced in terms of measured prognostic factors following such exclusions, this will not be sufficient to show balance on *all* factors. Even when rules for exclusion of participants are stated in advance (in statistical analysis plan or protocol), this does not preclude bias occurring, because no account is taken of the possible association between the reason for exclusion and outcome.

3.5.Complications in analysis

Nonadherence to any aspect of the trial protocol will present a challenge when analysing data. The chosen method of analysis is crucial to the interpretation of results from a trial, especially when dealing with deviation from protocol. In particular, deviation from treatment protocol will impact on the interpretation of trial analyses because the underlying assignment mechanism (randomisation), which forms the basis for unbiased hypothesis testing, no longer reflects the actual treatment received in some capacity.

The decision on how to approach analysis in the presence of such deviations is therefore crucial to the interpretation of trial results. If analysis is naively carried out according to treatment received rather than randomised allocation, the advantages of randomisation are lost such that the trial results and conclusions are potentially

rendered invalid. However, if all deviations are ignored in analysis and analysis is carried out according to randomisation, inference can be made only on the effectiveness of the treatment *policy* or *prescription*, rather than on the *biological* efficacy of treatment actually received. The most common alternative approaches to analysis are considered in the next section.

3.6.ITT, PP and AT analyses

Analysis in the face of treatment protocol deviations may therefore take a number of forms.

Analysis according to randomisation, in other words according to how the patient was intended to be treated (hence the term "**intention to treat**", ITT), ignores any such deviations, such that patients are simply analysed according to their randomised allocation regardless of whether they received or deviated from this allocation. In retaining randomised allocations, an ITT analysis maintains the balance afforded by randomisation, thus preventing selection bias and assuring a sound basis for statistical hypothesis testing. In the presence of any deviation from assigned treatment, however, the interpretation of such an analysis is limited to an assessment only of the effectiveness of treatment policy or of the treatment prescription, rather than a causal estimate of treatment received.

When the efficacy of treatment received is also of interest, alternatives to ITT are required, as inclusion of non-adherent participants in ITT analysis generally diminishes the estimated treatment effect and thereby resulting in a biased assessment of treatment efficacy (61). Analysts must however be mindful of the potential for bias resulting when analysing according to anything other than randomised allocations. A simple but statistically naïve method of analysis may involve analysis of patients in their

randomised group only if (or during the period during which) they follow their randomised allocation, thus excluding or censoring (terminating the time at which the participant was included in the analysis as being "at risk" of the event of interest at the point of deviation from the treatment protocol, which is possible when undertaking a survival analysis for an outcome which is time to some event) patients who deviate from randomised treatment protocol (hence the term "**per protocol**" (PP) analysis).

Exclusion of participants from analysis in this way affects both the internal and external validity of a trial (62). A patient's ability or willingness to persevere with treatment is highly likely to be correlated with their condition and other lifestyle factors; indeed, non-compliant patients have been shown to have worse prognosis than compliers in their respective randomised group, even when administered with placebo (9). As such, by excluding some definition of "non-compliant" participants, a PP analysis will affect the generalisability of a trial, as those who persevere with treatment protocol represent a non-random sample of the original group of trial participants.

Furthermore, such an analysis is likely to introduce selection bias and thus also affect the internal validity of a trial, because the various treatment protocols being compared will present different challenges to adherence. This is especially likely when the definition of "compliance" varies between treatment arms, reflecting the different adherence requirements of the treatment packages. As such, the residual compliant subgroups of each randomised group are unlikely to be comparable. Therefore, unless it can be demonstrated that the average prognosis of those who deviate from treatment protocol does not differ between randomised groups, a PP analysis will likely upset the balance provided by randomisation, casting doubt on the validity of its conclusions (60). However, given that those intermediate confounding factors which influence a patient's compliance status as well as their prognosis (and hence outcome) typically remain

unmeasured (and may arguably even be unmeasurable), it is often impossible to compare the profiles of these factors between groups. Therefore, even if the baseline characteristics, and rates and reasons for treatment withdrawals or changes, appear relatively similar between treatment arms, it is not possible to ascertain whether the compliant subgroups remain balanced with respect to unmeasured prognostic factors. The results of any such PP analysis are therefore highly likely to be unreliable because of these hidden confounding or selection effects (63).

Exclusion of patients due to deviation from treatment protocol reduces the statistical power of a study by reducing the sample available for analysis. A variation on (but one even more flawed than) PP analysis used to overcome this problem of reduced power is analysis according to treatment received (or an "as treated" analysis). **As treated** (AT) analyses, whereby patients are analysed according to the (predominant) treatment received, are never likely to be valid as randomisation is disregarded entirely (57).

3.6.1. Healthy user bias

This (so called "healthy user") bias associated with compliance analyses is supported by evidence demonstrating that those with better compliance behaviour tend to have better clinical prognosis that those with poor compliance, regardless of treatment received (e.g. active or placebo treatment). This effect has been demonstrated in varying clinical setting with disparate drug regimens, independently of whether or not there is an apparent clinical effect of drug (or of a drug/compliance interaction), suggesting generalisability of this phenomenon. (64) For example, in the Coronary Drug Project Research Group (65), non-compliers in the placebo group experienced nearly double the mortality rate compared to compliers (28% versus 15%, respectively), and in the Beta Blocker Heart Attack Trial, poor adherers had an increased risk of death whether they were taking active treatment (OR=3.1) or placebo (OR=2.5) (66).

Epstein (64) discusses potential reasons for this apparent relationship between compliance and outcome: first, the psychological impact of complying with medical instruction may actually enhance the wellbeing of the patient, either solely because of the placebo effect (expectation that the drug will work leads to improvement in a patient's condition) or because of the mediating effect of patients' changes in habits or actions which result from the positive feelings towards taking treatment (for example, if such positivity leads patients to alter other lifestyle habits which in turn influence their outcome). Secondly, the relationship between compliance and outcome may be spurious: either because of the method in which compliance information is collected (for example, if those faring better are more likely to provide compliance data or be labelled as "compliant") or if a patient's ability or likelihood to comply is determined directly by their prognosis (either physically or psychologically) or their innate personality traits (which in turn affect their outcome), and thus the apparent relationship between compliance and outcome merely reflects the underlying link between prognosis and outcome. Finally, the relationship between compliance and outcome evident in both treatment and placebo groups may be caused by different mechanisms – for example, in the active treatment group, this association may be due to the true effect of drug, whereas in the placebo group, other factors (such as those listed above) may have improved compliant patients' outcomes – though it may seem unlikely that these different mechanisms would lead to similar results between non/compliers in both randomised groups.

Thus, although it may tempting to base any validity of such a comparison on assessment of any differences in baseline factors between compliers and non-compliers, an apparent similarity in baseline prognostic factors is not sufficient to justify such an assumption. Indeed, analysis exploring the relationship of baseline prognostic factors

with compliance in a number of trials demonstrates that adjustment for important prognostic factors often fails to explain for the variation in treatment effect attributable to compliance features (65, 67, 68).

3.7.Analysis beyond ITT

Analysis according to ITT provides an unbiased estimate of treatment effectiveness, but only in terms of the conditions under which the trial was conducted and only if complete outcome data are collected on all randomised patients. As such, there may be may be a number of reasons why it may be necessary to consider analysis methods beyond ITT.

The choice of trial analysis should be determined according to the aim of the study and the ultimate research questions of interest. In many trials, the effectiveness of treatment will be of interest to ascertain whether the intervention being studied actually works in practice. Indeed, analysis according to randomisation is generally regarded as the most appropriate for primary trial analyses, as it mirrors the trial design and maintains all the advantages of randomisation (69).

However, it has long been recognised that trial results will inevitably be somewhat artificial; Bradford Hill in 11th edition of Principles of Medical Statistics wrote: "at its best … a trial shows what can be accomplished with a medicine under careful observation and certain restricted conditions. The same results will not invariably or necessarily be observed when the medicine passes into general use; but the trial has at the least provided background knowledge which the physician can adapt to the individual patient." Furthermore, ITT analysis will only be appropriate if the level of acceptability of treatment in a trial is generalisable to its future use in clinical practice: if non-acceptability in the trial setting is due to uncertainty about randomisation or

consent to trial procedures, then any problems regarding patient acceptability of prescribed treatment in the trial will not be relevant to clinical practice (53).

As such, it could be argued that ITT analyses from even the most pragmatic of trials may not even adequately estimate effectiveness, and thus such analyses are not providing answers to anyone's questions. This argument may be extended to suggest that instead a trial should provide an estimate of true efficacy, undiluted by the issues and complications presented by trial conditions, which can then be adapted to mirror alternative and more realistic projected situations. Indeed, it is arguable that efficacy is more useful to medical decisions than effectiveness; thus trials should be designed and analysed to provide causal estimates to inform clinical judgment (52). As such, it may arguably useful to collect and report information on the levels of treatment adherence in all trials, in order to facilitate accurate interpretation of the trial results and potentially to enable compliance-adjusted analyses.

Trials designed with explanatory aims tend to impose conditions with the aim of minimising adherence problems, for example by recruiting only patients who have demonstrated a commitment to compliance (identified perhaps using a pre-randomisation placebo screening period) and by specifying restrictions on treatment changes in the treatment protocol. However, the prevalence of treatment nonadherence observed in treatment practice will inevitably be carried into even the strictest explanatory trial setting. In addition, some trial features which aim to maximise validity of results may in fact impact negatively on adherence. For example, in the case of a blinded trial, if a participant thinks there is a chance they may be on control treatment, they may be less inclined to persevere, particularly if they have previously received the control treatment without success; in practice, however a patient may be more likely to persist with a novel treatment, given the associated expectation of their *known*

treatment. Therefore analysis of even the strictest explanatory trial will likely need to feature treatment protocol deviations that must be adjusted for in order to accurately estimate treatment efficacy.

Similarly, a trial may necessarily have a pragmatic design for practical, ethical or recruitment considerations, but the causal estimate of treatment may nevertheless be of primary interest. For example, treatment switches between randomised treatments may be permitted or encouraged in the protocol for reasons such as inefficacy (for example, on disease progression), toxicity or other adverse effects. Although such treatment changes are not technically deviations from treatment protocol, switches from original randomised treatment will dilute the ITT treatment effect as an estimate of treatment efficacy, because of merged treatment experience of the randomised groups. Therefore, depending on the research question, it may well be of interest to factor out such treatment changes in a causal analysis.

3.7.1. Differing research perspectives

As such, even when ITT is considered the primary analysis, there may be interest in secondary causal analyses, especially because of the typical mix of explanatory and pragmatic traits within a single trial, thus warranting several different analyses (70). Indeed, Godwin et al (42) advocate the collection of compliance information in both pragmatic and explanatory trials, but for different reasons: in an explanatory trial, compliance may be expected to be high, given the use of a homogenous population likely to comply with treatment and particular efforts typically often employed to maximise adherence to treatment protocol. If compliance data are collected and a high level of compliance is observed, the results of the ITT analysis will closely estimate the efficacy of treatment within that patient population. In a pragmatic trial, it is important to collect compliance data, in order to ascertain whether an intervention will be adhered

to in practice; if not, even a highly efficacious treatment may be rendered ineffective. However Godwin suggests that efforts to encourage adherence in a pragmatic trial should not exceed what would be given in usual clinical practice, in order to ensure a realistic level of compliance among the trial participants and thus externally valid results.

Therefore, even when ITT is considered the primary analysis, it is often appropriate to at least measure and report on, if not adjust for, compliance. The chosen method of analysis will depend entirely on the focus of the particular research questions of interest. The focus of the research question will in turn depend on the perspective of the party interested in the causal estimate.

Patient perspective

Adjustment for treatment deviation is important, not least for motivated **patients** who, when offered a new treatment will primarily want to know the long term and unadulterated treatment effect of treatment taken as directed, rather than the population average distorted by nonadherence. This will be especially important from a patient's point of view if the treatment changes or withdrawals are due to patient choice rather than side effects (71). Indeed, such interest in treatment efficacy is becoming increasingly important, as patient decisions based on informed choice in health care become more widespread (72).

Public health planning

The compliance rates or patterns observed in a trial are not guaranteed to reflect those in general practice. First, trial participants recruited into a trial may differ from the general target population, in obvious and less obvious ways. Recruitment of homogeneous patient groups according to strict eligibility criteria makes trial results less generalisable, but so too might the subconscious motivation provided by inclusion in

the trial. Even when trials participants' characteristics appear representative of their patient group, publication of positive trial results may increase patient's confidence in the drug when the trial is complete, thus impacting positively on compliance rates in general medical practice, compared with possible suspicion of trial patients aware of the unproven treatment efficacy (73). Indeed, if the treatment is widely publicised as a promising treatment, in practice patients may be encouraged and motivated to persevere such that the benefit seen in practice mirrors that of the true efficacy, rather than effectiveness, of the treatment in the original trial (74). In such cases, it would be of interest to adjust for these various compliance rates in order to obtain a more realistic estimate of treatment effectiveness in practice.

Furthermore, the policy of treatment change implemented in a trial (for example, the factors determining treatment changes) may differ from that used in practice. Although treatment switches will be necessary in clinical practice, the specific standard or non-trial second-line treatments that patients switch to in the trial setting may be obsolete by the time the trial results are published. An ITT treatment effect will relate only to the patterns and types of treatments used in the trial, which may be different to those expected in practice, adding to the argument for additional efficacy analyses to supplement ITT analysis.

Likewise, results that are generalisable to one community or setting may be less so in others with different resources or policies; as such, for the purpose of **public health planning**, it may be of interest to policy makers to obtain adjusted trial treatment effects assessing efficacy in a different setting which exhibits different patterns of compliance (73).

3.7.2. Anticonservative ITT estimation

Furthermore, although it is generally accepted that ITT provides a conservative estimate of treatment effect (by providing an estimate that is potentially closer to the null effect than the truth, thereby ensuring the true treatment effect is not exaggerated), this is not always true. In the case of the most common trial design, which aims to ascertain whether one treatment is superior to another and is thus called a superiority trial, a conservative estimate would mean that the treatment effect (e.g. the estimate of the difference in the effect of treatments) would be smaller than in truth, and thus is not exaggerating the true treatment effect. However depending on the pattern of deviations from treatment protocol, this may or may not be the case. For example, if the pattern of treatment changes makes the treatment experience of the two randomised treatment arms more similar than they would be if everyone had followed the treatment protocol (for example, if some patients switched to the alternative trial treatment), then in ignoring these treatment departures, the ITT treatment effect estimate would reflect this greater similarity in the treatment experience and would indeed provide a conservative estimate. However, if the pattern of treatment changes in the trial results in an exaggerated difference between the treatment experience of the randomised arms, the ITT estimate will be biased away from the null, providing an anticonservative estimate of treatment efficacy. This would occur if patients from the more efficacious treatment arm were able to switch to a non-trial treatment which was in fact even more efficacious, thus making the observed difference between treatment arms even greater than it would be if this type of switch had not occurred. The direction of bias may be difficult to predict particularly if numerous options are available to patients, for example when patients receive second-line treatments on cancer progression.

ITT is generally accepted as providing a conservative assessment of treatment efficacy, as non-compliance on the whole leads to greater similarity between treatment arms than originally intended (for example, when treatment switches occur between trial treatments or when premature discontinuations of treatment occur in both treatment arms) and as such, the inclusion of non-compliers leads to a diminished treatment effect compared to the true efficacy of one treatment relative to the other. However, there are a number of analysis scenarios for which this diluted treatment effect renders ITT an anti-conservative method of analysis.

Equivalence trials

When a trial is designed to demonstrate **equivalence** between two treatments, or non-inferiority of a certain treatment (as opposed to the standard superiority trial, designed to demonstrate that one treatment is superior than its comparator), the merging of treatment experiences of the different randomised treatment arms resulting from the inclusion of non-compliers in analysis will lessen the observed difference between treatment outcomes, thus increasing the probability that the null hypothesis (of a difference between treatments) will be rejected; in other words, ITT increases the probability of falsely concluding equivalence (or non-inferiority) between treatments.

Harms analyses

For similar reasons, the analysis of **harms** outcomes using an ITT approach is not recommended, as a diminished treatment effect in a harms analysis will be anti-conservative, making a harmful treatment more likely to be accepted as harmless. As stated by Lewis (75), "a pure ITT approach to the analysis of safety simply adds to the risk of failing to identify potential safety problems, and is therefore never advocated". Instead, the ICH GCP guidelines (61, 76) suggest that analysis of harms data should be

according to treatment received, in other words all participants who received at least one dose of a treatment should be included in that treatment group for harms analyses, though there is no consensus on this issue (62).

In particular, ITT analysis may distort the risk of harms that may be expected in compliant individuals. By averaging across all patterns of compliance in a trial, an ITT analysis will underestimate harms in patients who exhibit certain incorrect dosing patterns (for example, when drug holidays cause drug resistance or rebound or recurrent first-dose effects) while overestimating risk in the majority of patients who take the drug correctly (caused by a minority of patients who seriously deviate from treatment prescription) (20). If variable dosing patterns can cause serious harm, then disregard of non-compliance when considering the causes of such adverse events may lead to an incorrect conclusion regarding the safety of the intervention when taken as prescribed.

3.7.3. Statistical issues

There are numerous other statistical reasons why compliance information should be collected in order to allow analyses beyond ITT.

Power

The dilution effect of non-compliance in an ITT analysis affects the power of superiority trial analyses to a surprising degree. For example, a trial where only half of trial recruits were fully compliant may need up to five times as many to be recruited as for a trial with complete compliance (11) to achieve necessary power. As such, projections of anticipated compliance rates should be accounted for in sample size calculations; likewise, however, the **potential role of compliance in diminishing power of the ITT analysis** should also be investigated: if adherence to treatment

protocol is assumed to be high and is therefore not measured as part of the trial, the lack of an observed treatment benefit may simply be due to the dilution of the true treatment effect caused by poor adherence to the randomised treatment rather than due to true inefficacy of treatment (11). Adjusting for compliance may provide more realistic confidence interval limits for an observed treatment effect (68).

Meta-analysis

Another argument for reporting adherence information in trial publications is the need to consider variation in adherence rates across trials in a systematic review. Without consideration of the degree of compliance expected in a trial, the true efficacy of treatment may be masked by incorrect treatment-taking patterns; thus consideration of compliance in statistical analyses may also help to explain variations in treatment effects over time or between different participant subgroups or indeed between trials in a **meta-analysis** (MA) (73). Differences in compliance rates between trial settings will cause heterogeneity between ITT analyses, as ITT estimates mix the true efficacy with a measure of patient acceptability and compliance. Therefore, given that undiluted estimates of true efficacy are likely to be more constant across trials than ITT results, estimation of undiluted treatment efficacy may therefore help to reduce variability between trial results in a MA (77). Therefore, even when compliance adjustment is not relevant for trial aims, it is nevertheless important for trialists to report on the extent of non-compliance in their trial, using clinically relevant measures of compliance, such that potential meta-analysts can interpret any heterogeneity with respect to reported levels of compliance.

3.7.4. Drug development

Despite the prevalence of adherence problems in practice and the impact of adherence on treatment effectiveness, adherence to treatment has largely been ignored in the development of new treatments (22). This has been blamed on the unreliability of compliance measurements typically used in trials (for example, pill counts and biochemical measurements of drug, both of which may give distorted indications of longterm adherence) and the suspicion on the part of regulatory bodies regarding bias that may be introduced with any causal analyses that do not mirror randomised allocation. Indeed, Urquhart (22) argues that trial statisticians' reliance on ITT has fuelled this problem further. In addition, it may not be regarded in the interest of pharmaceutical companies to disclose adherence information regarding with their products, for fear of highlighting patient non-acceptance of, or difficulties with, treatment.

It is imperative, however, for compliance to be considered when policy makers are choosing between treatment options. A treatment found to be efficacious in strict trial conditions may not necessarily demonstrate treatment effectiveness under usual conditions of care. Thus one treatment may be deemed preferable to an alternative simply on account of patient acceptance, for example because of less frequent dosing, less invasive administration methods or combinations preparations (reducing the number of separate medicines taken) (78). Indeed, Urquhart (22) discusses how modern drug development rarely leads to a first ever treatment for previously untreatable conditions, but instead tends to identify treatments that may serve as alternatives to already available treatment modalities, increasing the importance of considering qualities such as patient acceptability and adherence.

Drug labelling

Adjustment for treatment changes is particularly pertinent as regards **drug labelling**, which is informed primarily from clinical trial data. Given the evidence for drug underdosing in general practice among patients with chronic diseases, it is not unlikely that drug trials will also suffer from such underdosing, leading to a diluted average ITT treatment effect relative to the full-dose effect, known as the 'dilution effect' (22). Drug labelling based on ITT results from clinical trials may therefore mislead those patients who intend to follow the treatment course as prescribed as to the likely drug outcomes (15, 22).

Cost-effectiveness analyses

Factors relating to adherence are also particularly pertinent for public health planning, given their impact on cost of treatment in actual practice. Nonadherence impacts on health care economics by its association with various related costs, for example, those associated with initial drug acquisition and utilisation of health care resources to deal with consequences of nonadherence; indeed, costs resulting from nonadherence may outweigh those related directly to the original treatment provision. For example, reduced efficacy of a treatment following nonadherence to the prescribed regimen may incur substantial additional costs due to requirements for extra clinical consultations, monitoring and testing, and hospitalisation; treatment costs may also increase because of unnecessary increases in dosage or switching to alternative (usually more expensive) drugs. Thus, a more expensive treatment may work out cheaper than its alternatives simply because of fewer adherence problems. As such, compliance should always be considered in health economic evaluations of drug cost-effectiveness (78), for example by exploring the impact of changes in non-compliance rates on trial cost-effectiveness analyses (69).

Pharmacokinetic and pharmacodynamic modelling

Compliance data may also be useful in pharmacokinetic and pharmacodynamic modelling. Pharmacokinetics (PK) may be simply defined as what the body does to the drug (in terms of the extent and rates of absorption, distribution, metabolism and excretion) as opposed to pharmacodynamics (PD) which may be defined as what the drug does to the body. PK/PD modelling combines these two disciplines to model the time course of the drug concentration and action in the body, for example through the use of dose-response curves. In providing exact information on both timing and quantities of doses, MEMS provide the required input parameters for PK/PD modelling, whereby the dosages of drug taken into the body over time are related to the drug concentration and action in the body. Availability of MEMS data therefore avoids the need to rely on the unrealistic and untestable assumption that all patients necessarily followed the treatment schedule according to prescription (22, 37). Furthermore, information on when doses were ingested along with biochemical measures of drug absorption allows assessment of potential variation between individuals in associations between these factors (23).

Pharmacogenetic modelling

Compliance information is also relevant for **pharmacogenetic** (PG) modelling. Genetic testing may be used to inform personalised treatment choices for patients by determining for a particular individual the appropriate dose of treatment (as in (79)), or by predicting the likelihood of treatment efficacy (as in (80)) or the risk of adverse drug reactions or toxicity (as in (81, 82)). Information on treatment discontinuations and compliance may be input into a PG model to identify patients for whom a certain treatment is particular efficacious and who therefore may be encouraged to persevere with treatment despite unpleasant side effects. Alternatively, PG modelling may use the

genetic profile of patients to identify those more able to tolerate adverse drug reactions, thus increasing their benefit-harm ratio and warranting prescription of otherwise overly toxic drugs (83).

3.8.Introduction to causal estimation

If trials are carried out perfectly, with perfect randomisation and blinding, full follow up and complete adherence on the part of all patients and treatment administrators, analysis according to randomisation will provide an unbiased estimate of the causal effect of treatment as received compared to control. However, as discussed in this chapter, complications often arise in trials as a result of nonadherence to treatment or follow up protocols, leading to changes to treatment or incomplete data. Before considering, in subsequent chapters, some examples of real-life trials exhibiting such compliance issues, this chapter will conclude with an explanation of why such deviations may cause problems when aiming to estimate efficacy of treatment.

In particular, this chapter ends with a discussion of the definition and difference between associational and causal effects, and how causal effects are most easily defined using a potential outcomes (counterfactual) framework. A brief introduction to the use of causal diagrams, which provide a pictorial aid to clarify the relationship between the treatment, outcome and other covariates associated with the particular causal scenario, will help to clarify the issues of confounding and selection bias and reveal how naïve methods (such as PP or AT analyses) introduce such biases.

Initially, therefore, it is necessary to provide an introduction to causal estimation and the framework within which such parameters may be estimated, beginning with the definition of a cause and an explanation of the difference between associational and causal inference.

3.8.1. Definition of a cause

The notion of cause dates back at least as far as Aristotle, who defined four types of causes (the material cause (that *out of which* the thing is made), the formal cause (that *into which* the thing is made), the final cause (that *for which* the thing is made) and the efficient cause (that *which makes* the thing), the last of these being that most relevant to statistical inference, while Locke in 1690 defined both the cause ("that which produces any simple or complex idea") and its effect ("that which is produced") (84).

There are two necessary conditions for the definition and estimation of the effects produced by a certain cause (logically referred to as "causal effects").

Firstly, the effect of a particular cause can only be meaningfully defined in relative terms to another cause. In other words, it takes at least two causes (or two versions or levels of a cause) to define an effect; thus in experimental studies, the treatment under study must always be compared to a relevant control condition, such that "experiments without control conditions are simply not experiments". Indeed, stating that "A causes B" inherently implies a comparison of the effect of A on B to some condition not involving A (84).

Secondly, the key notion that distinguishes a "cause" from an attribute or characteristic is the *potential* for all units in the population of interest to be exposed to *all* levels of the cause being compared. In other words, before a unit has been assigned a certain level of the cause, it must be technically possible to define and observe, in principle, every level a of the causal factor A under consideration; no outcome $Y_i(a)$ can be *a priori* "counterfactual" for any individual i (85). Thus Holland (84) explains how statements regarding the effect of an individual's traits (such as sex, race, eye colour) can only describe *observational* associations rather than be given *causal* interpretation. To

understand why this is so, it is necessary to consider the fundamental difference between causal and associational analyses.

3.8.2. Distinction between causal and associational analysis

In order to understand the difference between associational and causal analyses, it is vital to appreciate the difference between the underlying causal model and the process of observation (84). Thus consider a comparison of exposure to experimental $(A = 1)$ versus control treatment $(A = 0)$ on outcome Y. An individual's observed outcome Y_{Ai} may differ from their underlying unconfounded *potential outcome* under each potential treatment assignment (Y_{0i} or Y_{1i}), as their observed outcome may be influenced by factors other than just treatment received. Therefore, the association between treatment received and *observed* outcome may be contaminated by selection and confounding factors and thus may not reflect the true underlying causal relationship between treatment received and true *potential* outcome, the observed data (A_i, Y_{Ai}) may therefore differ from the underlying causal variables (A_i, Y_{0i}, Y_{1i}).

As such, in providing information only on the observed association between variables, results from standard statistical analysis methods (for example, regression or stratification methodologies) can only be interpreted in terms of descriptive statistics rather than providing any evidence of causality (84).

Such methods are used to estimate population parameters from which study samples are selected, and thus may be used to provide information on the observed relationship between variables by considering their joint distribution. For example, if $P(Y = y, A = a)$ denotes the proportion of individuals i in the population for which $Y_i = y$ and $A_i = a$, parameters estimated from this joint distribution simply describe the *observed*

relationship between variables at a single point in time; for example, the conditional distribution of Y given A

$$P(Y = y|A = a) = \frac{P(Y = y, A = a)}{P(A = a)}$$

describes how the distribution of Y changes with A. A typical parameter from this distribution may be obtained by regressing Y on A, providing the conditional expectation (or average) of Y given A, $E(Y|A = a)$, i.e. the expected value of Y given a specific value of A.

In contrast, statements regarding causality cannot be defined from a joint distribution alone. This is due to the potential bias arising when comparing different treatment effects observed in distinct sections of the population, as underlying inherent differences between individuals in each treatment group may distort the comparison. The definitions of, and reasons for, these forms of bias are most easily depicted and defined through the use of causal diagrams.

3.9.Causal diagrams

Causal diagrams are a form of directed acyclic graph providing a visual representation of causal or associational relationships between variables, first used by Wright (1921) to convey the direction of the causal relationship in equation modelling.

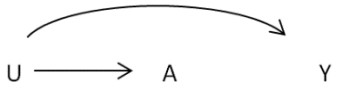
A directed acyclic graph (DAG) is made up of nodes (indicating measured or unmeasured variables) and directed edges (arrows between nodes). A DAG is called a causal DAG when the arrows have direct causal interpretation (i.e. arrows represent structural rather than merely associational relationships) and all common causes of each pair of variables are included on the graph. Causal DAGs (cDAGs) are necessarily acyclic (because a cause cannot affect itself, which would be implied from a cyclic

graph) and directed (because causal effects are by definition directional, *from* cause *to* effect).

Wright demonstrated how a causal diagram could be used to clarify the fact that the equation $\mathrm{Y} = \beta\mathrm{A} + \mathrm{c}$, intended to indicate the strength of the relationship between A (exposure or cause) and Y (outcome), could not meaningfully be rearranged to suggest that Y influences A i.e. $\mathrm{A} = (\mathrm{Y} - \mathrm{c})/\beta$, given that the causal diagram accompanying such an equation indicates the direction of the causal relationship with an arrow from A to Y.

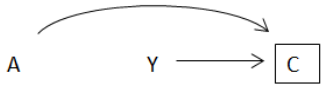
Hernan (86) summarises the theory underlying cDAGs, summarising the three causal structures by which apparent associations may be produced between treatment actually received (adopted treatment A) and outcome (Y):

i) treatment received (A) and outcome (Y) may share a common cause (U), leading to potential apparent association even when neither A causes Y (or vice versa) (**common cause**) (as per figure below)



ii) treatment received (A) and outcome (Y) may have a common effect (C), such that they will be conditionally associated when the association measure is calculated within levels of the common effect C (**conditioning on common effects**) as per figure below, where square around C indicates conditioning on certain value of C. In fact, the conditional association between A and Y may occur more generally, within strata of a common

effect C of two *other* variables – one of which is either A (or a cause of A) and the other which is either Y (or a cause of Y).



iii) the treatment received (A) may in truth cause outcome Y (cause and effect), thus the apparent association can be attributed entirely to the causal effect of A on Y.

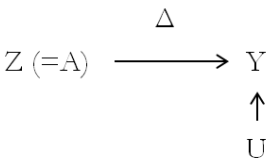
Note that these three causal structures do not include the arising of association simply by chance, because, in contrast to causal associations, such chance associations will diminish with increasing sample size.

Thus evidence of an association between two variables does not necessarily indicate a true causal relationship; instead apparent associations may arise due to a common cause or conditioning on common effects of the variables. Furthermore, categorisation into these three causal structures provides a basis on which to distinguish between the types of bias that arise when *associational* parameters are interpreted as *causal* measures, in particular differentiating between "confounding" and "selection" biases: the term "confounding bias" typically arises when exposure and outcome share a common cause (structure i), whereas "selection bias" occurs when analysis is conditional on common effects of exposure and outcome (structure ii).

Perfect randomised trial

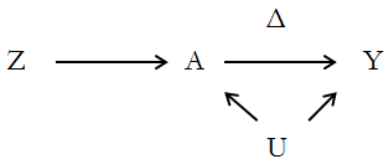
In the case of a perfect randomised trial, with no confounding of treatment assignment or nonadherence to assigned treatment, successful blinding and complete follow up, an

individual's actual treatment received (A) perfectly reflects their randomisation allocation (Z), which is assigned independently of all underlying prognostic factors (U). Thus, given that A and Y share no common causes and no conditioning is made on their common effects, the association measure (Δ) between A and Y is not affected by bias and can be interpreted as the causal effect measure.



Imperfect adherence to randomised treatment

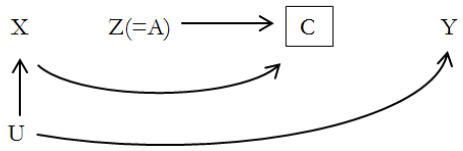
However, in the case of a randomised trial with imperfect adherence to treatment assignment, analysis according to adopted treatment (A) is confounded if there are underlying prognostic factors (U) which influence both an individual's likelihood of taking treatment and their outcome, as A and Y share a common cause (as per structure i above).



Missing outcome data

Furthermore, selection bias may occur if the trial is subject to missing data, even with complete adherence to assigned treatment; a so-called "differential loss to follow up" bias occurs when missing outcome data typically occur for reasons related to both prognosis and treatment (86). For example, side effects associated with treatment may

lead to censoring (if patients are too ill to attend follow up visits), as represented by arrow from $Z(=A)$ to C (denoting conditioning on censoring status) in the figure below. Furthermore, a patient's underlying condition and prognosis (U) (mediated by their observed symptoms X) may affect their likelihood of follow up (as demonstrated by arrows from U to X, and from X to C). Thus, the association between A and U via C means that, even when treatment does not affect outcome (evident by the lack of an arrow from A to Y), when analysis is conditional on follow up (C), an apparent association between treatment and outcome may result, arising solely because C is a common effect of a cause of the outcome (the underlying condition and prognosis, U) and of the treatment received (A).



Therefore, even in the case of full adherence to treatment assignment, trials may still be subject to selection bias because of differential loss to follow up. A similar selection bias occurs in the case of a per protocol (PP) analysis of a trial with imperfect compliance, where selection into the analysis set is determined according to compliance status (denoted again by C in the diagram above), if such compliance is related to any underlying prognostic factors (U) as well as being related to the assigned treatment protocol ($Z \neq A$). This selection bias introduces a non-causal association between Z and Y such that the PP analysis cannot be granted causal interpretation.

An as treated (AT) analysis attempts to deal with the loss of power associated with PP analyses, by allocating patients entirely according to their adherence rather than

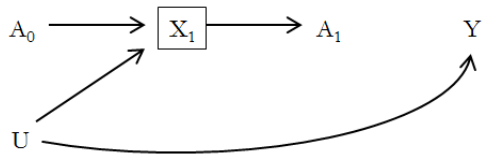
randomisation. Even in the case of purely random nonadherence, this analysis removes any balance afforded by randomisation, and therefore reflects an entirely observational comparison, given that randomised allocation is ignored entirely. As such, the simple A-Y association measure will be confounded in the same way as for observational studies, if underlying reasons for exposure to treatment are related in any way to prognosis (i.e. if treatment exposure (A) and outcome (Y) share common causes).

PP and AT analyses are therefore only valid under a simplistic exchangeability assumption, namely that those who do and do not comply are comparable, such that comparison between these groups is not subject to confounding or selection effects; only then will exclusion (or censoring) of those who deviate from randomised treatment (in the case of PP), or analysis according to treatment receipt (in the case of AT), not introduce bias. In other words, without appropriate adjustment for selection bias or confounding, PP and AT implicitly assume that non-compliance occurs completely at random (87). However, as demonstrated by the cDAG for a trial with imperfect compliance, given the likelihood of some *unmeasured* and *unknown* confounders, it will not be possible to account for the so-called "back-door" path (between U, A and Y) when estimating the causal effect (of A on Y) without invoking additional assumptions.

3.9.1. Problems associated with hidden bias or time-dependent confounding

Ten Have (88) distinguishes between two forms of (selection) bias, overt and hidden, which are attributed to observed and unobserved confounders, respectively. When aiming to estimate causal effects, standard statistical methods such as covariate adjustment in a regression model or propensity score analysis may be used to adjust for overt bias, given availability of known confounders, but these methods cannot address hidden bias.

Furthermore, standard adjustment for measured confounders may introduce bias if these confounders are time-dependent rather than fixed baseline measures. In the case of dynamic treatment regimens, common in epidemiological studies and trials where treatment is continuously being reassessed and varied according to evolving patient response, time-varying factors such as disease severity and adverse events will not only be affected by previous treatment received, but may also influence future treatment decisions and patient outcome. As such, these **time-dependent confounders** act simultaneously as confounders and intermediate (mediator) variables and cannot be adjusted for unbiasedly using standard methods. The reason for this (as demonstrated in the cDAG below) is that, in adjusting for confounding, selection bias is introduced.



By way of explanation, consider the trial scenario indicated in the cDAG above, where initial treatment (A_0) impacts on symptoms (X_1) which are used to decide on subsequent treatment (A_1). Unbiased estimation of the effect of overall treatment ($A = A_0 + A_1$) requires adjustment for X_1 to remove confounding bias in estimating the effect of A_1 (caused by common cause U of A_1 and Y), but in doing so, stratification (or regression, depending on the functional form of the covariates X_1) introduces selection bias for the estimation of the effect of A_0 (due to conditioning on X_1, which is a common effect of A_0 and U). Thus, standard methods to adjust for confounding factors affecting treatment decisions removes confounding at the expense of introducing selection bias (86).

Thus, given the inability of standard methods to adjust for unmeasured (hidden) or time-dependent confounding likely to be present in both trials and observational studies, alternative methods are required to provide unbiased estimation of causal contrasts of interest. Furthermore, estimation of treatment effectiveness may be subject to selection bias due to missing outcome data, such that the most basic ITT analysis (according to randomisation) may be rendered invalid. As such, causal methodology will be relevant in both pragmatic and efficacy trials subject to any form of deviation from treatment or follow up protocols.

3.10. Discussion

The complications introduced by nonadherence in the general medical setting (as described in Chapter 2) are compounded further in trials by the added dimension of necessary adherence by both the patient and the treatment provider to the treatment regimen specified in the trial protocol. Even the most pragmatic of trials require clear specification of the treatment practices followed in the trial to allow reproducibility, provide transparency and provide a basis for treatment evidence. Thus, the need to follow even a relatively flexible treatment protocol will necessarily limit the level of true concordance achievable in a trial setting (44). Serebruany (35) suggests that an element of concordance can be incorporated into a clinical trial context simply by the doctor explaining the possible adverse reactions that the patient may experience, in order to minimise treatment discontinuation, but concludes that the "impact of the concordance model on clinical trials is unclear and requires further consideration". Trialists would need to give detailed explanation of how 'concordance' affected treatment protocol, which would require detailed collection and succinct summaries of treatment patterns. Incorporating the spirit of concordance, with its evolutionary and particularly

unpredictable nature, is therefore even more challenging in a trial setting than in general clinical practice.

In this chapter, it has become evident that the nature and extent of nonadherence manifested in a trial will impact on trial conclusions, as a result of complications in statistical analysis when patients' outcome data are excluded or missing due to ineligibility, missing assessments or deviations from treatment protocol. However, even in trials of a predominantly pragmatic nature, there may be numerous reasons to consider the extent, and impact on analysis, of such deviations from protocol. Before considering methods to appropriately account for nonadherence, in the next chapter, we will consider real-life examples of some such trial scenarios where it may be of interest to look beyond ITT for answers to alternative research questions of interest beyond the effect of treatment as randomised.

3.11. Recommendations

Although ITT analysis is an important part of any trial, there is a potential danger that the spirit of ITT could be interpreted as an indication that collecting or reporting data on the degree of adherence to treatment protocol is unimportant. However even when there is no intention to formally investigate the relationship between treatment uptake and outcome, reporting information on the degree of intervention received is arguably important in order to assess the degree to which the intervention is even reaching the targeted population. Otherwise it may not possible to judge whether an unfavourable observed effect of treatment, for example, may be in part due to non-receipt rather than ineffectiveness of treatment, and if due to non-receipt of intervention, whether steps could be taken to improve uptake of treatment to potentially enhance treatment effect. Indeed, an ITT analysis can only provide information on the effectiveness of the

intervention as it was implemented in the trial (for example, the trial policies on treatment changes) and according to the level of adherence observed in the trial. In reality, the implementation of the intervention in clinical practice may differ, and adherence rates may improve (or diminish) directly because of dissemination of results from the trial itself and resultant changes in expectation of the intervention amongst users and clinicians. Thus even the interpretation of the effectiveness of an intervention may be limited from an ITT analysis. Collecting information on adherence and treatment changes that occurred in the trial could potentially allow subsequent statistical investigation into the impact of changes to administration of treatment or adherence rates on the treatment effect (89). Even in the case of so-called "large, simple trials" where emphasis lies on collection of limited data from a very large number of participants (90), collection of adherence information from just a small random subsample of patients may add to the clinical interpretation of trial results without greatly increasing the burden on participants or trial staff.

4. Expository trials demonstrating nonadherence

The previous chapter demonstrated how the problem of nonadherence in general clinical practice translates into problems in randomised trials, causing complications for analysis and interpretation of trial results. Before future chapters explore in more detail how deviation from protocol affects analysis and subsequent interpretation of conclusions drawn from trial data, and potential methods to appropriately account for such deviations, this chapter will present a number of expository trials demonstrating issues relating to various forms of departure from randomised treatment, thus providing examples of when it may be appropriate to consider estimation methods beyond ITT.

Trials will be described in term of their interventions, patient populations and the particular forms of treatment deviations that occurred, with a focus on how these treatment deviations relate to the trial research question and their subsequent impact on trial conduct, data collection and analysis. A number of these trials will be used later in

the thesis to illustrate the application of particular statistical methods to adjust for such nonadherence.

4.1.Introduction

This exposition includes seven trials from six different clinical areas, including HIV, cancer, epilepsy, community infant care, wound care and (two trials in) coronary heart disease (CHD). Treatment duration varies; the majority of the trials had a longitudinal treatment period, either medium term (wound care and cancer) or long term (epilepsy, HIV, hypertension and CHD), but there is also an example of short term treatment (community care). There are examples of trials which include treatment changes that were necessary and expected and thus were incorporated into the trial protocol (HIV, epilepsy, CHD and cancer); in contrast, other trials experienced treatment changes that were unplanned and undesirable, either as a result of treatment switches requested by patients or administered by external clinical staff (because of a belief that one treatment was superior (wound care)), logistical difficulties in distributing treatment (community care) or nonadherence to treatment dose or schedule by participating patients (CHD). Most of the trials considered analysis that accounted for changes in *prescribed* treatments rather than adherence by randomised participants, as patient adherence data were generally not collected.

The majority of the trials reported primary outcomes that captured time to some event, which is not surprising given the longitudinal nature of most of the trial treatments, but two trials included binary outcomes (one of which was the original primary outcome (community care) and the other which was included as a simplification of the complicated primary time to event outcome (epilepsy). All but one (cancer) of the trials are complete and have previously been published.

4.2.Trials

4.2.1. Vitamin A

The simplest trial in terms of both the types of treatment deviation and primary outcome is the vitamin A trial, first published by Sommer et al (91). This was a cluster randomised trial carried out in 450 villages in rural Indonesia, half of which were randomly assigned to the treatment arm and the remaining were designated as control villages. Preschool children in the treatment arm villages were given two large oral doses of vitamin A six months apart, and children in control villages did not receive any treatment. The primary outcome was mortality rate between four and twelve months of age.

Deviation from randomised treatment in the vitamin A trial did not occur as a result of a lack of cooperation on the part of the participant; instead non-receipt of trial drug occurred as a direct result of the failure of the trial drug distribution system to reach a substantial proportion of those randomised to receive treatment: 20% (2419/12,094) of those children randomised to the treatment arm did not receive vitamin A.

Table 2: Table I from (91)

Study group	Received treatment	Number of children	Number of deaths	Mortality (per 1000)
Control	-	11,588	74	6.4
Treatment	-	12,094	46	3.8
	Yes	9675	12	1.2
	No	2419	34	14.1

Information on those who did and did not receive treatment was captured, and thus it was possible to calculate the mortality rate among those allocated to treatment who did and did not actually receive vitamin A (1.2 per 1000 versus 14.1 per 1000 respectively) among those randomised to receive treatment. This initial comparison suggests that the

selection of patients in 'compliant' and 'non-compliant' groups is not simply a random process, and the reason why infants did not receive treatment (failure on the part of the distribution system) provides an explanation for this: Goetghebeur (92) points out that those infants who did not receive the treatment tended to be those living in a more difficult environment with less access to health resources. Thus the non-compliant subgroup had a higher baseline risk of mortality and were thus more likely to have died, irrespective of whether or not they received treatment.

This propensity for a more successful outcome among those who comply with treatment compared with those who don't comply with treatment does not, in itself, lead to a biased comparison between treatment groups. Bias arises only when differences between the two treatment groups, in terms of the mechanisms which lead to non-compliance, result in differential selection processes for compliant subgroups, such that the resultant compliant subgroups in the two treatment arms are not comparable (91).

However, this fact is also evident when one compares the mortality rate among the non-compliant subgroup of the vitamin A group to that observed in the group of infants who were randomised to not receive vitamin A. As children in the control arm were not administered with any placebo treatment, non-compliers in the active treatment received the same "intervention" (no treatment) as those in the control arm. Thus if one were to suppose that compliance to intervention was entirely random, the mortality rates between the control arm and non-compliers in the treatment arm would be expected to be broadly similar. However a comparison of these mortality rates demonstrates a mortality rate among non-compliant treatment arm infants (1.41%) which is more than twice the mortality rate of the overall control arm (0.64%). As

Sommer points out, this suggests that simply being able to access the vitamin A treatment strongly predicts the risk of mortality.

A simple ITT analysis, comparing the treatment groups as randomised, regardless of compliance with the randomised intervention, provides an estimate of the relative risk (RR, defined as the risk of death among the treatment arm divided by the risk of death in the control arm) of death, giving an indication of the so-called "programmatic effectiveness" of the vitamin A distribution. Using data from Table 2, the estimate of this relative risk is 0.38/0.64 = 0.59 (95% CI= 0.41, 0.86), thus implying that the vitamin A distribution programme used in the trial resulted in approximately a 40% reduction in childhood mortality.

However, given that a substantial proportion of the treatment arm did not receive Vitamin A simply because the trial treatment distribution system failed to reach them, causal analysis beyond ITT (such as that carried out by Sommer and Zeger) are necessary in order to estimate the biological action of vitamin A in preventing childhood mortality. This question of treatment efficacy is of particular interest in this case because, if vitamin A was proved to be efficacious, the treatment distribution method used in the trial would not be used in practice to provide supplementation to the Indonesian population. Instead a completely different method of circulation, such as fortification of a common daily food, would be used to reach the general population. As such, Sommer argues that by estimating the biologic efficacy of vitamin A supplementation, along with a likely rate of successful delivery of the actual method chosen for vitamin A distribution, the effectiveness of the chosen program could be determined; this would be a more useful measure than the effectiveness of the trial distribution program, which would not be used in practice.

Thus Sommer and Zeger (47) first proposed a method of analysis to estimate what they refer to as the 'biologic efficacy' of treatment, as opposed to the 'programmatic effectiveness' of treatment. As the use of a placebo was not permitted in this trial, it is not possible to ascertain which of the control children would have complied with placebo intervention. Such information is not, however, necessary for the method of analysis proposed by Sommer and Zeger, and the reason for this is that the compliant subgroups of the two randomised groups would not necessarily be comparable anyway. This is due to the fact that the compliance selection process is likely to differ between the two treatment arms, because the difference in treatment protocol between the two arms is likely to present a different challenge to adherence, thus resulting in compliant subgroups that have been selected by different processes and hence do not provide a valid or fair basis for comparison (47).

Instead Sommer and Zeger propose a comparison of outcomes among those who would have *complied with active treatment* if randomised to receive it. Although such a subgroup is directly observed in the vitamin A group, it is not possible to directly ascertain which patients would have complied with active treatment in the control arm, and thus it is necessary to compare the outcome among those observed to be compliant in the active treatment arm with an *inferred* control subgroup who would have complied with active treatment (47). The causal effect estimated using their proposed method (to be presented and discussed in the next chapter) was later referred to as the CACE estimate (complier average causal effect) and relates to the efficacy only among those who *would* comply with treatment.

4.2.2. CPD trial

The Coronary Drug Project (CDP) was a randomised, double-blind, placebo-controlled, six-arm trial that was carried out to evaluate the long term efficacy and safety of lipid-

influencing drugs in the secondary prevention of coronary heart disease (CHD) among middle aged males who had experienced a myocardial infarction within the previous three months (65). Participants were randomly assigned to the placebo arm (2789 men) or to an active treatment arm (each of which contained approximately 1100 men), and were followed up at four-monthly visits for a minimum (maximum) of 5 years, or until death. This summary of the trial focuses on only one active treatment arm (clofibrate) versus placebo. The primary outcome was all-cause five-year mortality rate.

Nonadherence to treatment protocol in the CDP occurred as a result of patients failing to take the correct dosage of their randomised treatment. Following the initial titration phase, patients in each treatment arm were required to take nine tablets daily (three tablets three times per day). Clinicians recorded any changes to prescribed medication at each four-monthly visit, along with their assessment of the patient's adherence to the drug regimen since their previous clinic visit. This assessment was made by counting or estimating how many capsules had been returned by the patient at each visit, along with information obtained from a discussion with the patient about whether they had experienced any problems in remembering to take (or side effects resulting from taking) their medication. This information was used to calculate the cumulative percentage adherence over the first five years of follow up (or until death, if this occurred before the end of five years follow up). Patients were divided into 'compliant' and 'non-compliant' subgroups according to whether or not they took at least 80% of their prescribed medication.

The adherence data used here was necessarily determined by clinician from interview with patient and pill counts, neither of which are methods that are considered particularly reliable (15). The adherence data suggested that one third of the patients

(33.5% (357/1065) in the clofibrate group and 32.7% (882/2695) in the placebo group) failed to take at least 80% of their prescribed medication.

Overall clofibrate did not appear to affect mortality when compared to placebo: five-year mortality in the clofibrate group was 20.0% compared to 20.9% in the placebo group. Given this negative result and the relatively complicated patient-administered treatment (3 x 3 tablets daily) and high prevalence of non-compliance (defined as taking less than 80% of prescribed medication over the course of the first five years of treatment), the study investigators and external parties became interested in whether the ITT result may have masked a true effect of treatment among those who received clofibrate as intended.

Table 3 Mortality rates according to adherence level.

Note that this table only includes those patients for whom cumulative adherence rate was available.

	Treatment group			
	Clofibrate		Placebo	
	Number of patients	% mortality (SE)	Number of patients	% mortality (SE)
Adherence				
<80%	357	24.6 (2.3)	882	28.2 (1.5)
≥80%	708	15.0 (1.3)	1813	15.1 (0.8)
Total	1065	18.2 (1.2)	2695	19.4 (0.8)

An initial inspection of the difference in mortality rates between 'compliers' and 'non-compliers' in the clofibrate group seems to support this hypothesis: the mortality rate among 'compliers' (15.0%) was substantially lower than among 'non-compliers' (24.6%). However a similar examination of the placebo group demonstrates a similar pattern: the mortality rate among placebo 'compliers' (15.1%) was almost identical to that of clofibrate 'compliers' and was nearly half that among those who did not comply to their placebo medication schedule (28.2%). Bearing in mind that those who did and did not comply with their placebo medication schedule were receiving the same "treatment"

(assuming that any placebo effect associated with compliance to prescribed treatment would have a minimal effect on the objective outcome mortality), this would suggest that the behaviour of patients as regards compliance with medication seems to be strongly related to their baseline risk of mortality. It would appear that there are confounding factors associated with compliance that affect outcome, such that those who are able or willing to comply with taking either intervention as prescribed tend to be less likely to experience the event.

These simple summaries according to adherence levels immediately demonstrate the dangers of using a patient response factor such as adherence to select patients for inclusion in analysis. A number of comparisons between subgroups defined according to adherence can be made; for example, comparing the mortality rate of only those with good adherence in the two treatment arms (15.0% in the clofibrate arm versus 15.1% in the placebo arm), comparing the mortality rate of only those with poor adherence in the two treatment arms (24.6% in the clofibrate arm versus 28.2% in the placebo arm), comparing the mortality rate for good adherers in the clofibrate group (15.0%) against the placebo group as a whole (19.4%), or comparing the outcome among poor adherers in the clofibrate group (24.6%) against the placebo group as a whole (19.4%). Indeed the investigators conclude that "one can justify almost any conclusion, depending on the analysis chosen" (65).

The CDP investigators assessed whether adjustment for differences in baseline factors between those with 'good' and 'poor' adherence in the placebo group helped to explain the difference in mortality rates between these subgroups. Although poor adherers tended to have a somewhat higher prevalence of some of the baseline risk factors when compared with good adherers, adjustment for 40 of these variables explained only a small proportion of the difference between these subgroups: the adjusted five-year

mortality rates were 16.4% (compared with 15.1% unadjusted) for good adherers and 25.8% (28.2% unadjusted) poor adherers.

Thus it is evident that the available baseline data did not capture the characteristics which were both predictive of outcome and associated with adherence. The baseline factors that were measured in this trial reflected clinical rather than behavioural or social measures, and it is possible that clinical factors may be less useful in explaining compliance behaviour. Without knowledge of which factors simultaneously influence adherence and outcome, it is not possible to ascertain the selection processes which have resulted in the compliant subgroups in each treatment arm and thus whether or not these subgroups provide a fair basis for analysis. Thus this trial provides a straightforward example of why analysis according to subgroups defined by compliance behaviour is not valid.

4.2.3. MRC hypertension trial

The MRC hypertension trial (93) was a randomised single-blind placebo-controlled trial set up to assess the long term efficacy of two antihypertensive drugs (diuretic or beta-blocker) compared with placebo in prevention of cardiovascular (CV) events and mortality in elderly patients.

Given the long term nature of the treatment protocol, the need to monitor and control each patients' blood pressure, and the known possible side effects of the prescribed active treatments, the trial protocol permitted changes to treatment from that to which patients were originally allocated. These protocol-permitted treatment changes reflected what would typically occur in clinical practice.

The nature of the treatment change depended on the underlying cause: if a patient's blood pressure persistently exceeded their target level, the dose of their (active)

randomised drug would be increased, and if this also did not achieve acceptable blood pressure control, the patient would be prescribed an additional drug (initially the alternative active trial drug). If this too was unsuccessful, a treatment external to the trial would be prescribed. If their blood pressure continued to be unacceptably high, the patient was managed outside of the trial protocol, but information on prescribed treatment and outcomes continued to be collected.

If a patient was experiencing unacceptable side effects on active randomised treatment, the randomised treatment would cease to be prescribed and an alternative treatment (commonly the other active trial drug) may have been given instead.

Other less common treatment changes included discontinued prescription of the randomised treatment when it was felt by the prescribing clinician to be unnecessary, or withdrawal from the study by patients for reasons which may have been associated with risk (for example, dislike of the trial treatments or assessments) or unrelated to risk (for example, if a patient moved away).

Note that the treatment changes recorded in the MRC hypertension trial relate only to changes to prescribed treatment. Information on patient adherence to prescribed treatment was not recorded, and thus could not be factored into the analysis. It is worthy of note that patient nonadherence may have had a similar magnitude of causal effect as changes to prescribed medication, but it was not possible to investigate this here.

When analysing those who were lost to follow up, it was possible to obtain their mortality data by Office of National Statistics "flagging", but it was unknown what treatment changes occurred following withdrawal. Such "loss to treatment follow up" (so-called by White (68) as these patients were not lost to *event* follow up, as time and

cause of death was ascertainable for all patients through flagging) was not an insignificant factor: White states than 21% of person years, and 41% of deaths, occurred in subjects who were lost to treatment follow up (defined as the period following withdrawal or 190 days after a subjects' last follow up visit).

Patients randomised to receive beta-blocker were more often prescribed supplementary drugs than those randomised to diuretic treatment, and this group experienced significantly more withdrawals than the diuretic group, both for reasons related to side effects and because of inadequate blood pressure control. A larger proportion of the beta-blocker group also discontinued their randomised treatment: including those who were lost to follow up, almost two thirds of the beta-blocker groups stopped their randomised treatment compared to approximately half of the diuretic group (cumulative percentages of those who stopped their randomised treatment, including those lost to follow up, were equal to 63% in the beta-blocker group and 48% in the diuretic group).

The ITT analysis of the MRC hypertension trial appropriately addressed the primary question of interest, namely whether there was any difference in the effectiveness of the treatment policies of starting treatment with one randomised intervention, followed by any treatment changes that became necessary. The effectiveness of the active treatments in preventing stroke and other CV events was apparent from the ITT analysis, as the risk of these events was significantly reduced in the active treatment groups compared with the placebo groups. When the two active treatment groups were compared directly in an ITT analysis, it was unexpectedly found that the rate of coronary events (and of all CV events) was significantly lower in the diuretic group compared with the beta-blocker group.

In light of the frequency of treatment changes that occurred in this trial, and the relatively high proportion of follow up time that was spent not taking randomised

treatment, it was naturally of interest to the trial investigators to investigate whether, and how much of, the unexpected ITT difference between the active treatment groups was in fact be due to differential rates of treatment changes that occurred in the different randomised groups.

White adjusted for changes in prescribed treatment in this trial using both naïve (68) and randomisation-based (89) methods (to be introduced in the next chapter). These analyses demonstrate the need for an understanding of the subtleties of the particular disease or treatment being studied before one can confidently produce and interpret findings from a causal analysis. Effective communication between the statistical analysts and the clinical experts for a particular trial is of paramount importance to ensure such nuances are not overlooked in the analysis, and to prevent statistical anomalies or biased interpretation of trial results.

4.2.4. Concorde

The Concorde trial was designed to compare two treatment policies with zidovudine (ZDV) among patients with asymptotic HIV disease. ZDV was known to be efficacious in delaying death when treating symptomatic HIV disease, but it was unknown whether treatment with ZDV earlier in the disease process would delay the onset of AIDS-related complex (ARC) or AIDS, or whether it could be potentially harmful due to increased toxicity. Thus patients were randomised either to start ZDV immediately following randomisation (immediate group) or to delay treatment with ZDV until the onset of symptomatic HIV disease (ARC or AIDS). The trial was double-blind; patients in the deferred group were administered with placebo from randomisation until they presented with symptoms of the disease, at which point they started to receive active ZDV treatment (94).

The treatment protocol intended patients to receive blinded treatment (active ZDV in the immediate group or placebo in the deferred group) until the point of disease progression, at which point they were to be prescribed open label ZDV without being unblinded to their original allocation. However, there were deviations from this intended schedule for a number of reasons. Some patients experienced a delay in starting their active or placebo treatment (though this was usually a relatively short period of delay). Patients could choose to stop taking their blinded medication, and this occurred usually because of adverse events or for personal reasons. In these cases of early cessation of treatment, patients were not unblinded to their original treatment and they were prescribed with open label ZDV if and when they presented with disease progression. If patients developed clinical or immunological deterioration (before disease progression) they stopped their blinded treatment and usually started open label ZDV within a few days.

Major changes were made to the treatment protocol one year into the recruitment phase. Although the original protocol specified that patients in the deferred group should not start active ZDV treatment until the onset of ARC or AIDS, the protocol amendment permitted subjects in the deferred group to begin taking open label ZDV before the onset of ARC or AIDS if they persistently presented with low CD4 cell counts. Another change which affected both treatment groups was introduced as part of this amendment, namely that patients with low CD4 cell counts were to be treated with primary PCP prophylaxis. (PCP is an opportunistic infection that affects AIDS patients, and recurring attacks of PCP may be fatal (95).) These changes brought the treatment protocol in line with what had become regarded as best clinical practice since the protocol had been originally devised.

A substantial proportion of patients experienced treatment changes in this trial. Approximately one third of each group stopped blinded therapy for reasons other than clinical deterioration; another one third of patients stopped open label ZDV treatment. A higher proportion of deferred group patients (48%) than immediate group patients (37%) started open label ZDV and these treatments tended to be started earlier in the deferred group.

There were three related survival outcomes of primary interest: time to disease progression (ARC or AIDS) or death, time to AIDS or death, and time to death. The ITT results from this trial suggested a potential benefit of early treatment with ZDV in preventing disease progression (to ARC, AIDS) or death (HR (95% CI) = 0.89 (0.75, 1.05), no effect on prolonging time to progression to AIDS or death (HR (95% CI) = 1.01 (0.82, 1.24), and a possible detrimental effect on overall survival (HR (95% CI) = 1.26 (0.93, 1.70)), but none of these results were statistically significant.

However, the amendments to the treatment protocol were a cause for concern: although it was recognised that the treatment protocol amendment was necessary to update the trial to standards of best clinical practice, the decision to allow commencement of treatment in the deferred group to a time point earlier than originally intended led to complaints that this had resulted in an unfair assessment and a diluted effect estimate of treatment with ZDV. The ITT analysis of the trial appropriately assessed the comparison of the treatment policies as implemented in the trial, but given the change to original protocol, this did not assess the treatment policies as they were originally *intended* to be implemented in the trial.

Thus, when the ITT results of the Concorde trial were published, providing no statistically significant evidence of a clinical benefit of starting ZDV prior to presentation of symptomatic HIV infection, the question arose whether these

discouraging results could have been caused by a diluted effect of starting ZDV treatment early in the asymptomatic phase of HIV disease, as a result of the change of treatment protocol leading to earlier starting of ZDV among the deferred group than had originally been intended.

Premature cessation of blinded or open label treatment, whether for clinical or personal reasons, would be expected to occur naturally when ZDV treatment was prescribed as part of normal clinical care. Thus this aspect of deviation from the randomised course of treatment was not considered to distort the interpretation of the trial, and accordingly the causal analysis did not attempt to factor out these treatment changes.

White (94) used this example to demonstrate the need to sensibly decide on the scenario that one wishes to simulate when adjusting for treatment changes: this is a clinical issue which must be determined in light of which treatment changes are inherently intrinsic to the treatment and disease pathway, and which changes would be usefully factored out in order to answer certain research questions. Once the relevant research question of interest has been carefully specified, this will then logically (if not easily) translate into a model that would reflect this adjusted scenario.

For example, in this trial, White points out that it that it would not be interesting or informative to factor out cessation of treatment that occurred when patients experienced side effects, as this is an expected course of events. Also, given that it was already known that ZDV was efficacious in treating advanced disease, it would not be sensible to simulate a comparison between the immediate group as treated against an altered deferred group who never received ZDV treatment, as this would not enhance clinical knowledge beyond what is already known of ZDV efficacy.

Instead the desired alternative treatment scenario of interest is that which would have occurred under the original treatment protocol, so that patients in the deferred group would not have received treatment prior to disease progression or death, but other treatment changes would have been expected as observed in the trial (for example, cessation of treatment due to side effects). Thus it was of interest to adjust only for the earlier start of ZDV in the deferred group, which could have contributed to the observed non-significant results by diluting the difference between randomised groups. This was particularly likely given that decisions to start open-label ZDV treatment in this trial were very strongly related to prognosis (74).

Thus this analysis sought to answer a well-defined question: what would the results have been if no participant in the deferred treatment arm had started ZDV before progression, but other treatment changes (such as stopping an active treatment after a side effect) were as observed?

Following on from an initial naïve analysis of these trial data using time dependent treatment covariates to simulate an AT survival analysis (68), White instead advocates and demonstrates use of a randomisation-based survival model (to be introduced in the next chapter), which also relates the event time to a time-dependent covariate but which is based on groups as randomised (rather than on the biased basis of actual treatment received) (89).

4.2.5. PACIFICO

PACIFICO is a randomised two-arm trial comparing two forms of chemotherapy for fitter older patients with follicular lymphoma, in order to assess which treatment is optimal in terms of efficacy in controlling the spread of the disease balanced against the

toxicity to patients associated with these types of treatment. This trial is currently still recruiting patients.

The PACIFICO treatment protocol specifies that cycles of inpatient chemotherapy be delayed, and potentially ceased entirely, if patients experience excessive toxicity. This is a pertinent issue in this trial, given the aim of balancing efficacy with tolerability of treatment, and detailed instructions are provided in the protocol regarding these permitted changes to the treatment. Clinicians are required to record details of the dose and time of treatment cycles in the trial case report forms (CRFs) to ensure that they follow the treatment protocol in this regard. However any deviations from the instructions given in the protocol are likely to reflect differences that will occur in real clinical care, and thus are accepted as an integral part of the variation of treatment according to patient's needs and symptoms.

The trial chemotherapy also involves patient-administered home therapy. Patients are asked to record in a daily diary the date, time and doses of self-administered treatment, also noting any occurrences when the absorption of the treatment may have been compromised (for example, if vomiting occurred soon after treatment was taken), though this is not expected to occur often. Patients are generally expected to comply closely with the doses and schedule of prescribed home chemotherapy as far as possible as they are keen to halt spread of cancer; furthermore, they are a relatively fit and able group who would not be expected to experience the typical logistical or memory problems that may be associated with self-administered treatment by the elderly.

The treatment changes that are, however, of particular interest in terms of causal impact in this trial are those relating to alternative treatment choices made at the point of disease progression; interest in the impact of this pattern of treatment changes at this stage of disease is typical of late-stage cancer trials. When patients are being treated for a

cancer that may later progress or relapse, it is often an ethical requirement to permit in the trial protocol a switch to alternative treatments when their disease worsens, which may or may not include the other trial treatment. In PACIFICO, the treatment protocol recognises the need to offer and experiment with other treatments if a patient's disease progresses, and thus the date, time and nature of these alternative treatments are recorded.

The primary outcome in trials where such treatment changes are expected is usually chosen to be progression-free survival (PFS) or relapse-free survival (RFS), defined as the time from randomisation to progression (or relapse) or death, whichever occurs first. Such an outcome is not affected by switches at progression or relapse, as the switches occur after the event of interest and thus have no causal impact. Determining date of progression, however, is potentially complicated and subjective, and may not necessarily translate into a survival advantage.

The traditionally more common outcome of overall survival (OS) would, however, be affected by such treatment changes, as merging of the treatment experience in the two arms following progression or relapse is likely to cause diminished treatment effects on OS. Despite the fact that such treatment changes are often seen in clinical practice, there may be a number of reasons for interest in obtaining a clean estimate of overall survival (OS). OS is objective and is usually the most important outcome for the patient, as well as being of primary interest for health economics and policy decisions.

Debate regarding PFS versus OS

Analysis of OS is often confounded by uncontrolled post-progression (or relapse) therapy, which is increasingly common thanks to the growing number of active anti-cancer agents. This has led to the use of PFS as a surrogate outcome for OS, as

treatment switches generally occur after the point of progression and, as such, time until progression is less heavily affected by changes from randomised treatment. However the use of PFS as primary outcome has its own disadvantages.

Disease progression is a categorical outcome first developed by WHO to describe changes that occur to a tumour during therapy, rather than as a measure intended to infer meaningful clinical benefit (96). Although potentially useful in early phase trials as an indication of potential treatment response, the use of PFS in phase III trials is controversial for a number of reasons.

Sharma (83) highlights how the debate regarding the most appropriate clinical endpoints in oncology trials, and in particular whether trials should be allowed to base their conclusions on shorter-term outcomes like PFS rather than ultimate long term outcomes such as OS, was sparked by the controversial fast track approval of the bestselling cancer drug Avastin. The license for sale of this drug was granted by the FDA based on trial results which suggested PFS, but not OS, improvement associated with the drug; however, following two subsequent trials which demonstrated no such improvement and which furthermore suggested a greater side effects profile, the FDA later rescinded its approval for the use of Avastin in metastatic cancers because of concerns that the drug's side effects outweighed its benefits, while the European Medical Agency reaffirmed its approval. The use of PFS is therefore argued to simply provide a means of lowering of the bar for declaring efficacy of new drugs, which in turn offers little in the way of meaning for patients.

Although there are a number of advantages when using PFS in place of OS (as progression usually occurs well before death, thus requiring smaller numbers and shorter follow up times, and progression is not confounded by effects of subsequent follow up therapies), there are a number of limitations associated with PFS.

Firstly, the practical implications in defining PFS are not straightforward; the theoretical definition of PFS is the time from defined origin (such as date of randomisation) to tumour progression or death. However, in practice, progression is usually identified via some form of imaging assessment, involving some subjective judgment on the part of assessors (often their treating clinician), especially in the case of equivocal or uncertainties in interpretation of radiologic assessments. As such, there are a number of biases and sources of measurement variability that may arise in obtaining the date of progression (for example, when times of evaluation differs between the treatment arms or when the knowledge of treatment received in some way affects a patient's recorded response or progression status) (97).

In particular, given that progression is not as objective an outcome as mortality, there are multiple factors which may introduce variability or bias into the assessment of progression (97). Thus, in order to make determination of time of progression as objective as possible, it is necessary that patients undergo regular and frequent assessments, which must be balanced and determined exactly equally for both treatment arms, to prevent any biases arising due to timing or prompting for such assessments. In order to protect from likely bias that may be introduced when subjective decisions are made by assessors or clinicians when determining progression (especially when factors such as "clinical deterioration" are included in the definition), blinding of clinicians and assessors is of great importance; however, this is not always possible given the disparate forms of treatment (for example in terms of treatment schedules and modalities). Alternatively the assessment of progression may be made by an independent imaging review group.

There are other potential concerns when using PFS in place of OS as a primary outcome. Firstly, it has been demonstrated that PFS is not a reliable surrogate for OS in

a number of cancers; only two tumour types have evidence providing support for the surrogacy of PFS for OS (96). Indeed Booth provides biological reasons why a certain treatment may demonstrate advantages in terms of PFS without impacting on OS. Furthermore, PFS is more likely to be affected by missing outcome assessment data and informative censoring than OS, as definition of progression relies heavily on the availability of the necessary assessment outcome data.

Change to treatment protocol

Another potentially interesting aspect of this trial with regards to treatment changes involves an important change to treatment protocol that occurred partway through the recruitment period. During the course of the trial, concerns regarding excess life-threatening toxicity due to one component (FC) of the experimental treatment were raised as a result of external evidence from another trial. As such, a "lighter" regimen (FC-lite) was introduced for all future patients recruited to PACIFICO. This amendment to treatment protocol would likely affect the treatment comparison for both primary outcomes, as fewer cycles and/or reduced dose of FC treatment will impact both on efficacy and tolerability of this treatment. Thus it would be of interest to estimate the causal effect of treatment adjusting for this protocol change, in order to attempt to answer the original question of interest.

Thus, causal analyses for the PACIFICO trial would focus on two main questions: first, what the treatment efficacy would be if all patients had *not* received alternative therapies on disease progression, and secondly what the treatment effect would have been if all patients randomised to receive FC had received the original (rather than the lighter) dose as per the original trial protocol.

4.2.6. Honey

Robson et al (98) carried out an open label RCT comparing wound healing rates in 105 patients randomised to receive either conventional wound dressings or dressings treated with a medical grade honey. Treatment and follow up continued until wound healing up to a maximum of 24 weeks, and the primary outcome was time to wound healing. The aim of the study was to compare a medical grade honey with conventional treatments on the rates of wounds healing by secondary intention (rather than by direct intervention, such as surgery).

Despite its relatively straightforward treatment protocol, this trial was subject to a number of deviations from treatment protocol and complications in follow up, including switches between trial treatments, switches to (known and unknown) non-trial treatments, LTFU and death during the course of treatment (a competing risk).

These treatment changes occurred for a number of reasons. In cases of deterioration of the wound, it was ethically necessary to allow patients to receive more extreme forms of treatment and as such some trial patients ceased randomised treatment and started receiving antibiotics or received surgery or radiotherapy as appropriate. Although information on these non-trial treatments were recorded whenever possible, the change of patient care often resulted in missing follow up information. When patients withdrew from the trial prematurely, or were moved to another hospital, it is likely that they continued to receive some form of conventional treatment, but treatment and follow up data were not available.

Expectation about honey treatment also lead to treatment switches from randomised conventional arm to honey. There had been considerable amounts of publicity regarding the use of honey in wound care by the time the trial was opened, as honey

was not at the time available on NHS prescription. This expectation, but lack of availability of honey (except in a clinical trial setting such as this one where the honey was provided by the manufacturers), initially created a recruitment incentive for both patients and clinical staff. However, this anticipated efficacy of honey treatment, coupled with the unblinded nature of the trial, frequently lead to disappointment when patients were not allocated to receive honey.

Patients recruited to the trial had typically been unsuccessfully treated with various forms of conventional dressings for their offending wound prior to recruitment into the trial. As such, given their natural expectation of honey and likely frustration with previous conventional treatments, along with the logistical impossibility of blinding patients to their randomised allocation (as patients necessarily were able to see and feel the applied dressing), patients were prone to disappointment when they were not allocated to receive honey. This lead to premature discontinuation of the trial by those who could not muster sufficient interest to continue attending the more frequent than usual clinic visits required for trial follow up. And occasionally, from the more bold discontented patients randomised to conventional treatment, there was the request to switch to honey treatment from the outset, which was granted in one case.

Treatment switches also occurred because of decisions of clinical staff external to the trial who were keen for their patients to receive honey, especially those younger fitter patients who would be expected to make good progress with honey treatment.

In contrast, some patients believed that honey treatment was causing additional pain to their wound which, as a result of their unblinded treatment, lead to a request to discontinue honey treatment and switch to conventional treatment.

Death during the course of the trial resulted in irretrievably censored treatment and follow up data. Patients requiring care for wounds healing by secondary intention are typically relatively elderly and frail, and as such a number of the recruited patients died during the course of follow up. Given that death precludes any subsequent occurrence of the event and causes of death may be related to deterioration of the wound being treated by the randomised treatments (and thus censoring at the time of death would potentially be informative), this event is defined as a "competing risk" and introduces complications for analysis.

Missing outcome data often accompanied the treatment changes in this trial. When patients withdrew consent from further trial treatment and follow up, either because they were discontented with their treatment, progress or with the trial follow up requirements, it was not possible to obtain information on the rate of wound healing. When the course of treatment was altered because the worsening severity of their wound necessitated a more extreme treatment choice, or when the patient moved hospital, outcome follow up was not possible on a practical level.

Although it was ethically necessary to allow patients to receive alternative or more powerful treatment if they experience side effects or wound deterioration, the primary aim of this trial was to ascertain the biological efficacy of honey treatment when compared to the best standard care. As such, it is of interest to estimate the causal effect of honey treatment compared with conventional dressings in the absence of any treatment changes.

4.2.7. SANAD

The SANAD trial was an unblinded randomised comparison of a number of Standard And New Antiepileptic Drugs primarily among patients presenting with newly

diagnosed epilepsy (although some patients who had unsuccessfully been treated with an anti-epileptic drug (AED) previously (not including any of the trial treatments) or who had relapsed following withdrawal of treatment after a period of remission were also recruited). Patients were enrolled into one of two arms of the SANAD trial:

Arm A of the SANAD trial was designed to compare the efficacy and safety of five antiepileptic drugs (AEDs) (carbamazepine (CBZ), lamotrigine (LTG), gabapentin, oxcarbazepine and topiramate) in terms of seizure control, tolerability, quality of life (QoL) and health economic outcomes (99), and was intended for patients for whom the standard treatment CBZ was considered to be most appropriate (usually presenting with focal epilepsy).

Arm B of the SANAD trial compared (sodium) valproate (VPS), lamotrigine (LTG) and topiramate (TPM) (100), and was intended for those for whom the drug valproate was considered more appropriate (most of whom had generalised epilepsy). This multi-way randomisation is the first complication affecting both arms of this trial, simply because many of the causal methods which can be applied to investigate or adjust for treatment changes will struggle to contrast more than two groups.

The second complication of this trial design was the need to balance efficacy and safety: thus if a drug was shown to be superior in terms of tolerability (side effects), it was deemed necessary only to demonstrate non-inferiority of the treatment's efficacy in terms of seizure control. The need to employ methods to determine non-inferiority (or equivalence) is complicated when treatment deviations occur, because as discussed in Chapter 3, ITT analysis is anticonservative in this case. Thus there was a need for estimation methods beyond ITT and PP, as neither of these methods are unbiased in this setting.

The third complication of the SANAD trial was the lack of blinding. As mentioned in the discussion of the honey trial, knowledge of which treatment has been prescribed can have an impact of the expectations and motivation of patients and clinicians alike. AEDs are known to often cause unpleasant side effects; a lack of blinding may mean that patients or clinicians are consciously or subconsciously looking for emergence of such symptoms and are therefore more likely to act on these side effects than they would be if they did not know which treatment was being prescribed. Furthermore, the inevitable disappointment if a hoped-for treatment is not prescribed may affect a patient's morale and potentially impact on their general health. Both of these consequences of open label prescribing may lead to more frequent decisions to switch treatments than if treatment allocation had been masked.

Changes to prescribed treatment are common in epilepsy, because the clinician must tailor the prescription to meet the individual's needs. The optimum dose of each drug varies between individuals, and must provide a balance between maintaining seizure control but limiting side effects to a tolerable level. This is achieved by a titration phase, in which the dose of the originally prescribed drug is increased or decreased in gradations until a balance between efficacy and tolerability is achieved. If a drug is not acceptable to a patient following such a titration phase, it may be necessary for an additional AED to be prescribed (if, for example, the original drug is not controlling seizures) or for the original drug to be dropped entirely and replaced with a different AED (if, for example, the original drug is leading to unacceptable side effects even at a low dose).

Thus the treatment protocol in SANAD was chosen to be entirely pragmatic, mirroring the ethical requirements for treatment changes that are part of everyday clinical care. The treatment protocol did not specify the initial dose of each randomised treatment, or

the rate of titration to reach the optimal dose of each trial drug. Instead decisions regarding dose and choice of alternative treatments were left to the discretion of the treating clinician.

Treatment changes were further complicated by the typical need for a loading phase (when starting a new treatment) or a withdrawal phase (when stopping a treatment) in order to slowly increase or reduce prescribed treatment doses, rather than simply ceasing treatment straight from, or starting treatment at, a potentially optimum maintenance dose. This meant that when a treatment switch took place, there were often overlapping periods when two AEDs were taken simultaneously, with the duration of this overlap varying considerably between patients.

Another complication of the SANAD treatment experience was the fact that some patients were still taking other treatments at randomisation. These patients were usually being withdrawn from a non-trial AED which had failed to prevent seizures, and the sometime lengthy withdrawal phase of this drug therefore overlapped with prescription of the randomised treatment.

Thus patients in SANAD could experience a number of different treatment changes and complications during their trial follow up, including changes to prescribed treatment dose, complete withdrawal from randomised treatment, addition of other trial (or less commonly non-trial) treatments to aid seizure control, switching to another trial (or non-trial) treatment or a continuing short term prescription of another treatment still being taken at randomisation.

This trial is therefore similar in a number of ways to the MRC hypertension trial mentioned previously. As in the MRC trial, the SANAD treatment protocol recognised that changes to prescribed treatment would be necessary, with treatment failure (change

to prescribed treatment) generally occurring in this clinical area for two competing reasons: either as a result of inefficacy of treatment (indicated by inadequate seizure control, ISC) or intolerability of treatment (due to unacceptable adverse events, UAE).

The relevance of the issue of changes to prescribed treatment in epilepsy is borne out by the fact that one of the primary outcomes in SANAD was time to treatment failure (defined as the time from randomisation to cessation of randomised drug (whether due to ISC or UAE or both) or time to addition of a supplementary AED, whichever occurred first). ISC typically resulted in addition of an alternative treatment, most commonly one of the other trial AEDs, whereas UAEs led to cessation of randomised treatment and prescription of an alternative treatment.

Both of these causes of treatment failures are related to dose but with opposite directions of association. Dose is related not only to the efficacy with which a certain treatment will prevent further seizures, but also to the likelihood of adverse reactions: as dose increases, the efficacy of treatment is expected to increase, but so too will the likelihood of unpleasant side effects. Thus it was necessary to analyse time to treatment failure using methods allowing for competing risks: a blanket analysis of time to treatment failure, without regard for the underlying cause, could lead to a masking of treatment effects as a result of cancelling out of the opposite effects of treatment by the two competing causes of treatment failure (198).

The second primary outcome was time to 12 month remission (T12mR), defined as the time from randomisation to reaching a 12 month period free of seizures. This outcome is of great practical importance to patients, as one year's remission is required before a patient regains the right to their driver's license. However, as we shall see in future chapters, it is a rather complicated non-standard summary of repeated events data.

Time to first seizure (TFS) was one of a number of secondary outcomes. Although this outcome may initially seem a sensible and obvious choice for a more straightforward primary efficacy outcome (than T12mR), it does not necessarily provide a fair comparison of treatments, given that seizure control is going to be related to some degree to the initial dose of treatment. When clinicians prescribe a drug that is well known to them, they are likely to be more confident in increasing the treatment dose as part of the titration phase and thus may achieve higher doses of drug more quickly than when they are prescribing a newer drug with a less familiar side effect profile. Thus the comparison of TFS across treatments may be confounded to some degree with the familiarity of the drug, in favour of older treatments, as seizure control depends substantially on the initial drug dose.

Indeed any treatment comparison in terms of seizure control is likely to be influenced by this dosing issue; however, dose is likely to have a more direct impact on this outcome than on T12mR, which depends more on a sustained delay of seizures and is thus less affected by the initial dose.

Also, as in the MRC hypertension trial, the dose data collected in SANAD related only to prescribed treatment rather than patient compliance. Thus, any causal inference from these data is necessarily interpreted in terms of the explanatory effect of treatment that would be observed if there had been no prescribed changes to randomised treatment. Ideally this causal research question would also relate to the effect of compliance with randomised treatment schedule; indeed the effect of participant non-compliance with prescribed treatment, especially given the side effects associated with treatment, may have had a causal impact equal to or greater than that of actual treatment prescribed, but without the necessary adherence data, this theory cannot be investigated.

In the same way that mortality information was available (via flagging) when a patient had incomplete treatment follow up data as a result of premature discontinuation in the MRC hypertension trial, and patients in the Concorde had missing treatment information despite continuing to be followed up for outcome information, SANAD patients for whom treatment follow up data were missing or sparse nevertheless tended to be followed up for seizures. In such cases it was necessary to assume, as in the MRC trial, that no treatment changes had occurred unless they were specifically recorded.

When treatment changes occur prior to achieving remission, the assessment of each randomised AED is distorted and it becomes unclear which AED should be attributed with success or failure in achieving remission. It is therefore of interest to assess the impact of uninterrupted use of each AED in achieving remission; thus the causal question of interest in terms of estimating treatment efficacy in this trial is as follows: What is the relative benefit of each drug in achieving a minimum T12mR, in the absence of any changes to prescribed treatment other than changes of dose of randomised treatment? Potential analyses investigating this (and variations of this) causal research question are explored in more detail in Chapters 8 to 10.

4.2.8. Magnetic

In the MAGNETIC trial (101), over 500 children presenting to A&E with a severe asthmatic attack were randomised to receive three doses of magnesium or placebo 20 minutes apart. The primary outcome was asthma severity score at 60 minutes following randomisation (ASS60). Treatment deviations occurred in the timing and number of doses received: 24 (10%) in the magnesium group and 29 (12%) in the placebo group received doses that were more or less than 20 minutes apart (with a maximum margin of error of 5 minutes). Such deviations in terms of timing are arguably going to occur in clinical practice, given the pressured A&E setting, and thus it would not be sensible to

factor out such deviations. It may, however, be reasonable to expect greater control over the number of doses given in practice, and thus one could argue that it would be of clinical interest to estimate treatment efficacy that adjusts for the cases where less than three doses of treatment (as specified by the treatment protocol) were administered. There were seven (3%) of such cases in the magnesium arm and 12 (5%) in the placebo arms.

As will be discussed in the next chapter, adjustment for dose deviations could be implemented using instrumental variable (IV) regression, with randomisation as the exogenous variable; endogenous variables would be actual number of doses of active treatment received along with other baseline characteristics that are considered to correlate with number of doses received but which do not directly influence the outcome. Note that the number of active doses of treatment received would be set as zero for all patients in the placebo arm, regardless of their adherence to the treatment protocol. Thus the only patients for whom adjustment was required in the case were the seven patients in the magnesium arm who received less than three doses.

Unfortunately six of these seven patients withdrew from the trial prior to collection of the primary outcome. The likelihood of missing outcome data among patients who deviate from treatment protocol is not unusual. It does however mean that it is not possible to use any form of adjustment for deviation from dose protocol without imputing values for the missing outcome data (of those who received less than three doses in the magnesium arm).

The authors of the HTA report (102) presented three sensitivity analyses investigating the impact of missing outcome data generally (not just among those with treatment protocol deviations). The reasons for missing data were recorded, and were most often

due to withdrawal from trial or a serious adverse event. The primary ITT analysis assumed that the outcome data were missing at random.

The first sensitivity analysis imputed an ASS60 score of 0 if the recorded reason for missing data related to a good clinical status (for example, withdrawal due to improvement in condition, as was the case in three children), and imputed a value of 9 if withdrawal was due to poor clinical status (as was the case for one child). When the reason for missingness was unknown or was unlikely to be related to clinical status, the outcome data were left as missing. These imputations did not affect the statistical significance of the outcome.

The second sensitivity analysis involved multiple imputation to create five sets of imputed values, the average and corresponding variance of which were used to create the final data set and adjusted mean differences of ASS60 between the two treatment groups. Again the difference in group outcomes remained statistically significant.

Finally a joint modelling analysis of the time to withdrawal (T0, T20, T40 or T60, with those who did not dropout prior to T60 being censored at T60) and the longitudinal ASS data was performed to account for the non-ignorable nature of the missing outcomes. The results suggest that those who dropped out from the magnesium group did so because of an improvement in condition, whereas those who withdrew from the placebo were experiencing a worsening of scores prior to dropout.

Thus an informed adjustment for the reduced dose that was received by seven of the magnesium group patients would be likely to further increase the difference between the magnesium and placebo groups. However given that these patients are such a small percentage (3%) of the magnesium group, it is unlikely that such an adjustment would make any appreciable difference to the conclusions.

4.3.Discussion

This exposition of nonadherence in trials has demonstrated how treatment receipt may differ from randomised allocation for a number of reasons, including logistics, changes to trial protocol and (especially in the case of long term treatment regimens) patient and/or clinician choice. Changes to prescribed treatment may be necessary for clinical reasons, for example when patients are experiencing toxicity or adverse reactions. Furthermore, when faced with chronic disorders, patients themselves will often seek their own optimal course of treatment to fit in with their individual practical and medical needs. When clinicians are willing to work with patients to find this optimal course of treatment, concordance may be achieved between patient and clinician; however, this adaptive flexible approach to treatment introduces complications for randomised trials, in a similar way to all other treatment deviations from original randomised treatment. The potential for very different treatment experiences across patients within the same randomised group prompts the question of how any useful causal interpretation may be obtained from trials featuring such treatment deviations.

The approach to causal analysis therefore begins with identification of the particular causal research question of interest, which must differentiate treatment deviations or characteristics that are accepted as inherent within the treatment process (and thus should not be factored out in analysis) from those that do not reflect the intended treatment experience (and thus would usefully be adjusted for in analysis). Thus, by ignoring those treatment switches which correspond with the *intended* trial scenario (for example, clinically indicated changes in treatment prescription due to side effects) while adjusting for other deviations that do not reflect the scenario of interest (such as changes to trial protocol, patient nonadherence or logistical difficulties with treatment delivery), causal analysis may in fact provide estimation of treatment effects which are

conceptually intermediate between efficacy and effectiveness of treatment (73). As will be discussed further in future chapters, the approach taken to the causal question will therefore vary depending on the perspective of the interested party; for example, highly motivated patients' interpretation of which treatment changes should be accounted for (in order to provide a realistic efficacy estimate) may well differ from those of policy makers (perhaps interested in pragmatic effectiveness of treatment in a setting that varies somewhat from that presented in the trial), and therefore their corresponding causal questions will reflect different handling of treatment deviations. Once defined, it is then necessary to consider how to translate this question into causal contrasts prior to undertaking analysis; this issue will be discussed and demonstrated in future chapters.

4.4.Conclusion

The literature review of the previous chapter and description of expository trials in this chapter have demonstrated how treatment deviation may manifest itself in a variety of clinical trial settings, and that, despite the best efforts of trialists, the problem of deviation from treatment protocol is to some extent inevitable and will not be resolved; indeed changes to treatment are sometimes necessary features of the protocol.

Given that it is often important to estimate (some aspect of the) effect of treatment as received rather than simply as randomised, there is a need for appropriate alternative statistical methods which overcome the inadequacies of commonly employed methods such as PP or AT analyses, in order to provide unbiased estimates of such causal effects, defined by the causal question of interest. The next chapter therefore summarises a review of the statistical literature regarding potential methods to adjust for treatment deviations, using a number of the trials described in this chapter to illustrate methods.

5. Statistical methods to adjust for deviation from randomised treatment protocol

5.1.Introduction

Following the brief introduction to causal analysis and causal diagrams provided at the end of Chapter 3, this chapter will present an overview of causal methods to estimate treatment effects for a range of outcome measures and varying complexity of compliance patterns and associated patient-level covariates, with a particular focus on methods to analyse time to event (TTE) outcomes, as such outcomes are often

particularly relevant in trials with longitudinal treatment and follow up periods which are typically most likely to suffer from complications due to nonadherence or loss to follow up.

In keeping with the practical focus of this PhD, this overview of statistical methods is written with the intention of being a guide for researchers, directing them to the most appropriate causal method to use, according to their particular trial and compliance scenario. Thus technical details will be kept to a minimum; instead discussion will focus on application, highlighting important issues that researchers should consider as part of the causal estimation process. The key statistical methods discussed in this chapter will be illustrated in this and following chapters using a number of the expository trials introduced in Chapter 4.

5.2.Potential outcomes framework

Having considered in Chapter 3 the definition of a causal effect and how biases may be introduced using associational statistical methods due to unmeasured or time-varying confounding, it may now be apparent that causal effects can in fact only be defined in terms of comparing, for each individual, their potential outcomes resulting from each of the relevant treatment conditions while holding all other factors constant, such that any outcome difference can then be attributed solely to the differential exposures rather than to any potential selection or confounding effects. Such a within-patient contrast cannot of course ever be directly observed (in other words, it is a latent comparison) because only at most one of the conditions can be observed for a given individual at a particular time. Even crossover trials (which aim to estimate within-patient effect estimates by randomising patients to receive both the treatment and control treatment in sequential treatment periods separated by a suitable washout period) are unable to

provide true causal contrasts, as time (and thus potentially other factors) differs between the different treatment exposures periods.

5.2.1. Counterfactuals

When comparing potential outcomes for a given individual, all outcomes except that corresponding to the actual treatment received are said to be counterfactual (i.e. they are entirely hypothetical and cannot be observed) while that which is observed is said to be factual (85). The entire collection of all possible outcomes, both actual and counterfactual, for the target population is called the set of potential outcomes (103).

The notion that a causal relationship is necessarily described in terms of a within-patient difference between outcomes under counterfactual conditions (i.e. comparing the effect on outcome of one form of treatment with what would have happened under an alternative form of treatment) is not new. In 1748, the philosopher Hume stated "we may define a cause to be an object followed by another… where, if the first object had not been, the second never had existed."; in hypothesizing about what would have occurred under conditions contrary to those that occurred in reality, this statement is termed "counterfactual conditional" (103). However, it took another 175 years before this idea of comparing actual and counterfactual outcomes for the same individual was first formalised by Neyman (in 1923), who introduced the idea of multiple prospective outcomes under differing treatment conditions in the context of randomised agricultural experiments. His notation (relating to multiple plots and seed varieties) and terminology ("potential yield") has since been applied to randomised trials, providing a crucial basis for design, analysis and interpretation of studies investigating causal relationships (85).

5.2.2. Barriers to the use of causal estimation methods

Methods based on the potential outcomes (PO) framework are relatively new and continue to be developed, with particular methods varying according to (outcome and compliance) data and focus of the causal question. However a number of barriers, both historically and in present day, have undermined the use and acceptance of causal analysis techniques by mainstream statisticians.

One important proposed reason for the apparent resistance to wider use of causal estimation methods based on the PO framework is the requirement of underlying additional, but untestable, assumptions to allow inference on a causal rather than merely associational basis. This is due to the need to consider multiple, partially latent, counterfactual outcomes for each individual, which not only underpins the difference between associational and causal estimation, but also, in the words of Holland (84), is the "fundamental problem of causal inference", given that it is impossible to observe more than one outcome for each individual at a particular time point. The apparent impossibility of this task does not render causal analysis entirely futile, but instead requires that additional assumptions are made.

Causal assumptions are often treated with scepticism, despite often being more plausible than the underlying (but perhaps unappreciated) assumptions required for valid interpretation of naïve associational analyses according to treatment receipt (for example, PP analysis assumes the simplest but least plausible assumption that subjects who do and do not comply with their assigned treatment protocol are comparable with regard to prognosis). Pearl (104) argues that this reticence to accept causal, in contrast to associational, analyses may be due to the fact that assumptions which underlie associational analyses, such as those regarding probability density functions or conditional independence between variables, are somewhat less transparent than causal

assumptions regarding how one variable may cause another, and therefore the latter tend to be more open to counterargument.

Another potential barrier to the use of causal estimation is that, in contrast to those relating to associational relationships, assumptions associated with causal relationships cannot be expressed using usual statistical notation. Given that all causal analyses are conditional on a given set of assumptions beyond those of standard associational methods, clear notation is important in order to allow judgment of whether these assumptions are credible. For example, statistical dependence between outcome Y and adopted treatment A (expressed as $Y|A$) cannot be distinguished from causal dependence using standard statistical notation alone. Thus, supplemental explicit notation has been developed to facilitate differentiation between associational inferences and causal claims or assumptions, as well as to clarify the notion of counterfactuals, both in terms of outcomes and treatment receipt.

Such a causal "language", however, has developed only relatively recently. Although Neyman introduced the concepts relating to causal analysis in the 1920s, half a century passed before Rubin (105) and Holland (84) helped to reignite interest in causal techniques by clarifying the notation for counterfactual or potential outcomes which underlie causal inference. The importance of Neyman's innovative notation was that it allowed causal effects to be expressed independently of any particular probability model (85). For example, it became apparent that when carrying out causal analyses, it was possible to consider counterfactual outcomes as functions of treatment received and counterfactual treatment receipt variables in terms of randomisation assignment.

5.2.3. Notation and definition of typical causal estimands

Using such notation, the causal contrast can be expressed as a relative measure (for example, a relative risk or odds ratio for a binary outcome) or absolute difference (for example, difference in means of a continuous variable) (103). Assuming a simple scenario, where outcome or response (Y) is to be compared under two conditions (experimental ($A = 1$) versus control ($A = 0$) treatment) without interference between units (in other words, the effect of treatment on an individual does not vary according to treatment received by other individuals), the causal effect for a given individual i can be defined in terms of a contrast between their (potentially unobserved or counterfactual) responses under the two conditions, for example the absolute difference $Y_{1i} - Y_{0i}$ (106).

In other words, to describe a causal estimate, one needs not one (Y_i) but two (Y_{1i}, Y_{0i}) response variables for each individual, denoting their response had they been assigned to treatment or control respectively (84). Note that this notation can be extended to a situation where treatment is received repeatedly over time, such that $Y_{\bar{a}i}$ denotes the (potentially counterfactual) outcome that would have been observed if subject i had followed treatment regimen a for their entire follow up period, say up to time k ($0 \leq t \leq k$) i.e. $\bar{a} = \{a_1, a_2 \ldots a_k : 0 \leq t \leq k\}$. The causal contrast comparing continuous treatment to no treatment would then be written as $Y_{(\bar{a}=\bar{1})i} - Y_{(\bar{a}=\bar{0})i} = Y_{\bar{1}i} - Y_{\bar{0}i}$ (107).

Furthermore, given that the treatment a participant receives is assumed to be influenced by their randomised treatment allocation, again assuming binary treatment receipt, the two potential treatments that a patient may receive can be indexed according to whether they were assigned to experimental ($Z_i = 1$) or control treatment ($Z_i = 0$), i.e. $A_i(Z_i =$

$1) = A_{1i};\ A_i(Z_i = 0) = A_{0i}$. (Note that the subscripts denoting individual i can be dropped without ambiguity, given that outcomes values are assumed to be independently and identically distributed over the entire sample being studied.)

Given that it is impossible to observe both potential outcomes for a given individual, rather than attempting to determine the causal effect of treatment for a particular individual i (known as the **individual causal effect (ICE)** and defined by the difference $\Delta_i = Y_{1i} - Y_{0i}$), it is instead sensible to estimate some form of **average causal effect (ACE)** for the population of interest. For example, the average within-patient causal difference $ACE = E(Y_{1i} - Y_{0i})$ is (by usual laws of probability) equal to the difference in averages $E(Y_{1i}) - E(Y_{0i}) = E(Y_1) - E(Y_0)$ between the treatment and control observations for all units in the population of interest. If one can assume a *constant effect* of treatment on all units, the ACE as an estimate of the underlying true constant treatment effect is relevant and informative for all individuals (84).

5.2.4. Underlying assumptions

However, given that outcome data are available for only a small sample rather than all of the population, it is necessary to accept certain assumptions before the ACE, or any other causal parameter, can be estimated.

If treatment assignment is randomised and adhered to perfectly, randomisation provides independence between outcome and allocation of treatment; thus, under the assumption that the sample is generalisable to the target population, the outcome data $E(Y_j) = E(Y|Z = j)$ can be used to estimate this average quantity ACE by $E(Y|Z = 1) - E(Y|Z = 0) = E(Y_1) - E(Y_0)$, a result first demonstrated by Neyman (85). However any selection process which violates the independence between allocation of treatment and outcome (such as that typically introduced by deviation

from randomised treatment or by observational rather than experimental assignment of treatment) invalidates this estimation (84).

As such, the underlying **treatment assignment mechanism** is of fundamental importance. The causal methods to be discussed presently rely on **randomisation** to provide balance on all baseline variables (including all potential random variables and thus counterfactual outcomes) between groups. Methods for observational studies rely on the assumption that the randomisation balance can be simulated by conditioning on all relevant variables which confound the *a priori* independence between treatment allocation and outcome, invoking the so-called **strong ignorability** assumption (where "strong" implies an unconfounded and therefore causal interpretation), also known as the assumption of **no unmeasured confounders** (NUC), such that the relationship between the potential outcomes and received treatment can be assumed to be independent, given adjustment for all such confounders. In the case of **dynamic treatment regimens**, the **sequential ignorability** assumption would be required, adjusting for relevant (time-varying) covariates at each time point, to simulate sequential randomisation of treatments at each time k (107, 108). This assumption can be interpreted as, conditional on all pertinent baseline and time-varying confounders, those who do and do not change treatment at time t have the same probability of outcome, such that the decision on whether to change treatment is independent of underlying untreated outcome given these variables.

Formally, randomisation ensures that the probability that an individual with potential outcome Y_0 and Y_1 is assigned a certain treatment is a constant that does not depend on their potential outcomes Y_0 and Y_1, such that $P(Z|Y_0, Y_1) = P(Z)\ \forall\ Y_0, Y_1$, whereas the NUC assumption states that the probability of assignment is independent of all missing (Y_{mis}) and observed (Y_{obs}) outcomes, and can be assumed to be random once

conditioned on observed covariates X, i.e. $P(Z|X, Y_{obs}, Y_{mis}) = P(Z|X) \; \forall \; Y_{mis}, Y_{obs}$ (109).

Another basic assumption commonly invoked for causal analysis, necessary to allow statistical inference based on the assumption that individuals' observations are statistically independent, is the stable-unit-treatment-value assumption (**SUTVA**), which assumes that a full set of potential outcomes exist for each person (even though only one is observed) which are independent of outcomes and treatment status of all other subjects (110). Note that the SUTVA differs from the usual assumption of identical independent distributions of outcomes (i.e. that the outcome of an individual is unaffected by outcomes of other participants) required for standard statistical analysis, as it relates to stability of treatment received as well as outcome status. This assumption is necessary for the notation Y_{ik} to be sufficient in denoting the outcome of unit i with treatment k; otherwise, notation would need to include information on treatment received by all other units and the basic assumption of statistical independence between subjects (necessary for all statistical inference) would also fail (111). The SUTVA means that the observed outcome for unit *i* will equal one of the potential outcomes for that unit, no matter how the treatment was received (63).

This assumption also incorporates the notion of treatment stability, i.e. that each level or form of treatment is given to different units in an identical manner (so that there are no hidden versions of treatments), or equivalently that even if the treatment delivery varies slightly from individual to individual, the effect (i.e. potential outcomes) will be consistent. This assumption is required in order that the following equality will always hold: $Y_i = Y_{1i} * A_i + Y_{0i} * (1 - A_i)$ (112).

The SUTVA may be contravened in studies of infectious disease or when behavioural interventions are affected by interaction between individuals (85) or in cluster randomised behavioural or educational trials where participants may observe or influence each other's outcomes. Treatment stability may not hold in cases where natural variation occurs in the treatment process, for example when the rapport between therapist and individual affects the delivery of treatment for psychological disorders, or when clinicians' expertise influences the quality of surgical or medical treatment, or when practitioners' experience with one patient influences their subsequent treatment of another patient. In practice, some treatment variation is likely to occur, but such heterogeneity is assumed to remain within a reasonable range (113).

The **exclusion restriction** (ER) is a commonly employed assumption which implies no direct effect of treatment assignment, such that the *offer* of treatment itself has no influence on outcome, but instead any effect of randomisation on outcome is entirely due to its effect on treatment *receipt*. An alternative interpretation is that a patient who receives treatment will experience the same outcome whether or not they were originally randomised to receive it (114). In a continuous compliance setting, this assumption may be considered a "zero dose-response" assumption, such that without treatment, assignment will have no effect (115). The exclusion restriction is so called because the assignment mechanism is assumed to influence outcome "exclusively" through its effect on actual treatment received (112).

The exclusion restriction is likely to hold in double-blind trials (116) where the psychological effect of treatment assignment is (theoretically) negligible, but when participants are aware of the treatment they have been allocated to, there is potential for a strong placebo effect (or expectation of treatment) and thus this assumption may not hold. Unblinded behavioural intervention studies may be particularly prone to

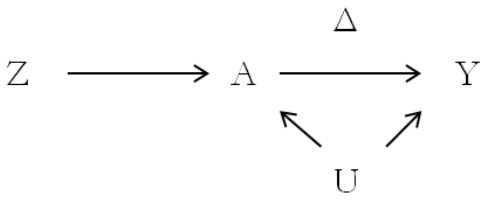
invalidation of this assumption, given that the impact of being assigned a certain behavioural intervention may affect an individual's outcome for reasons other than the intervention itself (for example, due to changes in a patient's mood or motivation levels, or because of an underlying rapport with therapist). Furthermore, the exclusion restriction will not hold if additional treatments or extra medical attention tend to be given in a certain treatment arm.

Before moving on to consider various approaches to causal methods which make these assumptions, it is worthy to note that, in making explicit the nature of relationships between variables and the underlying implicit assumptions regarding these relationships required for causal inferences, cDAGs implicitly demonstrate a number of key features of the causal techniques to be discussed.

5.2.5. cDAGs demonstrate underlying features and assumptions of causal techniques

First it is interesting to discover that the graphical representation of a trial with imperfect compliance using a cDAG implicitly depicts two such key assumptions that are often required in order to allow causal interpretation.

Pearl (104) points out that the underlying causal assumptions are encoded in a cDAG by the *missing* rather than the indicated links in a causal diagram; the existence of an arrow in the diagram suggests the potential for a causal connection (of unknown strength) between variables, whereas a *missing* arrow indicates a definite lack of causal link. Therefore, the randomisation (or independence) assumption is demonstrated by the lack of any common causes of Z with any other factor; as such, assignment Z can be assumed to be entirely independent of any underlying prognostic factors U.



Secondly, lack of a direct arrow between randomisation Z and outcome Y indicates that assignment Z has no direct influence on outcome Y except through the actual treatment received A (or equivalently, the effect of randomisation on outcome is entirely mediated by treatment exposure). This depicts the ER assumption, as it excludes the assignment mechanism variable *Z* from being a determining argument of the (A, Y) association.

Along with these two assumptions (randomisation and ER), this cDAG introduces another important feature of causal analysis, namely that of **instrumental variables**. The depiction of an arrow from Z to A indicates that there is a potential correlation between Z and A; if it can be assumed, not only that the effect of randomisation is entirely mediated by treatment receipt (in other words, the ER assumption holds) but also that randomisation has a *positive* effect on treatment exposure (such that randomisation to experimental treatment leads to a greater probability of receiving that treatment than randomisation to control treatment), randomisation is said to fulfill the role of an instrumental variable (IV) in this context.

Exploitation of randomisation as an IV is a common feature of the various causal techniques to be introduced in this chapter. The idea behind IVs is to avoid correlation between an endogenous variable (like compliance) and the error terms in a regression model, by substituting this endogenous variable (so called because, in a model regressing outcome on treatment received, treatment received would be correlated with the error terms; such endogeneity would render biased coefficients of the mediator)

with another "instrumental" variable (say, randomised treatment) which is correlated with the original predictor (compliance) and with outcome, but which does not directly affect outcome. Thus, the problem of hidden confounding may be circumvented by assuming a priori that randomisation per se has no direct effect on outcome, but rather its impact on outcome is entirely via treatment received (as per the ER assumption), which then (in conjunction with the assumption of random allocation) leads naturally to the use of randomisation allocation as an IV. Typically IV methods use a two-stage least squares (2SLS) process to estimate these coefficients, whereby the outcome is regressed on the predicted values of the mediator variable (obtained by first regressing this endogenous mediator variable (treatment received) on the IV (randomisation)), thus providing unbiased estimation coefficients of the effect of the original mediator variable on outcome.

The use of randomisation as an IV is implicit in many causal methods, but is most commonly associated with the more traditional method involve structural equation modelling.

5.3.Traditional causal estimation: SEMs

Traditionally, causal estimation has been approached through the use of structural equation modelling (SEM), where "structural" refers to the underlying structure of the causal relationship (as opposed to mere association) as demonstrated in a cDAG. Introduced by Wright in the 1920s, this method has been developed and used since the 1960s, primarily by sociologists and economists. SEM is very similar to classical simultaneous equation modelling, with features such as an underlying linear structure, constant coefficients and reliance on error terms to encompass the effect of missing variables (117). The focus of this regression approach is to estimate direct and indirect

(for example, through adherence) effects of treatment on outcome using instrumental variable (IV) techniques but with implicit reliance on a number of assumptions, firstly that the functional forms assumed by the model are correct, and secondly that there are no missing influences (U) on the outcome. In other words, conditional on the included covariates (X), there remains no confounding between mediator (treatment adherence) and outcome (thus invoking the NUC, or strong ignorability, assumption).

In general, SEM has not been widely accepted by statisticians as a causal estimation method because of a number of important inadequacies: first, the lack of transparency regarding underlying assumptions (expressed, for example, in terms of correlation between error terms) makes it is difficult to interpret, assess or communicate the validity of results and implications of the model results. Furthermore, given the likelihood of hidden or time-varying confounding bias, SEM is often viewed with suspicion because of a lack of confidence that full information has been collected on all confounding variables in order to satisfy the NUC assumption.

More recently, however, newer causal inference methodologies have been developed, with explicit underlying assumptions and clearly notated causal contrasts, thus allowing researchers to more easily understand the implications of the underlying assumptions and how the causal contrasts relate back to the research question of interest. Indeed, the most important difference between available approaches to causal estimation lies in their transparency with respect to their underlying assumptions, affecting the ability of researchers to interpret and evaluate the validity of each method.

These newer methods, to be discussed in the next section, are based on the PO framework but are not entirely unrelated to SEMs; indeed SEMs are actually implicitly based on the assumption of POs, although this is not obvious, while the use of

common assumptions (such as randomisation, or NUC, and ER) means that modern methods often also rely on randomisation as an IV (118).

5.4.Causal estimation methods

Having established the notation and assumptions required for causal methods based on the PO framework, two approaches to such estimation are now introduced, both of which are highly suited to trial data, as they rely on the randomisation assumption of balance between groups in terms of underlying untreated (or control) outcomes. The first approach employs this assumption to estimate treatment efficacy (i.e. effect of actually receiving treatment) for all patients, while the second estimates the effect of randomisation (i.e. the ITT effect of being assigned treatment) within a (latent) "complier" subgroup.

Other methods, originally developed for observational studies, rely on the assumption that all time-varying confounders between treatment receipt and outcome are known and measured, such that appropriate adjustment for these factors removes the selection bias associated with analysis by treatment received. Such methodology can be applied to randomised trials, and will be introduced following the discussion of the two randomisation-based approaches involving comparison of potential outcomes.

5.4.1. Principal stratification approach (CACE)

Continuing to assume only (time-fixed) binary (rather than quantitative or time-varying) compliance and a binary treatment contrast, the potential outcomes framework provides a basis for classifying individuals according to their underlying *potential* compliance status under each randomised treatment.

The principal stratification (PS) approach is based on stratifying the population into latent subgroups (**principal strata**) based on this potential compliance behaviour and then fitting (unbiased) ITT estimates (known as **principal effects**) within these principal strata.

A patient's principal stratum is considered to be a characteristic of the subject, effectively representing their underlying willingness to become exposed to the various treatments being contrasted. This characteristic is considered a baseline covariate, which is therefore unaffected by randomisation allocation, and as such can be used as a stratifying variable without introducing bias (unlike for other confounders measured after randomisation) (119). As such, methods based on PS can be thought of as latent class models, whereby the problem of deviation from randomised assignment is avoided by considering the (unconfounded) estimation of effect of treatment assignment (i.e. ITT effect estimate) within categories of patients according to their principal compliance status. The challenge presented by this approach is that an individual's status is never fully observed (i.e. it is latent) because it refers to their compliance behaviour on both possible treatments, whereas it is possible to observe only their observed compliance to their assigned treatment. As such, it is necessary to invoke additional assumptions in order to categorise patients into one of the possible principal compliance classes (119). Hence in this approach, counterfactuals enter with respect to definition of compliance class, defined assuming all-or-nothing compliance.

5.4.2. Compliance classes

The (latent) compliance classes can be defined by cross-classifying each patient according to the treatment that they would receive under each possible randomised allocation, in order words cross-classification of A_0 and A_1, where $A_0 = 0$ ($A_1 = 1$) if a

patient, when randomised to control (experimental) treatment, would adhere and actually take control (experimental) treatment; $A_0 = 1$ if instead they would instead take experimental treatment when randomised to control and $A_1 = 0$ if instead they would take control treatment despite receiving allocation to experimental treatment.

The table below shows how this cross-classification provides the definition of the underlying compliance classification status (C) of each individual:

Table 4 Compliance classification

	Treatment received when randomised to control arm	
	Control treatment ($A_0 = 0$)	Experimental treatment ($A_0 = 1$)
Treatment received when randomised to experimental arm		
Control treatment ($A_1 = 0$)	Never taker (N)	Defier (D)
Experimental treatment ($A_1 = 1$)	Principal complier (PC)	Always taker (A)

In other words, an **always taker** (A) takes experimental treatment regardless of their randomised allocation ($A_Z = 1 \; \forall \; \mathrm{Z}$), a **never taker** ($N$) will always takes control treatment ($A_Z = 0 \; \forall \; \mathrm{Z}$) and a **defier** ($D$) will always take the opposite treatment to that which they were randomised ($A_Z = 1 - Z \; \forall \; \mathrm{Z}$). A **principal complier** (PC) is defined as an individual who would always take the treatment to which they were randomised ($A_Z = Z \; \forall \; \mathrm{Z}$).

The term "principal complier" is used in preference to "complier" in order to distinguish between those who have simply been *observed* to comply with their randomised allocation only and those who would comply with *either* randomised allocation. Those observed to comply with experimental arm allocation may be always takers or principal compliers, whereas those observed to comply when randomised to control treatment may be never takers or principal compliers. As such, observed

compliance status alone does not provide full information on a patient's underlying principal stratum. More formally, given that, for each individual, we are able to only observe their compliance status A_{z_i} under their randomly assigned treatment z_i, their compliance status $A_{(1-z_i)}$ is unobserved and therefore their underlying compliance class is unknown.

Once these compliance classes have been identified, unbiased causal estimation is thereafter achieved by estimating the (unbiased) effect of assigned treatment (i.e. according to randomisation) only in the subgroup of principal compliers who would comply with whatever treatment they were randomised to (and thus distinct from observed compliers). Given that no deviation from randomisation occurs in this subgroup, analysis according to randomisation equates to analysis according to treatment received for this subgroup of patients, and thus this effectiveness estimate can be granted causal interpretation.

Thus the PS approach involves estimation of ITT within latent compliance classes known as principal strata. The resulting (unbiased) ITT estimate of treatment effectiveness calculated within the principal strata of potential compliers can be interpreted as a measure of treatment efficacy relating to such patients, known as the **complier average causal effect (CACE)**. This is because, for this universally compliant subgroup, treatment received is always equal to randomised treatment; as such analysis according to randomised treatment (ITT analysis) within this group of patients equates to (unbiased) analysis according to treatment receipt. However, rather than relating to the entire population, the CACE is therefore generalisable only to the (unobservable) subgroup of patients who would naturally comply with their assigned treatment.

5.4.3. CACE vs PP vs AT: how PO formulation clarifies deficiencies

The key to PS methods is using the potential framework to define patients' potential compliance status, thus revealing the distinction between a principal complier (compliance under *both* treatments) and an observed complier (compliance under the treatment actually assigned). Little (109) demonstrates how the PS formulation can be used to clarify the biases introduced by naïve analysis methods such as PP or AT.

Using the notation $Y(Z = j) = Y_j$ to denote a patient's potential outcome when randomised to treatment j, the CACE estimate is denoted as

Equation 1

$$CACE = E(Y_1 - Y_0 | C = PC)$$

where the conditioning on $C = PC$ is identical to conditioning on $A_1 = 1$ and $A_0 = 0$. The CACE is a valid causal effect because it is a summary measure of individual-level effects ($Y_1 - Y_0$) in a subpopulation of interest, namely compliers.

In contrast, AT analysis, which classifies subjects by the adopted treatment $A(Z) = A_Z$, estimates

$$AT = E(Y_Z | A_Z = 1) - E(Y_Z | A_Z = 0)$$

This estimate does not carry causal interpretation, as it is not an average of individual-level causal effects; instead it compares outcome averages for different groups with different characteristics. In particular, it compares the average outcome of those adopting experimental treatment (namely all always-takers, principal compliers who were assigned experimental treatment and defiers who were assigned to the control treatment) with the average outcome of those adopting control treatment (namely compliers randomised to receive control treatment, defiers randomised to experimental treatment and never-takers who were assigned either treatment).

Similarly, PP analysis, in estimating

$$PP = E(Y_Z|A_1 = 1) - E(Y_Z|A_0 = 0)$$

is not an average of individual-level causal effects. In comparing only those subjects who actually adopted their assigned treatments, a PP analysis compares the outcomes of those who were assigned and received experimental treatment (compliers and always-takers assigned to experimental treatment) with the outcomes of those who were allocated to and received control treatment (that is, compliers and never-takers randomised to control treatment). In order for the PP analysis to be valid, unrealistic assumptions are required: the mean outcome for those who comply with experimental treatment is assumed to be the same both for compliers in the experimental arm and always takers from both treatment arms, and similarly the mean outcome for those who comply with control treatment is assumed to be constant, regardless of whether they are non-compliers from the experimental arm or never takers from both treatment arms: i.e. outcome is unrelated to compliance: $Y \perp C|C = N\ or\ PC, Z = 0$ and $Y \perp C|C = A\ or\ PC, Z = 1$ (109).

5.4.4. Heuristic derivation of CACE: assumptions required

Although CACE in Equation 1 above is simply the ITT estimate among the class of "compliers" and thus a valid causal estimand, it is not immediately apparent how it can be estimated, given that underlying principal compliance status (C) of individuals is not known (109). The answer lies in being able to assume certain plausible restrictions. Indeed, the heuristic derivation of the CACE estimate provides a useful platform for demonstrating the relevance and use of the assumptions most commonly applied in causal analysis.

In particular, in order to estimate the CACE, a number of standard assumptions (SUTVA, ER and randomisation) are required, along with two further restrictions (monotonicity and non-zero compliance). The following explanation of CACE estimation demonstrates how and why these assumptions are implemented.

First note that, assuming randomised allocation of treatments, the usual ITT estimate is equal to

$$ITT = E(Y_1 - Y_0)$$

which, by the usual laws of probability, equates to

$$\begin{aligned} ITT = E(Y_1 - Y_0) & \\ &= E\big((Y_1 - Y_0)|C = PC\big)\, P(C = PC) \\ &+ E\big((Y_1 - Y_0)|C \neq PC\big)\, P(C \neq PC) \end{aligned}$$

Note next that, if the exclusion restriction holds, the ITT treatment effect (i.e. difference in outcome between randomisation to treatment and control arms) in never-takers and always-takers will always equal zero. This follows because the exclusion restriction states that randomisation has no effect beyond that of treatment received; thus, given that, in these compliance subgroups, treatment receipt does not vary with randomisation, the difference in potential outcomes for never- and always-takers will always equal zero; i.e. $Y_0 = Y_1$ if $A_0 = A_1$.

The next assumption required is that of **monotonicity**, which implies that there are no defiers (subjects who would always take alternative treatment to that randomised) This assumption is named 'monotonicity' because treatment assignment is assumed to work in one direction only, such that $A_{1i} \geq A_{0i} \forall i$ or equivalently $P(A_{1i} \geq A_{0i}) = 1$. If

these two assumptions hold, it can be noted that the second term of the equation above is equal to zero, because

$$E\big((Y_1 - Y_0)|C \neq PC\big)\, P(C \neq PC)$$

$$= E\big((Y_1 - Y_0)|C = A\big)\, P(C = A)$$

$$+E\big((Y_1 - Y_0)|C = N\big)\, P(C = N)$$

$$+E\big((Y_1 - Y_0)|C = D\big)\, P(C = D)$$

and $E\big((Y_1 - Y_0)|C = N\big) = E\big((Y_1 - Y_0)|C = A\big) = 0$ (by ER) and $P(C = D) = 0$ (by monotonicity). Thus

$$ITT = E(Y_1 - Y_0) = E\big((Y_1 - Y_0)|C = PC\big)\, P(C = PC) = CACE * P(C = PC)$$

As such, the CACE estimate is simply the ITT estimate (of the difference in outcome between randomised groups) divided by the proportion of principal compliers (which, in the final assumption of this analysis (namely, the **non-zero compliance** assumption), is taken to be positive; in other words, it is assumed that there exists at least one individual in the trial who would always comply with their randomised treatment allocation). Although unknown, this proportion can be estimated very simply and unbiasedly if one invokes the randomisation and monotonicity assumptions. In Table 4, it may be noted that the observed proportion of patients randomised to the experimental treatment arm who actually receive this treatment (denoted here as p_1) is an unbiased estimate of the combined (population) proportions of compliers and always-takers; similarly the proportion (p_0) of control arm patients who adopt experimental treatment is an unbiased estimate of the population proportion of always-takers (as there are assumed to be no defiers). As such, the true proportion of compliers $P(C = PC)$ can be estimated by the difference in the observed proportions of

experimental and control arm patients who receive experimental treatment ($p_1 - p_0$), which is equal to the ITT difference (i.e. when analysing according to randomisation) of experimental treatment receipt.

Therefore, under the above stated conditions, the CACE estimate is simply a ratio of two ITT effects: the ITT effect (of randomisation) on outcome divided by the ITT effect (of randomisation) on treatment receipt. For example, with observed mean outcomes, $\bar{y}_0$ and $\bar{y}_1$, from the control and experimental treatment groups respectively, the CACE estimate is given as

Equation 2

$$CACE = \frac{\bar{y}_1 - \bar{y}_0}{p_1 - p_0} = \frac{ITT\ effect\ on\ outcome}{ITT\ effect\ on\ treatment\ receipt}$$

The use of assumptions (namely SUTVA, randomised assignment, exclusion restriction and monotonicity) in this derivation of CACE links it to IV estimation, as randomisation satisfies the necessary conditions of an IV in this setting. As such, this ratio is sometimes called the **instrumental variable (IV)** estimator and was demonstrated by Angrist (117) to be equivalent to the IV estimand.

Newcombe (53) was the first to describe PS (IV) methods to adjust for non-compliance from randomised treatment, demonstrating the equation above for a continuous outcome. Newcombe described the link between the explanatory (CACE) and pragmatic (ITT) estimates of treatment effect in terms of an attenuation factor (difference between the proportions in each randomised group who receive experimental treatment, or equivalently the difference between group proportions receiving control treatment), and pointed out that counterintuitively, this attenuation factor depends, not on the *sizes* of the selection biases associated with deviation from assigned treatment but instead only on the *frequency* of such deviations.

5.4.5. Demonstrate method for binary outcome: Sommer

Sommer and Zeger (47) then demonstrated a similar idea adjusting for binary non-compliance with a binary outcome, their methodology having been motivated and illustrated by the Vitamin A trial (described in Chapter 4). Given that children in the control arm of this trial were not given the option of active treatment (such that only two PS exist in this setting: by definition, patients are not able to be always-takers or defiers), their method (later referred to CACE) relates to assessing outcome among those who would have complied with *active* treatment if randomised to receive it.

As such, patients are assumed to either be principal compliers or non-compliers (with active treatment). Although such a subgroup is directly observed among those randomised to vitamin A, it is not possible to directly ascertain which patients would have complied with active treatment in the control arm, and thus it is necessary to compare the outcome among those observed to be compliant in the active treatment arm with an *inferred* control subgroup who would have complied with active treatment. Thus, as with all causal methods, their analysis relies on a number of untestable but potentially plausible assumptions (with the degree of plausibility needing to be assessed according to the particular clinical setting).

The derivation of their CACE estimate, given in terms of the relative risk (R) of death among those children who would receive treatment if allocated to it, is demonstrated here as an example of how these assumptions are applied in practice, highlighting issues that may need similar consideration by researchers wanting to apply such methods. (Note that the original article displayed an incorrect formula without derivation – as such, these workings have been included here to demonstrate my own work and understanding of the concepts).

Table 5: Table IIa in (91)

Control group: principal compliers				Treatment group: principal compliers			
	No	Yes	Total		No	Yes	Total
Alive	m_{00}	m_{01}	$m_{0\cdot}$	Alive	n_{00}	n_{01}	$n_{0\cdot}$
Dead	m_{10}	m_{11}	$m_{1\cdot}$	Dead	n_{10}	n_{11}	$n_{1\cdot}$
Total	$m_{\cdot 0}$	$m_{\cdot 1}$	M	Total	$n_{\cdot 0}$	$n_{\cdot 1}$	N

Using Sommer's notation given in Table 5, the relative risk estimate can be expressed as

Equation 3

$$\hat{R} = \frac{n_{11}/n_{01} + n_{11}}{m_{11}/m_{01} + m_{11}}$$

However, given that we do not know which control patients would have complied with active treatment, it is not possible to observe m_{01} and m_{11}, so these values must be estimated using the two key assumptions, namely randomisation and ER.

By the randomisation assumption, we assume that $\frac{m_{\cdot 0}}{M} = \frac{n_{\cdot 0}}{N}$ (given that randomisation balance allows us to assume that the proportion of patients who would comply with active treatment is assumed to be equal between groups), such that

Equation 4

$$\frac{m_{00} + m_{10}}{M} = \frac{n_{00} + n_{10}}{N}$$

$$m_{00} + m_{10} = M[\frac{n_{00} + n_{10}}{N}]$$

By the ER assumption, we assume that $\frac{m_{10}}{m_{00}+m_{10}} = \frac{n_{10}}{n_{00}+n_{10}}$ (because, in the absence of treatment, the probability of outcome is assumed to be independent of randomised allocation), such that

$$m_{10} = [m_{00} + m_{10}][\frac{n_{10}}{n_{00} + n_{10}}]$$

Substituting in from Equation 4

$$m_{10} = M[\tfrac{n_{00}+ n_{10}}{N}][\tfrac{n_{10}}{n_{00}+n_{10}}] = n_{10}\tfrac{M}{N}$$

Thus Equation 4 becomes

$$m_{00} = [n_{00} + n_{10}]\frac{M}{N} - n_{10}\frac{M}{N} = n_{00}\frac{M}{N}$$

Thus we can calculate estimates of m_{01} and m_{11} as follows:

Equation 5

$$\hat{m}_{01} = m_{0.} - \hat{m}_{00} = m_{0.} - n_{00}\frac{M}{N}$$

Equation 6

$$\hat{m}_{11} = m_{1.} - \hat{m}_{10} = m_{1.} - n_{10}\frac{M}{N}$$

In order to estimate the relative risk among compliers, given in Equation 3,

$$\hat{R} = \frac{n_{11}/n_{01} + n_{11}}{m_{11}/m_{01} + m_{11}}$$

$$\hat{R} = \left[\frac{n_{11}}{n_{01} + n_{11}}\right]\left[\frac{m_{01} + m_{11}}{m_{11}}\right]$$

Using the estimated values of $\hat{m}_{01}$ and $\hat{m}_{11}$ given in Equation 5 and Equation 6, we obtain

$$\tilde{R} = \left[\frac{n_{11}}{n_{01} + n_{11}}\right]\left[\frac{\hat{m}_{01} + \hat{m}_{11}}{\hat{m}_{11}}\right]$$

$$\tilde{R} = \left[\frac{n_{11}}{n_{01} + n_{11}}\right]\left[\frac{\left(m_{0.} - n_{00}\frac{M}{N}\right) + \left(m_{1.} - n_{10}\frac{M}{N}\right)}{m_{1.} - n_{10}\frac{M}{N}}\right]$$

$$\tilde{R} = \left[\frac{n_{11}}{n_{01} + n_{11}}\right]\left[\frac{m_{0.} + m_{1.} - (n_{00} + n_{10})\frac{M}{N}}{m_{1.} - n_{10}\frac{M}{N}}\right]$$

$$\tilde{R} = \left[\frac{n_{11}}{n_{01} + n_{11}}\right]\left[\frac{m_{0.} + m_{1.} - (n_{00} + n_{10})\frac{M}{N}}{m_{1.} - n_{10}\frac{M}{N}}\right]$$

$$\tilde{R} = \left[\frac{n_{11}}{n_{01} + n_{11}}\right]\left[\frac{M - n_{.0}\frac{M}{N}}{m_{1.} - n_{10}\frac{M}{N}}\right]$$

$$\tilde{R} = \left[\frac{n_{11}}{n_{01} + n_{11}}\right]\left[\frac{M\left(1 - \frac{n_{.0}}{N}\right)}{m_{1.} - n_{10}\frac{M}{N}}\right]$$

$$\tilde{R} = \left[\frac{n_{11}}{n_{01} + n_{11}}\right]\left[\frac{M\left(\frac{N - n_{.0}}{N}\right)}{m_{1.} - n_{10}\frac{M}{N}}\right]$$

$$\tilde{R} = \left[\frac{n_{11}}{n_{.1}}\right]\left[\frac{M\left(\frac{n_{.1}}{N}\right)}{m_{1.} - n_{10}\frac{M}{N}}\right]$$

$$\tilde{R} = \frac{n_{11}}{\left[m_{1.}\frac{N}{M} - n_{10}\right]}$$

Note that this (corrected) formula appears in (47).

A more intuitive set of calculations may be given as follows. First, the compliance information in treatment arm and the balance produced by randomisation is used to estimate the number of children in the control who would be non-compliant. Given that the proportion of non-compliant patients is assumed to be equal in the two

randomised groups i.e. $\frac{m_{.0}}{M} = \frac{n_{.0}}{N}$, the number of children in the control who would be non-compliant is assumed to be equal to the proportion of non-compliers in the treatment arm multiplied by the number of patients in the control arm, i.e. $m_{.0} = M\ \frac{n_{.0}}{N}$. Thus in this example, $m_{.0} = 11{,}588 * \frac{2419}{12{,}094} = 2317.8$.

The inferred number of control children who *would* comply with intervention is therefore equal to $m_{.1} = M - m_{.0} = 11{,}588 - 2317.8 = 9270.2$.

Under the ER assumption, the mortality risk of the non-compliers in the control groups is assumed to be equal to the mortality risk among non-compliers in the treatment arm (as neither of these groups receive any treatment) i.e. $\frac{m_{10}}{m_{.0}} = \frac{n_{10}}{n_{.0}}$ and thus $m_{10} = m_{.0}\frac{n_{10}}{n_{.0}} = 2317.8 * \frac{34}{2419} = 32.6$.

Thus, by definition, the number of non-compliant control children expected to survive is equal to $m_{00} = m_{.0} - m_{10} = 2317.8 - 32.6 = 2285.2$, and thus the number of compliant control children expected to die and survive can be calculated by subtracting the expected number of non-compliant children who will die and survive, respectively, from the known number of control children who died and survived, respectively.

$m_{01} = m_{0.} - m_{00} = 11{,}514 - 2285.2 = 9228.8$

$m_{11} = m_{1.} - m_{10} = 74 - 32.6 = 41.4$

Table 6

	Control compliance		
	No	Yes	
Alive	2285.2	9228.8	11,514
Dead	32.6	41.4	74
Total	2317.8	9270.2	11,588

	Treatment compliance		
	No	Yes	
Alive	2385	9663	12,048
Dead	34	12	46
Total	2419	9675	12,094

Thus the estimated risk reduction among compliers can be calculated as

$$\tilde{R} = \frac{n_{11}/n_{01} + n_{11}}{m_{11}/m_{01} + m_{11}} = \frac{12/9675}{41.4/9270.2} = 0.277$$

Thus the causal effect of treatment among compliers can be interpreted as leading to a risk of mortality among those who take treatment equal to just 28% of the risk among those who do not take treatment, i.e. treatment is estimated to reduce the risk of mortality by more than 70%.

5.4.6. Interpretation of CACE

The attractive simple nature of PS methods must be weighed against their inherent limitations, the first being the implicit restrictions on interpretation of the CACE estimate.

It is imperative to appreciate that the resulting CACE estimate applies only to the (theoretical) subgroup of "principal compliers", the exact interpretation of which must be determined in the light of the trial setting. Rather than relating to the entire population, the CACE is therefore generalisable only to the (unobservable) subgroup of patients who would naturally comply with their assigned treatment.

As such, the CACE relative risk estimated above must be viewed in light of the barriers leading to (principal) non-compliance in the vitamin A trial. This causal effect of treatment relates only to those who would have had the ability to receive treatment if they were allocated to the treatment arm, and therefore cannot be generalised to those who were not able (often for logistic reasons) to receive treatment (principal non-compliers). Given that it was evident that children who were unable to receive treatment tended to have worse baseline health, if the effect of treatment varies according to the baseline mortality risk, it may not be appropriate to use the CACE to

generalise the treatment effect to the entire population. However, as pointed out by Sommer, there are no data available from this trial to test whether such an interaction exists.

Furthermore, Sommer points out that this reduction includes both the *biologic* impact of vitamin A and any *placebo* effect of having active treatment (compared to no treatment). Without a placebo, it is not possible to differentiate between these two effects and thus it is not possible to isolate the biological impact of treatment. However, such an effect would also be included in the ITT estimate of any trial that does not involve a placebo, and given that the outcome is entirely objective, one would expect the placebo effect to be minimal.

Although interpretation of any CACE estimate is limited in that it relates only to the (unidentifiable) PS consisting of participants who would "comply" (as defined by trial conditions) in future, such an estimate is likely to be of particular interest to motivated patients and their treating clinicians. Furthermore, PS estimates from *all* the different compliance classes may be combined as a weighted average, assigning weights to each PS estimate, depending on (and assuming availability of) reliable estimates projecting the likely compliance profile in the population to which the results are to be generalized, thus providing an overall estimate of the likely average treatment effect for the whole general population. This weighted average may be of particular interest for policy makers who want to determine the benefit if all potential individuals were assigned treatment and is likely to be more reliable than the usual ITT effect estimate, which is realistic only when the population compliance patterns match those demonstrated in the trial (112).

However, Shrier (113) points out that a patient's PS membership may differ between trial and real-life conditions. For example, if participants in the control arm are not able

to receive active treatment in the trial setting, the latent subgroup of always-takers is assumed not to exist, whereas in practice the active treatment may be freely available, such that, in practice, patients may be categorized into different principal strata to those available in the trial. Indeed Shrier (113) describes how the real-life conditions under which a treatment is available may alter a patient's PS: if treatment is free and readily available, a patient may be an always-taker, whereas if it is expensive or difficult to access, they may be a never-taker; if affordable or available with a small degree of effort, they may be a principal complier; as such, if it is plausible that a patient's underlying compliance status may change with conditions (for example, in terms of treatment availability and patient motivation), differences between trial and real-life settings may render even the basic (rather than overall weighted) CACE ungeneralisable.

5.4.7. Limitation to PS methods

The other, most notable, limitation of PS methods is the requirement for binary compliance classification. Although, in theory, PS estimation embraces a broad class of models which includes general forms of non-compliance (such as time-dependent, continuous or multivariate), they are unbiased generally only for binary non-compliance and binary treatment comparisons, and thus at most four latent compliance classes (depending on the restrictions imposed by the trial design on whether control (or experimental) arm patients are able to receive the alternative treatment). For example, as in the vitamin A trial, if the trial treatment protocol prevents control patients from receiving experimental treatment (such that non-compliance only occurs among experimental arm patients who either do or do not receive experimental treatment), only two compliance classes actually exist, as patients do not have the opportunity, within the trial setting, to be always-takers or defiers.

Any extension to PS methods, for example to allow for partial compliance or comparison of (compliance with) more than two treatments, add to the complexity of models by requiring more principal compliance categories (for example, where compliance-type is defined by the time at which an individual would stop treatment if randomised to treatment) and more complicated assumptions in order to ensure identifiability of the causal estimates (69). For example, Sommer suggests that their method could be extended to allow more than two categories of compliance or continuous compliance by imposing some parametric structure on how the probability of outcome depends on compliance in the control group (for example, by assuming a logistic regression model that expresses likelihood of outcome for the control group as a function of compliance). Goetghebeur (115) demonstrates this by estimating causal effect for ordered binary categories (ordinal) under the assumption of a monotone dose-response relationship between treatment received and outcome. However, in general, methods to accommodate partial or time-varying treatment adherence require complex assumptions and are not fully developed (113).

Furthermore, although model-based extensions to PS methodology are possible (for example, allowing for missing outcome data ((120-123) or including covariates for model-based estimation (124) using maximum likelihood (125) or Bayesian techniques (126, 127), inference using such models tends to be particularly sensitive to distributional assumptions, and still necessarily rely on classification of compliance into discrete (usually binary) compliance categories.

This is a major limitation, given that trials most likely to be subject to treatment nonadherence are often those with long-term treatment and follow up schedules, adherence to which may not be adequately summarised using binary compliance classes. A further disadvantage is that in such trials, interest typically lies in assessing TTE

outcomes (thus capturing more information on the temporal aspect of treatment effect on outcome) rather than simpler binary or continuous outcomes; however, PS methodology has not been extended to analysis of such survival outcomes.

These limitations lead us to now consider the second approach to estimation of causal effects using the PO framework, which was not only developed originally for use with TTE outcomes but also which can adjust for more flexible measures other than simply binary compliance.

5.5.Efficacy-based: structural modelling

The alternative causal methodology to PS based on the PO framework is that of structural modelling. Models for the distribution of potential outcomes are called **structural** (meaning causal) because of the underlying structure (as depicted in a cDAG) in a causal relationship, rather than mere association between variables.

Structural models (SMs) make no assumption regarding the relationship between compliance and outcome, but instead model the relationship between observed and underlying untreated outcomes in terms of the relevant compliance measure (for example, cumulative dose of drug taken or time to treatment deviation) with estimation based on the assumed balance between randomised groups in terms of baseline (untreated) outcomes. As such, the concept of POs facilitates causal estimation using these models by defining the (potentially latent) subject-specific untreated outcome as the reference, against which the observed (treated) data may be modelled with explicit estimation of the causal parameter (73).

Thus in contrast to the CACE estimate, which relates only to those patients who would naturally comply with their assigned treatment, a structural model provides a causal estimate relating to the entire population. However, the validity of such an estimate may

be equally questionable, if it is not feasible to believe that everyone in the population could be induced to comply.

Compared to PS methods, this more general modelling approach to causal estimation means that SMs are able not only to incorporate more flexible measures of compliance but also to accommodate survival outcomes. Indeed after the concept of structural modelling was first introduced by Robins (128), SMs were developed by Robins and Tsiatis (129) specifically for TTE outcomes, in order to address the problems associated with adjustment for time-varying treatments using standard methods of survival analysis when treatment changes are typically indicative of patient prognosis.

5.5.1. Standard survival methods

Analysis of survival outcomes, which record the time from randomisation to the outcome event of interest (for example, time to disease progression or death) is complicated by the need to allow for censoring when follow up occurs prior to observation of the endpoint of interest, such that individuals are included in analysis only until the time at which their follow up ceases.

Standard survival analysis methods assume uninformative censoring (i.e. censoring occurs completely at random), such that those who prematurely cease follow up can be considered a random sample of those remaining in the trial and thus censored at the point of loss to follow up without introduction of bias. One such method is the Cox proportional hazards (PH) model, the most commonly used model for survival outcomes which avoids the need for any distributional assumptions of underlying survival times by simply estimating the relative hazard (or hazard ratio, HR) between treatment groups, which is assumed to remain constant over the trial follow up period (under the proportional hazards (PH) assumption).

For example, in order to investigate the effect of a binary treatment Z, the basic Cox model $h(t) = e^{\beta Z} h_0(t)$ relates a patient's (unspecified) underlying hazard $h_0(t)$ with their observed hazard $h(t)$ at time t through an estimated treatment HR e^{β} (expressed in terms of an exponential variable, given that such a ratio is necessarily positive). However, given that changes to treatment are unlikely to occur randomly with respect to a patient's prognosis, censoring patient's analysis at the point at which deviation from randomised treatment occurs (as in an "on treatment" (OT) version of a PP analysis) will not appropriately protect against bias.

Similarly, an AT adjustment for a time-varying treatment covariate $\bar{A}(t)$ (a summary measure such as cumulative dose or time spent on treatment until time t, or actual treatment received at time t) in a time-varying Cox model will reflect only the associational rather than causal relationship between treatment and outcome, given that the estimated HR e^{α} fails to distinguish between causal and potential selection (or confounding) effects (130).

$$h(t) = e^{\alpha \bar{A}(t)} h_0(t)$$

Furthermore, attempts to adjust for time-varying confounders $X(t)$ (which are not only affected by prior treatment but which also influence future treatment choices as well as outcome) will not help to remove such biases, given that, in removing some of the effect of treatment received, the resulting treatment effect estimates would reflect neither the overall (unadjusted) nor direct (causal) effect of treatment on outcome (130). Indeed, in adjusting for such time-varying confounders (TVCs) which are also affected by prior treatment, further bias is introduced, as the TVC lies on the causal pathway between treatment exposure and outcome (at a particular time t); therefore explicit

control for this confounder will distort the treatment effect estimate, as the effect of treatment on the outcome acting via the confounder will be removed (131, 132).

5.5.2. Accelerated failure time model

An alternative standard survival model which is less common, but may be especially useful when the PH assumption does not hold, is the accelerated failure time model (AFTM). In contrast to the Cox PH model in which factors are assumed to affect the likelihood of the outcome occurring at a particular time point, the explanatory variables in an AFTM are assumed to act multiplicatively on the time scale, thus instead affecting the *rate* at which an individual proceeds along the time axis (133). If $S_0(t)$ and $S_1(t)$ denote a patient's probability of surviving beyond time t on control and experimental treatments, respectively, the relationship between these times, assuming an AFTM with (again necessarily positive) acceleration factor (so-called because of its effect on the rate at which lifespan is used up), e^{β}, would be denoted $S_1(t) = S_0(t/e^{\beta})$. Thus $e^{\beta} < 1$ (i.e. $\beta < 0$) would imply an acceleration in time to event with experimental treatment compared to control, such that, with a positive (negative) event, such as healing (death), $e^{\beta} < 1$ implies the experimental (control) treatment is preferable (74).

5.5.3. Rank-preserving structural failure time model

The structural version of the AFTM was introduced by Robins (129); this model is structural in the sense that it makes a direct within-patient (rather than between patient) comparison of their observed (Y_i) and (potentially latent) untreated survival times (Y_{0i}). The structural AFTM is also known as the rank-preserving structural failure time model (RPSFTM), "rank-preserving" because, if the untreated event time for individual i is less than that for individual j, the predicted treated event time for individual i would

also be less than that for individual j (thereby implying an assumption of no treatment-covariate interactions)(74).

The acceleration factor (AF) is estimated using a process called G-estimation, based on the assumption of balance produced by randomisation between treatment groups in terms of patients' (potential) untreated survival times U_{0i}. The model assumes that an individual's observed survival time Y_i is related to their underlying baseline survival time U_{0i} through an AF e^{β} assumed to vary with treatment received $A_i(t)$ (130).

$$U_{0i}(\beta) = \int_0^{Y_i} e^{\beta A_i(t)}\, dt$$

Thus the RPSFTM relates observed outcome Y_i to Y_{0i} (counterfactual or potential outcome that would have been observed without treatment) through treatment effect e^{β} (hence structural as it relates potential outcomes on the *same* individual).

Assuming discrete intervals of time, and binary treatment receipt (i.e. $A_i(t)$ is a binary time-dependent variable denoting receipt of experimental ($A_i(t) = 1$) or control ($A_i(t) = 0$) treatments over time), where a patient's observed survival time T_i can be expressed in terms of the amount of time spent on experimental (T_{1i}) and control (T_{0i}) treatments, where $T_i = T_{0i} + T_{1i}$ (134), the relation between treatment receipt and underlying untreated outcome can then be simplified as

$$U_{0i}(\beta) = T_{0i} + e^{\beta} T_{1i}$$

such that the counterfactual (untreated) survival time U_{0i} can be interpreted as the sum of time spent on control treatment and the time spent on experimental treatment multiplied by the AF e^{β} (135). Thus, an AF equal to e^{β} implies that every unit of time spent on experimental treatment would equate to e^{β} units of time on control treatment.

Thus $e^{\beta} < 1$ (i.e. $\beta < 0$) would imply an acceleration in time to event with experimental treatment compared to control (treatment decreases event time), whereas $e^{\beta} > 1$ (i.e. $\beta > 0$) would imply deceleration in time to event with experimental treatment compared to control (treatment extends event time). The optimal value of β is determined using an unusual test procedure known as G-estimation.

5.5.4. G-estimation

Given the reliance on unobserved counterfactuals, the acceleration factor e^{β} cannot be estimated using usual methods for associational models. Instead, a method known as G-estimation is employed, based on the assumption that the untreated survival times U_{0i} can be considered a baseline feature unaffected by post-randomisation treatment or confounding, which on average will differ only randomly between the randomised groups, i.e. $U_0 \perp Z$. Thus, in exploiting the expected independence between the underlying untreated outcome and randomisation (the IV) under the randomisation and ER assumptions, G-estimation is the general version of IV estimation for time-varying treatments.

Causal estimation is based on finding a value of the treatment effect parameter (β) that attains baseline balance in terms of potential untreated outcome U_{0i} between randomised groups (69).

In practice, G-estimation is carried out on a trial and error basis, testing a range of proposed values of the acceleration $e^{\widehat{\beta}}$ factor to determine which achieves the best balance between randomised groups in terms of their (estimated) values of $U_{0i}(\hat{\beta})$. By specifying a plausible range of values for β, and intervals within this range, the user defines a grid over which to carry out a search for the optimal value of $\hat{\beta}$ (134).

Balance between randomised groups is assessed, for each $\hat{\beta}$, according to the test statistic $Z(\hat{\beta})$ corresponding to the hypothesis that $\hat{\beta} = \beta$, whereby the point estimate of $\hat{\beta}$ is chosen as that for which $Z(\hat{\beta})$ is closest to 0 (indicating equality between treatment arms in terms of $U_{0i}(\hat{\beta})$). The hypothesis test used for this procedure is usually the same as that used for the original ITT analysis, for example the log rank test, Wilcoxon or Cox model; given the nonparametric nature of these tests, $Z(\hat{\beta})$ is a step function; thus in practice, values of $\hat{\beta}$ are plotted against corresponding $Z(\hat{\beta})$ values in order to identify $\hat{\beta}$ for which $Z(\hat{\beta})$ is equal to 0. Similarly, the 95% CI for $\hat{\beta}$ consists of all values which fail to reject the null hypothesis at 5% (i.e. CI limits for $\hat{\beta}$ are those for which the hypothesis test p-value is equal to 0.05) (136). It may be useful to consult graphical displays of $Z(\hat{\beta})$ against $\hat{\beta}$ to check that a unique value of $\hat{\beta}$ exists, and its associated confidence limits are estimable (134). For example, if no values of $e^{\hat{\beta}}$ rejects the null hypothesis at the 5% level, the upper CI limit will not be estimable.

In order to check whether the estimation procedure has worked well, it is useful to compare the HR estimated between control group and experimental group counterfactual survival times (in other words, the HR between the groups adjusted as if none of the patients had received experimental treatment): if the estimation procedure has worked well, this value should be very close to 1. This can be further investigated by checking that the KM graphs of these modelled counterfactual event times for each group are similar.

G-estimation preserves ITT p-value

Given that $U_{0i}(0) = T_i$, the hypothesis $\beta = 0$ simply corresponds to the null hypothesis of the ITT analysis, as by definition the test of $\beta = 0$ is the straightforward

test of $T \perp Z$, thus demonstrating how G-estimation preserves the p-value associated with ITT analysis (leading to the term "randomisation-based efficacy estimator" (RBEE) (69). This in turn means that the 95% CI for β will include 0 if and only if the ITT analysis is not significant at the 5% level.

5.5.5. Assumptions

It is generally not possible to identify the RPSFTM causal parameter without making some assumptions that are themselves non-identifiable (129). Firstly, use of a RPSFTM implies the (untestable) assumption that the underlying **acceleration model** is correct.

Assumptions underlying **G-estimation** are that randomisation must have adequately balanced the alternative treatment arms with respect to prognostic characteristics, and that the effect of randomisation is mediated entirely by treatment receipt (such that those who receive a given duration of treatment will experience the same treatment effect, regardless of whether or not they were *randomised* to receive that treatment), i.e. the **ER assumption**. Given that, despite randomisation, the potential remains for important differences at baseline in small and larger trial, it is possible to adjust for baseline covariates within an RPSFTM analysis, in order to improve baseline balance and increase power (135).

Furthermore, the treatment effect (relative to the duration of treatment) is assumed to be constant regardless of the point at which treatment is received (known as the "**common treatment effect**" assumption). For example, in a trial where treatment switches occur for clinical reasons, the treatment effect is assumed to be the same for those who were randomised to receive treatment (and thus received treatment from randomisation) as for those who only began treatment when switching was clinically indicated (for example, from time of disease progression). This may be unlikely in

reality, given that the treatment effect for those receiving treatment at a more advanced stage in their disease may vary from that experienced when treatment is given earlier (137).

Similarly, the **constant treatment effect** (which gives rise to the title "rank-preserving") implies that the effect of treatment is constant between individuals.

Although it is possible to use multiple parameter versions of the RPSFTM in order to relax the assumption of a constant treatment effect by incorporating more than one treatment covariate, these models typically suffer from a lack of power and do not deliver meaningful causal estimands (94, 95, 138). For example, inclusion of $p > 1$ covariates as in the equation below would require a p-dimensional test statistic $Z(\psi)$, so that, if $p = 2$, the point estimate could be determined as the value at which both the logrank and Wilcoxon tests are zero. However these estimates are likely to be unstable and unreliable (89).

$$U_i(\psi) = \int_0^{T_i} \exp\left[\sum_{j=1}^{p} \beta_j X_{ij}(t)\right] dt$$

It is also possible to adjust for baseline covariates in order to allow more precise estimation of causal effects, but given that each selection of included predictors will alter the resulting causal estimand, it is important to consider this choice carefully. Goetghebeur (139) demonstrates methods to check (both graphically and quantitatively) whether a baseline covariate can be omitted from a model.

White (94) demonstrates how parameter estimates can be used to construct the data sets which, according to the accelerated life model, would have been observed under an alternative treatment scenario, and how this 'counterfactual' dataset can be used to estimate a '**corrected' hazard ratio**. The causal AF estimand can be used to construct

the corresponding counterfactual dataset that, under the assumption of accelerated life model, would have been observed under the "corrected" treatment scenario. These counterfactual data can then be used to estimate the "corrected" HR, if it can be assumed that the survival times follow a Weibull distribution $h(t) = \phi\gamma t^{\gamma-1}$ (where ϕ denotes the scale parameter and γ the shape parameter) (140), as this model provides direct correspondence between the accelerated lifetime model and Cox model, such that the RPSFTM estimand β can be expressed in terms of the Cox parameter θ, as $\beta = \theta/\gamma$. Therefore if one can assume the exponential distribution ($\gamma = 1$), the HR (e^{θ}) and AF (e^{β}) will be equal.

Rank preserving

When analysing RPSFTMs, it is generally necessary to apply certain key assumptions, namely the randomisation, SUTVA and exclusion restriction assumptions. Additional, (or variations of these) assumptions may be required for a particular analysis, depending on the degree of extrapolation required (for example, if one is interested in estimating the causal effect not only in those who received treatment but also to those who do not receive the treatment) or complexities presented by the nonadherence scenario.

For example, when the exclusion restriction assumption is not plausible, it is not possible to estimate the causal treatment effect using SMs without invoking alternative no interaction assumptions, which differ subtly from the exclusion restriction in terms of the use and definition of "randomisation allocation". For the no interaction assumption, Z defines different but observed groups of subjects (i.e. as a conditioning argument) whereas for the exclusion restriction, Z defines different (potential) randomisation conditions for the same individual (112). Again, this assumption may not be realistic: for example, if switches occurs early on for those who, due to underlying

disease state or psychological reasons, are unable to tolerate side effects or perceive treatment inefficacy, these underlying traits may lead to different treatment effects from those who were originally randomised to receive treatment.

5.5.6. Censoring

A major limitation of the RPSFTM is that censoring that is non-informative (on the original T-scale) may become informative on the U-scale through its inherent association with treatment received (89). Consider the uninformative censoring time C_i for individual i (which is assumed to be known for all individuals whether or not their event time was censored; for example, a fixed maximum follow up time for all patients, or an individual maximum follow up time for each patient based on the difference between their date of entry and the final date of follow up, i.e. when censoring occurs for administrative purposes). The censoring time (D_i) for $U_{0i}(\beta)$ is therefore

$$D_i(\beta) = \int_0^{C_i} e^{\beta A_i(t)}\ dt$$

which, through its dependence on treatment received $A_i(t)$, may in turn depend on prognosis.

White (141) explains how this censoring introduces selection bias on the counterfactual scale: in restricting analysis to subjects with known death times, analysis becomes conditional on having an observed survival time less than time C_i, and this survival time (and hence the probability that it is less than C_i and hence observed rather than censored) is assumed to be affected by treatment; a positive or negative treatment effect may increase or decrease the probability that the survival time of an individual is censored, and where treatment crossover occurs, treatment received is likely to be associated with prognosis (142). Thus for those individuals who had an unfavourable

treatment history (for example, control group patients who did not cross over onto a beneficial new treatment, or control group patients that did crossover onto a detrimental new treatment) and experienced events a short time before C_i, a more optimal treatment path may have delayed their event such that it was not observed (i.e. if event had been delayed until after C_i) and as such it would not have been included in analysis. White's proposed solution to remove the dependence of censoring time $D_i(\beta)$ on $A_i(t)$, in order to allow unbiased estimation of the causal parameter, is to restrict analysis to those whose survival time would have been observed whether they had been treated or untreated, in other words by recensoring at the minimum of all possible censoring time over all possible evolutions of $A_i(t)$ **in participant i's allocated treatment group**:

$$D_i^*(\beta) = \min_A D_i(\beta)$$

In particular, if $A_i(t)$ could have been always 0 and could have been always 1, then

$$D_i^*(\beta) = \min(C_i, C_i e^{\beta})$$

such that if $Y_{0i} > D_i^*(\beta)$, then Y_{0i} is replaced by $D_i^*(\beta)$ and its censoring indicator is replaced by 0. This method requires therefore that a potential recensoring time C_i is provided for each individual (for example, as a fixed maximum follow up time for all individuals or a maximum follow up time for each individual based on the difference between their date of entry and the final date of follow up).

The disadvantage to this recensoring approach is the obvious need to ignore some follow up data, with the impact of recensoring depending on the magnitude of the acceleration factor and the distribution of event times (141). Recensoring will be particularly problematic if the treatment effect is not constant over time; in the case of a

treatment-time interaction, recensoring of events that occur late in follow up will distort the overall estimated treatment effect, due to over-representation of earlier compared to later events. As such, White (89) warns that recensoring can lead to distorted conclusions if results are taken at face value without further consideration of the plausibility of the proposed causal effects, especially if the effect of treatment is not constant over time. Furthermore, the changes to original data due to recensoring introduce a undesirable difference between the RPSFTM and original ITT analysis (74).

Alternative solutions to the problem of recensoring include specifying a relationship between U_0 and A_t (but this may well lead to bias if this parametric model is misspecified) or using a method based on inverse probability weighting (89), to be introduced in the next section. Generally, the need to recensor observations is one of a number of limitations associated with SMs; in the next section, the advantages and disadvantages of this methodology will be discussed in relation to the alternative randomisation-based PS techniques.

5.5.7. Limitations of SM/advantages over PS techniques

SM offer greater flexibility than PS methods in terms of the compliance measure or summary of treatment received included in the model, which in turn means a more accurate reflection of the causal contrast of interest. Thus binary, ordinal and continuous measures of compliance can be accommodated in SM via an appropriate functional form for the treatment received variable in the model.

Furthermore, the flexibility of structural models allows extensions to include covariates and investigation of potential interactions between treatment group and baseline covariates (as demonstrated by Robins (143)); therefore these models are useful for assessing whether or not the effect of received treatment is modified by baseline

covariates. Inclusion of baseline covariates can also reduce the variance associated with the causal parameter, as long as the covariates are correlated with both the outcome and treatment compliance (144).

However, SM is generally limited to valid estimation of only one causal parameter, which limits the scope of investigation especially for trials with multiple forms of treatment deviation. When aiming to estimate more than one structural parameter, it would be theoretically possible to include baseline predictors of compliance in order to ensure identifiability; in particular $(k-1)$ baseline covariates are required in order to identify k parameters but these parameter must be good predictors of both compliance and outcome (144); in practice such estimation is not stable or reliable (94).

Additionally, given the reliance on an underlying model structure, SM requires assumptions that once again cannot be empirically validated (114). In particular, a structural model for the effect of treatment received on outcome (for example, in terms of a linear function of duration or cumulative dose of treatment) requires the correct specification of dose-response relationship. Robins (145) demonstrated however, that although these models require specification of such assumptions, they are relatively robust to model misspecification and will provide unbiased model estimates and standard errors even if the functional relationship between outcome and covariates differs to that assumed by the model. Goetghebeur (139) demonstrates graphical methods to assess the appropriateness of the assumed structural model.

Finally, in respecting randomisation (as demonstrated in section 5.5.5), estimated causal effects using SMs will, by definition, have the same significance level as original ITT effect. Even with incorrect model specification, this approach guarantees valid tests of the null hypothesis (of no effect of treatment) (112). This may be considered an advantage (given that interpretation may be unclear if the compliance-adjusted analysis

suggested greater statistical significance than the original ITT analysis) or a limitation (if there is a need to regain the power lost due to deviation from randomised treatments) (94).

Limitations of the RPSFTM in particular, such as the need for recensoring to overcome the problem of informative censoring on the counterfactual survival time scale and the implications of the common treatment effect, will be addressed in detail by considering application of this method to a number of the expository trials introduced in Chapter 4.

However the RPSFTM is only one of a whole class of structural models developed for a variety of outcomes and compliance scenarios in both trial and observational settings.

5.5.8. History of SM for alternative outcomes

An alternative structural AFTM proposed by Robins (131), for observational data, adjusts for time-varying covariates (rather than confounders) as an alternative to the Cox PH model with time-dependent covariates, both of which may be applied, for example, in the case of changing exposure to environmental factors over time. Robins and Greenland (95) extended these methods for a trial setting, thus making use of randomisation as the IV through G-estimation and therefore circumventing the reliance on the assumption of NUC for identification of the causal parameter.

Robins (145) then presented the theory behind the general class of structural nested mean models (SNMMs) for causal estimation in the setting of a repeated continuous outcome, where nested structural models are necessary when treatment exposure *and* confounders vary over time (118). With repeated outcomes (or survival times split into discrete time periods and analysed sequentially using pooled logistic regression, PLR), the basis of structural nested models is that patients contribute to each discrete follow up period for which they remain under follow up (i.e. prior to censoring or event

occurring in the case of survival outcome), such that the trial is considered as a sequential randomised trial, each 'trial' comprising of subjects who remained 'at risk' (uncensored) at the end of the previous 'trial', each of which is nested within the previous trial (130). This approach addresses the bias introduced by time-varying confounders, as each nested 'trial' can be treated independently to all previous 'trial' periods.

Robins (145) proposed general methods (for various outcome types) to estimate effect of treatment received using structural mean models (SMMs). SMMs are a special case of SNMMs for single measure rather than repeated measure outcomes; in the case of a singly measured linear outcome, it is possible to estimate the causal parameter using linear regression rather than through the iterative process proposed by Robins for the general form of SNMM with repeated outcome measures (144).

Robins described additive (and multiplicative) SMMs to adjust for treatment deviations, where the causal parameter is expressed as the difference (ratio) of the expected observed and untreated (relevant control condition) outcome as a linear (exponential) function of unknown parameters, potentially adjusting for baseline covariates. Examples of additive (multiplicative) SMMs include linear (log-linear or logistic) SMMs which contrast linear models for the difference in means (log-ratio of means or log-odds-ratio of success in exposed subsets of the population) of observed and untreated (potential) outcomes (146). These methods were investigated further by Fischer-Lapp (144) and Ten Have (147) for situations where control group patients are and are not (respectively) allowed to switch to experimental treatment, focusing on the effect estimate for those who actually took treatment. Without a constraint ensuring that the probability of treatment success is bounded between 0 and 1, valid estimation of the causal parameter in logistic SMM was not achievable until Vansteelandt (148)

introduced the idea of generalised SMMs, allowing generalisation to incorporate any monotone, continuously differentiable link function (thus allowing for non-linearities by replacing the linear or log-linear links) and incorporating an associational model to appropriately constrain the predicted probability from logistic SMMs.

Robins (149) introduced analogous structural nested models (SNMs) for estimating causal effects of time-dependent treatment in observational settings, which without the advantage of randomisation are identifiable only under the assumption that all relevant confounders have been accounted for (via NUC assumption), once again addressing the problem of time-varying confounding via the use of 'nested' observation periods. SNMs estimate the magnitude of causal effect of a final brief exposure of a time-dependent treatment given in an observational setting at time t as a function of past time-dependent treatment and prognostic factor history.

In a trial setting, however, the balance provided by randomisation can be exploited such that the causal parameters for linear or log-linear SMMs, as well as RPSFTM, can be estimated by G-estimation.

As with this structural modelling approach, which may be applied to observational settings under the assumption of NUC (in place of randomisation), the next causal estimation method to be discussed can be used in both observational and trial settings. Indeed, it was originally developed for use in observational studies, but has since been adapted and applied to randomised data as an alternative method to adjust for treatment deviations. In contrast to the nested approach of SNM, these observational-based methods allow for time-dependent confounders by indirectly adjusting for such variables via the use of inverse probability weighting.

5.6.Observational methods

The basis of the previous two approaches for randomised trial data is that potential outcomes and compliance are independent of randomisation; this independence may in fact be true only after conditioning on covariates, in other words by the NUC assumption (110). Furthermore, the randomisation-based methods discussed in the previous section are generally only useful when adjusting for limited forms of treatment deviations in order to estimate a single treatment effect (for example, experimental versus control therapy) for trials with only two randomised arms. When the causal research question requires estimation of multiple treatment effects (for example, in an active-controlled trial with several randomised treatment arms, or in the case of an equivalence trial where patients may terminate treatment altogether rather than simply switch between treatments, requiring independent estimation of both treatment effects), these randomisation-based estimation procedures tend to be imprecise and unreliable (73). In such cases, it may be helpful to employ alternative methods, developed originally for use in observational studies which rely on the NUC assumption.

Indeed, the distinction between randomised and observational settings tends to blur as follow up lengthens and hence potential for treatment deviations increases (87). It is therefore natural to consider methods originally intended for observational settings for application in trial settings, while remembering that trial data have the advantage of randomisation as an IV (although the strength of this IV weakens with increased treatment deviations over time). Thus the final approach to be considered for causal estimation is an observational approach based on the concept of propensity score weighting.

5.6.1. Inverse probability weighting

An individual's propensity score is their probability of receiving treatment (at a particular time) conditional on pertinent measured confounding variables, usually estimated by regressing treatment received (usually a binary measure, therefore using logistic regression) on all such confounders. This propensity score can then be used as a covariate in a regression model or to stratify or match treatment and control subjects, thereby eliminating bias caused by this set of confounding covariates, when assignment mechanism can be assumed to be unconfounded after conditioning on these observed confounders (119).

However, as demonstrated by the cDAG in section 3.9.1, when covariates vary over time and are affected by previous treatment, as well as influencing future treatment and outcome, they are known as time-dependent confounders and cannot be adjusted for using usual methods (such as regression or standard methods involving propensity score matching). Instead unbiased estimation may be achieved a method known as inverse probability weighting (IPW), which uses propensity scores to simulate a pseudo-population where causal association between treatment received and outcome can be inferred, by excluding or artificially censoring follow up when patients deviate from treatment conditions being assessed. The selection bias introduced by this artificial exclusion or censoring is addressed by weighting the remaining (uncensored) patients by the inverse of their probability of remaining uncensored. These probabilities are determined conditional on all factors that jointly predict outcome and treatment change, thus removing dependence between outcome and censoring. Thus the assumption of NUC (or of sequential randomisation) is key, as is that of **positivity** (that the probability of remaining on treatment is above zero for all possible covariate combinations at each time point) in order to ensure weights (inverse of these

probabilities) are estimable throughout follow up. Under these assumptions, in this simulated pseudo-population, treatment can be considered to be randomly assigned conditional on past covariate history and the treatment effect on outcome is the same as it would be in the original (unweighted) study population (111). This indirect adjustment for time-dependent confounders through their effect on an individual's probability of receiving treatment at any particular time point avoids the confounding and selection biases that result when time-dependent confounders are included directly as variables or via propensity scores in the model.

The reasoning behind this methodology is that, if one is able to determine all factors that jointly predict treatment change and outcome, weighting up the analysis of time to outcome (by the probability of remaining uncensored), while censoring at the point of any treatment change, assigns appropriate weight to individuals with the same characteristics of those previously censored. Given the assumption that all predictors of the *outcome* have been included in the weight-determining model, this weighting provides a pseudo-population that mimics the same likelihood of occurrence of outcome events that would have been observed if none of the patients' follow up had been censored due to treatment changes.

5.6.2. Model type

Like SM, IPW methods are flexible and vary according to whether the compliance measure is binary (i.e. all-or-nothing, as required for PS methods), quantitative (for example, a summary continuous measure such as dose) or time-varying (such as time to withdrawal from randomised treatment). The exact method used will depend, not only on the compliance measure, but also on the outcome type.

In a trial setting where compliance is recorded using a binary variable (receipt/non-receipt of randomised intervention), IPW methods can be used to adjust for selection bias that would be introduced in PP analyses due to informative exclusion of patients who did not receive their assigned treatment. With binary compliance, an individual's propensity score reflects the impact of an individual's time-invariant (i.e. baseline) covariates on their probability of receiving assigned treatment, which is estimated by logistic regression (of the binary treatment receipt variable) on all such confounding baseline variables. The corresponding inverse weights, known as inverse-probability-of-treatment weights (IPTW), are then applied on a patient-specific basis in the outcome model.

When treatment receipt and outcome are time-varying (for example, in a trial with a survival outcome where compliance is defined by time to withdrawal from randomised treatment), time-varying weights (indicating inverse probability of remaining on randomised treatment, and thus uncensored, at each time point) may be calculated using a time-varying Cox (or pooled logistic regression, PLR) weight-determining model; the inverse of these patient-specific time-varying probabilities (known as inverse-probability-of-censoring weights, IPCW) are then applied as time-varying weights for each individual's contribution (while they remain uncensored) in the Cox (or PLR) outcome model.

Similarly, if outcome is recorded at the end of follow up as a binary or continuous response to treatment, and time-varying (binary) treatment is recorded at discrete time intervals, the overall weight for each individual can be calculated by multiplying the probability of receiving treatment at each discrete time point (estimated by logistic regression, regressing on all important baseline and time-varying covariates). Under the assumption of conditional independence, it is unnecessary to allow for correlation from

observations on the same subject, because of full conditioning on treatment history; in other words, all patients included in analysis at time t are known not to have experienced treatment change prior to this point, such that their history is uninformative (and therefore there is no need for clustered standard errors). These patient-specific overall IPTW are then applied to each individual's observations when carrying out logistic or normal linear regression (of binary or continuous response variables, respectively) including in the estimation sample only those patients who fully complied with their randomised assignment.

When compliance is recorded using a continuous measure (such as cumulative dose received), however, the more general approach of IPW estimation using MSMs is required to accommodate the underlying dose-response relationship.

5.6.3. MSMs

The class of marginal structural models (MSMs), originally developed by Robins (111) for observational studies, uses the PO framework to estimate causal effects while simulating ignorable treatment assignment mechanisms using inverse probability weighting (IPW) techniques. Similarly to SNMs, MSMs are based on regression modelling using the PO framework, for example employing linear or logistic regression for continuous or binary outcome data (respectively) to model the relationship between treatment and outcome, in terms of the within-patient difference between a patient's observed and (potentially latent) untreated baseline outcome. However, in contrast to the nested approach of SNM, MSMs (and associated IPW) methods allow for time-dependent confounders by indirectly adjusting for such variables via the use of IPW, thus rendering the treatment receipt exogenous (which, in the context of cDAGs, would be demonstrated by a lack of any arrows *into* the (potentially time-varying)

treatment variable, except potentially from previous treatment); this is achieved by creating a pseudo-population in which treatment history has the same causal effect of treatment history as in the original population but prognostic factors do not predict treatment. Valid causal estimation is then achieved by simply applying standard regression models to the weighted population (111). Thus these models are both *structural* and *marginal,* as they model the structural (or causal) effect of treatment on the marginal distributions of the counterfactual outcomes, $Y_{\bar{a}}$ (rather than their joint distribution, as no account is taken for correlation between different potential outcomes).

The IPCW method is therefore a special case of MSM which can be applied with randomised trial data where compliance to randomised treatment is measured as time to "deviation" from randomised treatment (with "deviation" defined according to the research question of interest). Both MSMs and IPCW methods relate to individual-based treatment contrasts, but IPCW methods compare only two potential treatment histories; for example, in the case of a randomised two-arm trial (experimental versus control treatment) subject to treatment changes, the IPCW would compare treatment with experimental (identified from the weighted experimental treatment arm, censored at treatment change) versus treatment with control (identified from the weighted control treatment arm, censored at treatment change).

Alternatively, MSMs can be used when binary treatment receipt varies over the course of the trial, such that the time following first treatment change is not ignored in analysis; instead the observed time-varying treatment covariate is included in the model (as in a standard time-varying treatment model). However, the bias introduced by use of this time-varying treatment covariate is countered again by the use of inverse probability weights. However, in this case, weights are calculated equal to the inverse probability of

receiving actual *observed* treatment at time k, known as inverse probability of treatment weights (IPTW). Thus, if at each time point k, patient may receive experimental treatment ($A = 1$) or control treatment ($A = 0$) and estimated probabilities (p_{ik}) reflect the chance that individual i receives experimental treatment, at any time point at which they receive *control* treatment, the IPTW would be equal to $1/(1 - p_{ik})$ rather than $1/p_{ik}$.

Furthermore, MSMs are able to compare a wider range of potential treatment histories (for example, the effect of being treated for d days) by modelling causal effects across potential treatment histories; in other words, by making a comparison of counterfactual outcomes resulting from different potential treatment histories in the same individual (for example, assuming a causal effect of d days of treatment equates to βd), where estimation of outcomes under each potential treatment history is obtained by weighting the data according to each patient's corresponding IPTW. Thus, in an MSM, it is necessary to specify an appropriate functional form of relationship between dose and outcome, which again must reflect the causal question.

Different versions of MSM can be applied, depending on the form of the outcome variable, such as linear or log-linear (or logistic) models for continuous or binary (with log or logit links, respectively) outcomes. In assuming a certain functional form (for example, linear) for the relationship between treatment received and outcome, information from those who did not fully comply with treatment is utilised to provide an estimation of the dose-response relationship.

5.6.4. IPW method in practice

The use of IPW methods in practice requires a number of analysis steps, which, given the pertinence of TTE outcomes when considering causal estimation techniques, will be

described below for survival outcomes, under the assumption that compliance is recorded as a time-vaying binary measure (or as time to deviation from randomised assignment), thus requiring IPTW (or IPCW) estimation respectively.

It is necessary first to define the exact treatment regimens to be contrasted, in particular identifying the point at which deviation from these treatment regimens is considered to have taken place. In the case of IPCW, this definition of "treatment deviation" is then used to artificially censor individuals when they deviate from their original regimen. For IPTW models, this definition distinguishes between time spent on experimental and control treatment, thus informing the binary treatment receipt variable. It is then necessary to identify all important baseline and time-varying confounders that are considered to impact, not only outcome, but also the likelihood that an individual will "deviate" in this way from their assigned treatment allocation.

In order to adjust for bias associated with this artificial censoring (in the case of IPCW modelling) or use of time-varying treatment covariate (in the case of an MSM with IPTW modelling), a weight-determining (WD) Cox (or logistic) model is used to calculate the time-varying (or time-fixed) probability that each patient is censored (or excluded), adjusting for all identified important predictors of treatment deviation. The estimated probabilities derived from the WD model are then used to calculate IPCW (or IPTW) for each patient at each time point. The time to event outcome is then assessed (using a relevant model corresponding to type of outcome data) in the population of patients weighted by their IPCW (or IPTW) until the time they were artificially censored due to treatment deviation (or, in the case of binary compliance, only in those who received randomised treatment).

5.6.5. Qualities of weighted pseudo-population

This pseudo-population with w_{ik} copies for each "compliant" patient i at time k, has two important properties which allow valid estimation of causal estimate: first, covariate history does not predict the status of treatment change at time k and secondly the causal effect in this pseudo-population is the same as that of the original population (111). Thus it is valid to use standard regression techniques (such as ordinary time-dependent Cox or linear regression) in this population to obtain unbiased causal estimates (136).

Heuristically, it may be appreciated that weighting the observation for (uncensored) individual i by W_{ik} at time k essentially allows for the loss at or before time k of the $(W_{ik} - 1)$ individuals who shared the same values of all confounding factors ($X_{i0}, \bar{X}_{ik}$) as individual i at time k) but who were artificially censored at time k. Given that, under the assumption of NUC, these $(W_{ik} - 1)$ individuals are also assumed to share the same (time to event) outcome as individual i, the estimated parameter in the weighted population is equivalent to that which would have been observed in the original population but without treatment deviation, and as such may be granted causal interpretation.

In other words, the causal effect of treatment estimated from the pseudo-population is equal to that in the original population if no-one had been informatively censored, under the assumption that censoring occurs at random given covariate history. Thus when exclusion or artificial censoring of patients is applied to those who (or at the times when patients) deviate from their assigned treatment, weighting of the remaining patients creates a pseudo-population of patients who exhibit the same causal risk of event but who did not undergo changes to their treatment.

5.6.6. Need for stabilised weights

Weights are referred to as inverse probability weights because the denominator of the weight is essentially the probability that the individual complies with their randomised treatment (at time k, in the case of time-varying compliance).

However, given that the chance of treatment deviation increases over time, and as such the probability that a patient remains (artificially) uncensored decreases over time, the inverse of this probability can become very large, leading to very high weights for certain patients with prolonged follow up and extreme values of time-dependent prognostic factors strongly associated with outcome. These patients will dominate the weighting analysis by contributing large number of copies to the pseudo-population, leading to unstable parameter estimates with large variability. Using large samples of patients helps to cut down on occurrence of such sampling fluctuations, but in addition Robins and Finkelstein (150) recommend using a "dampening" numerator to stabilize the weights, equal to the probability of being censored at time k dependent on baseline covariates only.

Whereas unstabilised weights are simply the inverse of the conditional probability of having remained uncensored until time t conditional on baseline and time-dependent covariates, stabilised weights are the conditional probability of having remained uncensored until time t given baseline covariates, divided by the conditional probability of having remained uncensored until time t given baseline and time-dependent covariates. The stabilised weight will be equal to 1 for all t if the history of the included prognostic factors for failure does not impact upon the hazard of censoring at t, such that there would be no informative censoring and treatment switching would be at random.

Thus, in an IPCW model, the stabilised weight for individual i, $W_{ik} = 0$ at the time at which patient deviates from their assigned treatment (i.e. such that they are then dropped from analysis); otherwise their (stabilised) weight is denoted

$$W_{ik} = \frac{\Pr(C_{ik} = 0|X_{i0})}{\Pr(C_{ik} = 0|X_{i0}, \bar{X}_{ik})}$$

where C_{ik} indicate whether individual i has been artificially censored (due to treatment deviation) ($C_{ik} = 1$) or remains under observation ($C_{ik} = 0$) at time k, X_{i0} are fixed baseline covariates, X_{ik} are time-varying covariates at follow up time k, assumed to influence both treatment crossover and final outcome and $\bar{X}_{ik}$ is the history of these relevant time-varying covariates up to follow up time k.

5.6.7. IPW methods to adjust for LTFU

As well as adjusting for treatment deviation, IPW methodology may be used to account for bias introduced by informative loss to follow up; for example, when (complete case) pseudo-ITT analyses are carried out, the bias caused by missing outcome determination can be removed using IPW methods, assuming that data are available on all time-varying determinants of loss to follow up. Furthermore, it is possible to adjust for treatment deviation as well as loss to follow up in the same analysis, by simply multiplying a patient's time-varying "treatment deviation" weights by the corresponding "loss to follow up" weights to obtain overall weights (based on assumption of conditional independence) (151).

5.6.8. Considerations regarding use of IPW methods

There are a number of important limitations that must be considered when using IPW techniques.

First, the correlation introduced by the use of IPWs renders normal standard errors invalid; as such, robust variance estimation or bootstrapping is required for confidence interval estimation.

IPWs are unestimable if there are levels of any covariates that ensure (in order, where the probability equals one) that treatment switching will occur (152). Indeed, IPW methods require that there remains a reliable pool of uncensored patients on which to base weight estimation at each time point and for each combination of confounding covariates. If the probability of treatment deviation for any such combination is especially high (or low), the stabilised weights will be subject to wild fluctuations, in turn leading to unreliable causal estimates.

And finally, but perhaps most importantly, especially in the context of randomised trials, is the assumption of NUC. In order for IPW methods to be implementable, it is vital for data to be available on all potential confounders (in other words, those baseline or post-randomisation, potentially time-varying, variables that both predict treatment change and outcome). It is never possible to be entirely confident that all data on all confounders have been collected and correctly allowed for; indeed it is highly unlikely for the NUC assumption to be entirely true, but IPW methods will work well even when this assumption is only approximately true (142), provided no important confounders have been omitted from the model.

However, the NUC assumption may represent a key limitation of the IPCW, particularly in a trial context. IPW methods were developed for observational studies which are typically more suited to the longitudinal data collection process required to meet this assumption. In contrast, datasets collected for trials are typically much smaller than those for observational studies, and key (time-varying) predictors of treatment switching are often not collected in trials (for example, reasons for patient switching

preferences). Personal or psychological traits which affect a patient's ability to comply (and in turn affects their outcome) may be very difficult to record reliably longitudinally, especially in the strict setting of a randomised trial (while observational studies may provide more scope for long term follow up of such factors). Moreover, data collection on such indicators is often stopped during follow up for practical reasons (for example, upon treatment discontinuation or disease progression), thus further hampering the applicability of the IPCW method. Overall, when data on predictors are sparse, the NUC becomes less plausible and the IPCW method becomes less stable (142).

5.7. Considerations regarding choice of adjustment method

Having introduced the reader briefly to the field of causal analysis, it is useful to reflect briefly on the requirements and limitations of each method, in order to ensure appropriate application and interpretation, as well as to guide researchers in their choice of causal methodology.

Table 7 Main differences between causal analysis approaches

1. Key to causal analysis is appreciation of the **PO framework**, which underlies most causal estimation techniques.

2. Traditional **SEMs** are disadvantaged by **lack of transparency** regarding underlying assumptions and the likely problem of hidden bias and time-varying confounding.

3. **Principal stratification** techniques are straightforward to apply in simple, all-or-nothing compliance scenarios but **cannot be extended** easily to allow for (more usual) **time-varying treatment** deviation or survival outcomes.

4. **Structural models** provide a more flexible approach to causal analysis, accommodating more general compliance measures and using a nested approach to

adjust for time-dependent confounding, but are restricted to causal estimation of **single causal effects** and thus cannot be used to adjust for more than one form of treatment deviation.

5. The assumption of **NUC** underlies the observational-based approach of **MSM**s (and associated **IPCW methods**), which account for time-dependent confounding by the use of inverse probability weighting; thus the key requirement for unbiased IPW estimation is **availability of data on all relevant confounders**, which may present a particular challenge in trial settings.

5.8.Conclusion

The overview of statistical methods to adjust for deviation from treatment protocol in randomised trials and observational studies summarised in this chapter has revealed a number of available approaches to causal estimation, with a particular focus on survival analysis methodologies, given that settings typically associated with treatment deviations (such as the necessarily dynamic treatment regimens in the treatment of cancer or longterm treatment for chronic disease) often use TTE outcomes to assess treatment efficacy. Having determined that such methodologies exist, a review of recently published trials is summarised in the next chapter, in order to determine the extent to which such treatment deviations actually occur in trials and whether trialists are aware of the available causal analysis methods in order to appropriately adjust for such treatment changes. The thesis will then move on to explore in more detail the practicalities and challenges arising from the application of the two key methods identified in this review of causal methods for adjustment for deviations to randomised trials when analysing survival outcomes, namely the IPCW and RPSFTM.

6. Review of nonadherence in published RCTs

6.1.Introduction

In previous chapters, it has been demonstrated that nonadherence is a problem in general clinical practice as well as trials, with specific examples given of departure from randomised treatment in (a select group of) trials and a discussion of the resultant complications in trial conduct and analysis and potential statistical methods to appropriately handle such departures in analysis. This chapter now describes a review of recently published RCTs carried out with the aim of determining the extent to which nonadherence occurs in randomised trials, as well as the degree to which trialists appreciate the importance of the issue of nonadherence, in terms of the quality of both their reporting of nonadherence and use of appropriate methods to adjust for such treatment deviations.

6.1.1. Reporting of adherence information

It is well recognised that clinical trials often fail to report important features of design and analysis that are necessary to ascertain the methodological quality of the trial (153-159). Lack of adequate trial reporting led to the development (and two further updates) of the CONSORT (*Con*solidated *S*tandards *o*f *R*eporting *T*rials) statement (41, 160, 161), which aimed to improve reporting of trial methodology and results by providing authors with a checklist of items recommended for inclusion in trial publications. CONSORT recommends reporting "for each group the numbers of participants who were randomly assigned, received intended treatment and were analysed for the primary outcome.... Knowing the number of participants who did not receive the intervention as allocated or did not complete treatment permits the reader to assess to what extent the estimated efficacy of therapy might be underestimated in comparison with ideal circumstances." Use of CONSORT flow diagrams to illustrate patient flow through the trial has become increasingly common (162), but the information presented in the flow diagram may not be detailed enough to ascertain the true extent and nature of any nonadherence to treatment protocol that have occurred in a trial.

6.1.2. Aims

As such, this review aimed to ascertain

1) the extent and nature of nonadherence to randomised treatment reported in published trials, 2) how such nonadherence is handled in statistical analyses of benefit and harms outcomes, and 3) how well such nonadherence, and the statistical methods to deal with it, are reported in published RCTs.

6.2.Methods

6.2.1. Terminology

In this review, the term "nonadherence to randomised treatment" is used to indicate any deviation on the part of patient or treatment provider to the trial treatment protocol, or any treatment change agreed with medical staff but not permitted by the trial protocol. Protocol-permitted treatment changes can also present problems of interpretation but these were not considered here. For example, discontinuation of randomised treatment for safety reasons as specified in the protocol, or premature stopping of a trial according to a predefined stopping rule, were not be considered as examples of deviations from treatment protocol in this review.

For the purposes of assessing adequacy of trial reporting, the definitions given by Vrijens (3) for the three quantifiable phases of patient adherence to prescribed medication (initiation, discontinuation and implementation) were extended for use in a clinical trial setting, where the intervention could be nonpharmacological and where adherence to treatment protocol can be influenced by treatment providers when administering or prescribing treatment as well as by participants, because of the required adherence to the trial treatment protocol.

Thus adherence is defined as the degree of correspondence between a participants' intended randomised treatment prescription and their actual history of treatment received over the course of the protocol-specified treatment period, and is split into the following phases: *initiation* of randomised treatment occurs when the participant begins their randomised intervention and *discontinuation* is defined as the permanent cessation of randomised intervention, with the intervening time being defined as the *implementation*

period. Persistence is defined as perseverance with randomised treatment as per the trial protocol until the end of the protocol-defined treatment period.

Given that adequacy of trial reporting on adherence to treatment protocol depends on the duration and complexity of trial specific interventions, trials were classified according to whether the trial interventions were given at a single or multiple time points. Trials with treatment given at multiple time points (referred to as "longitudinal treatment periods") were divided into "short term intervention trials" involving treatment given at a few discrete time points (for example, single daily dose over the course of a week) or continuously over a short period of time (for example, infusion given for one hour), and "long term intervention trials".

6.2.2. Benefits versus harms outcomes

Analyses of harms data are particularly affected by departures from treatment protocol because of the anticonservative nature of ITT analyses for harms outcomes. Therefore it was of interest to investigate the handling of benefit and harms outcomes separately in this review; as such, outcomes were categorised according to whether they were assessing benefits or harms.

Outcomes related to harms (for example, mortality or a composite cardiovascular outcome – including events such as stroke, myocardial infarction and death) but which were likely to be caused by the disease itself rather than side effects of the treatment (and therefore the purpose of the drug was to prevent such events) were categorised as benefit outcomes.

6.2.3. Key points to be reported

In order to judge the quality of trial reporting, and statistical handling of treatment protocol nonadherence, it was important to identify key points that should be explicitly reported in each publication. Table 8 presents recommendations for explicit reporting of information relating to adherence to treatment protocol.

All trials should report the number of randomised participants and the number of participants who initiated their randomised intervention. Longitudinal treatment period trials were also expected to have reported on treatment persistence: in the interest of conservative assessment of adequacy of reporting, short term intervention trials were required only to have reported on the number of patients who completed intervention (or the number who discontinued randomised intervention prematurely). Long term interventions are, however, more likely to result in treatment interruptions or deviations, so patients who were still taking treatment at the end of the protocol-determined treatment period may not have been fully adherent to the treatment schedule for the whole duration of treatment. Thus publications of trials with long term interventions were judged according to whether they had supplemented their reporting of the number of participants who persisted with intervention according to the treatment protocol (or conversely the number of participants who prematurely discontinued randomised intervention) with some measure of participant and/or treatment provider adherence to treatment protocol (as appropriate, depending on whether administration of treatment was by the participant or the treatment provider) over the implementation period.

Table 8 Recommendations for explicit reporting of information relating to adherence to treatment protocol

Quoted justification is taken from CONSORT 2010 elaboration document (41) unless stated otherwise.

	Report for all trials	Report according to treatment duration			Report reasons for	Justification
		Single (one-off) intervention	Short term intervention	Long term intervention		
1. Randomisation	[1] Randomised					"Crucial count for defining trial size and assessing whether a trial has been analysed by intention to treat"; necessary to determine whether all trial participants received treatment and were included in analysis.
2. Adherence to treatment protocol						
a. Initiation		[1] Initiated (or received) randomised intervention	[1] Initiated randomised intervention	[1] Initiated randomised intervention	Those not initiating randomised intervention	"Knowing the number of participants who did not receive the intervention as allocated or did not complete treatment permits the reader to assess to what extent the estimated efficacy of therapy might be underestimated in comparison with ideal circumstances."
b. Completion/ Persistence[2]			[1] Completed randomised intervention	[1] Persisted with randomised intervention as required by treatment protocol	Those who did not complete / persist with randomised intervention	
c. Adherence over treatment period [3]						
i) Method			Description of method used to measure adherence over treatment period (and of an additional method to check reliability if trial involves participant-administered intervention) [4]			If participant compliance data are collected, the reliability of the method used to record compliance should ideally be checked by use of another method (for example, treatment diaries backed up by counts of remaining tablets at the end of each course of treatment) (14)
ii) Justification for definition			Justification for any reported definition of adherence (for example, if a threshold is used to define adequate adherence) [4]			"If patients are to be divided into "compliant" and "noncompliant" groups, the division should ideally be made on the grounds of the relationship of the compliance level to the therapeutic response or outcome" (11)
iii) Results			Measure of participant and/or treatment provider adherence with randomised intervention (as appropriate) [4]		Any participant or treatment provider nonadherence [4]	
3. Analysis						
a. Analysed	[1] Analysed				Any exclusion of participants from analysis	"Attrition as a result of loss to follow up, which is often unavoidable, needs to be distinguished from investigator-determined exclusion for such reasons as ineligibility,

	Report for all trials	Report according to treatment duration	Report reasons for	Justification
				withdrawal from treatment, and poor adherence to the trial protocol... Participants who were excluded after allocation are unlikely to be representative of all participants in the study."
b. Analysis set composition	How analysis sets differ from randomised groups		Any difference between analysis sets and randomisation groups	"Erroneous conclusions can be reached if participants are excluded from analysis, and imbalances in such omissions between groups may be especially indicative of bias."

[1] Report numbers of participants in each randomised intervention group satisfying condition listed in each cell.
[2] Persistence is defined as perseverance with prescribed treatment until the end of the treatment period.
[3] Adherence is defined as a measure of the degree of correspondence between prescribed treatment and actual treatment received by participant.
[4] Note that, depending on the complexity of treatment, it may not be necessary for trials with short term intervention to report on adherence over the treatment period.

Methods used to assess participant or health care provider adherence to treatment protocol should be described, and clinical justification given for any definition used to define adherence. Ideally clinical trials involving participant-administered intervention should use a reliable measure, and ideally a combination of measures, to accurately record participants' adherence data (14). Therefore, publications of such trials with a long term intervention period in this review were judged according to whether they collected and reported details of the methods used to assess, and check reliability of, participant adherence.

It was also recorded whether any explanation was given regarding the use of any adherence threshold to split participants into "good" and "poor" compliers. Although such dichotomisation is not recommended (20), any threshold used to create such groups should be specified in the protocol; otherwise, there may be suspicion that an optimal adherence cut-off has been selected on the basis of the results to allow the most favourable adherence rates to be reported, as the "definition of non-compliance is malleable and could be inadvertently manipulated for benefit of investigators" (35).

Analysis sets for both benefits and harms outcomes should be defined explicitly, with reasons given for the exclusion of any patients from analysis. If there is any difference between the defined analysis sets and the intervention groups as randomised, this should be stated clearly and any potential resultant bias should be discussed.

6.2.4. Missing data

The issue of missing outcome data is related to non-compliance, as the reason for missing data may well be related to nonadherence to treatment: those who withdraw from treatment may also withdraw from follow up (68). Thus when trialists apply methods to overcome the problem of missing data, this may be viewed as an attempt to

also deal with the issue of non-compliance, especially if outcome data are missing entirely as a result of treatment cessation. Thus it was noted, as part of this review, whether trialists reported missing outcome data and any methods used to adjust for this missingness.

In particular, it was noted whether sensitivity analyses were performed, whereby missing outcome values are imputed assuming various missingness patterns (for example, MCAR or MAR, i.e. dependent on cofactors), recommended as a method to assess the robustness of conclusions from complete case analyses (163).

6.2.5. Selection of reports

A search of Medline (using terms randomi$ed controlled trial$ or controlled trial$ or controlled clinical trial$ or RCT$) was carried out in order to identify trial reports published in the British Medical Journal (BMJ), Journal of the American Medical Association (JAMA), Lancet and New England Journal of Medicine (NEJM) during 2008. Of the 698 articles obtained from the initial search, 281 (49 from BMJ, 53 from JAMA, 84 from Lancet and 95 from NEJM) remained following deletion of duplicates, comments, systematic reviews and meta-analyses. A sample of 100 trial reports were randomly selected from these articles, 16 of which were from the BMJ, 20 from the Lancet, and 32 each from the Lancet and NEJM.

6.2.6. Secondary publications

Five of the selected studies were secondary publications of recent trials (original publications were in NEJM, American Journal of Medicine and Neurology). In these cases, for the purposes of reporting adherence data, the combination of both the primary and secondary publications of the same trial were considered as the unit of analysis.

6.2.7. Data extraction

Characteristics of the trial publications were recorded using a piloted, standardised form. In cases of any doubt or ambiguity, the paper was reviewed by a second reader. Scrutiny of each trial publication allowed determination of

- whether a CONSORT flow diagram was presented, and if so, what level of detail regarding deviation from treatment protocol was provided
- whether procedures were put in place to ascertain and improve patient compliance with the treatment protocol
- whether patient-related adherence data were collected and whether any attempt was made to check on the reliability of methods used to record compliance (for example, treatment diaries backed up by counts of remaining tablets at the end of each course of treatment)
- the extent and nature of nonadherence to treatment protocol that was reported in the trial
- method of benefit analysis used (for example, ITT or PP) and its appropriateness, given the aims, design and hypothesis of the trial
- whether harms analyses appropriately allow for departures from treatment protocol (for example, when patients did not receive their allocated treatment or switched treatments)
- any methods used to specifically adjust for, or examine effects of, departures from treatment protocol
- extent of missing outcome data and methods to deal with missing data

6.3.Results

6.3.1. Characteristics of RCTs

The characteristics of the 100 randomly selected randomised controlled trials are shown in Table 9. The vast majority (92%) of the studies in the sample included a superiority hypothesis only. Ten trials were randomised at a cluster, rather than individual, level. Only one of these was a drug trial; the remaining involved nonpharmacological interventions. Sixty nine trials involved a drug treatment; the majority of which were labelled as (or characteristic of) phase III trials, and 44 (63.7%) of these drug trials used a placebo for the control arm. Only one nonpharmacological trial (which aimed to investigate various aspects of the placebo effect using placebo acupuncture) involved a placebo.

Table 9 Characteristics of 100 randomly selected RCT articles

Characteristic		Number (=%)	
Journal			
BMJ		16	
JAMA		20	
Lancet		32	
NEJM		32	
Hypothesis			
Equivalence		1	
Non-inferiority		7	
Superiority		92	
Design			
Cluster		10	
Parallel		90	
(of which factorial)		(11)	
Blinding			
None		29	
Single		16	
	Investigator only		1
	Assessor only		13
	Laboratory staff only		2
Double		55	
	Patient and investigators[1]		26
	Patient and assessors		2
	Investigator and assessors		1
	Statistician and assessors		1
	Laboratory staff and assessors		1
	Patients, investigator and assessors[2]		22
	Patients, statistician and assessors		2
Prespecified statistical analysis plan			
Stated		15	
Not stated		85	
Intervention type[3]			
Nonpharmacological intervention		35	
	Surgical		3
	Behavioural		14
	Device		7
	Model of care/guidelines		7
	Other		4
Drug		69	
	Phase I or II		4
	Phase III[4]		65
Control intervention			
Active intervention		55	
Placebo		45	

[1] Note that this includes 12 placebo trials which were described as "double-blind" but did not explicitly state that the patient and investigators were blinded.
[2] Note that this includes two placebo trials which were described as "double-blind" but did not explicitly state that the patient and investigators were blinded.
[3] Four trials involved two types of intervention.
[4] Forty-nine of these did not specifically state "phase III" but were powered and controlled comparisons indicative of phase III.

6.3.2. Extent and nature of reported nonadherence to treatment protocol

Ninety-eight publications reported at least one form of departure from treatment protocol (see Table 10). Direct comparison of the extent of departure from treatment protocol across trials is not straightforward, as trials differed greatly in terms of type and duration of intervention, definitions used to define nonadherence and level of reporting. However the distribution of percentage of patients displaying some form of deviation from treatment protocol (or the average degree of nonadherence in a single trial), based on information reported in the trial publications, can also be seen in Table 10.

Table 10 Reported forms of nonadherence from protocol

	Number (=%)
Reported some form of nonadherence to treatment protocol[1]	**98**
Participants who did not initiate allocated treatment	39
Participants with incomplete treatment (among those who initiated allocated treatment)	78
Participants who switched trial treatments	12
Participants who started open label treatment (not as per protocol)	7
Participants who started disallowed/non-trial treatment	4
Evidence of contamination between treatment groups	3
Other forms of nonadherence to treatment dose/schedule	23
Nonadherence on the part of treatment providers	12
Did not report any nonadherence to treatment protocol	**2**
Percentage of patients experiencing/displaying some form of nonadherence to treatment protocol	
None reported	2
0-5%	23
5-10%	10
10-20%	22
20-30%	11
30-50%	12
>50%	9
>0% but unclear[2]	11

[1] Trial publications may have reported the number of participants in more than one of the categories listed.

[2] For example, trial report states only that the treatment providers were not adherent to some degree, or unable to distinguish withdrawal from treatment for legitimate reasons (for example, death or treatment changes permitted by protocol following adverse events) from withdrawal due to nonadherence.

6.3.3. Quality of reporting on departures from treatment protocol

The vast majority of trial reports (96%) included a CONSORT flow diagram. Table 11 summarises the quality and completeness of reporting on randomisation, adherence to treatment protocol and analysis in the CONSORT flow diagrams and the text. All 100 trials stated the numbers randomised, but only 58 publications stated how many patients actually initiated their allocated treatment. All trials provided some information on the number of participants included in analysis of the primary outcome, but this information was not always provided for secondary outcomes, particularly when a large number of outcomes were analysed. Forty-three trial reports included an explicit explanation of the composition of the analysis sets used for benefit outcomes, 48 trials labelled the analysis sets (47 "ITT" and 1 "PP") without further explanation of how the analysis sets were composed, and no details on the composition of benefit outcome analysis sets were given in the remaining nine trials.

Table 12 provides a breakdown of persistence and adherence information reported in the 88 studies with longitudinal treatment periods. The majority (81, 92%) provided some information on treatment completeness, but this was sometimes incomplete or vague.

Overall, reporting of treatment initiation and completeness was judged to be adequate in only 7 (11%) of 66 trials with long term interventions (requiring reporting of treatment initiation, persistence and some measure of adherence over the implementation period) and 8 (36%) of 22 trials with short term interventions (requiring reporting of treatment initiation and completion). Reporting of treatment receipt was judged to be sufficient in 10 (83%) of 12 trials with a one-off treatment.

Table 11 Reporting of key points in 100 trial reports

	Report for all trials	Report according to treatment duration		
		Single (one-off) intervention (n=12)	Short term intervention (n=22)	Long term intervention (n=66)
1. Randomisation	100 (100%)	-	-	-
2. Adherence to treatment protocol				
a. Initiation	-	10 (83%)	15 (68%)	33 (50%)
b. Completion/ Persistence	-	-	12 (55%) [17 (77%)][1]	31 (47%) [51 (77%)][1]
c. Adherence over implementation period				
i) Method	-	-	1 (50%)[2]	21 (47%)[2]
ii) Justification for definition	-	-	0 (0%)[3]	0 (0%)[3]
iii) Results	-	-	5 (23%)[4]	22 (33%) [28 (42%)][4]
3. Analysis				
a. Number analysed	100 (100%)[5]	-	-	-
b. Analysis set composition	43 (43%) [91 (91%)][6]	-	-	-

[1] Number (%) of trials that fully reported [partially reported] on persistence or completion of randomised treatment.
[2] Number of short term (2) or long term (45) intervention trials with patient-administered treatment (2) is used as denominator for %.
[3] Number of short term (1) or long term (18) intervention trials with patient- administered treatment that reported adherence definition is used as denominator for %.
[4] Number (%) of trials that fully reported [partially reported] some measure of nonadherence on the part of patient or treatment provider.
[5] Number (%) of trials that reported the number included in analysis of primary outcome.
[6] Number (%) of trials that fully reported [partially reported] analysis set composition; partially reporting trials stated that analysis was by "intention to treat" (n=47) or "per protocol" (n=1) but did not explicitly explain composition of analysis sets.

Thirty-three trials with intervention given at multiple time points reported information on treatment interruptions (two trials) or a measure of average adherence over the treatment period, on the part of the participant (28 trials) and/or treatment provider (four trials; one trial reported on adherence of both the participant and treatment provider), but six of these reported this information for the whole trial or combinations of intervention groups rather than by individual intervention group. Reported measures included the percentage of patients who achieved a particular arbitrary level of adherence in terms of the proportion of doses received or the proportion of time patients were supplied with randomised drug over the course of the trial.

Table 12 Breakdown of persistence and adherence reporting in 88 trials with longitudinal intervention periods

	Short term intervention (n=22)	Long term intervention (n=66)	Total
Persistence			
Fully reported[1]	12	31	43
Partially reported only[2]	5	20	25
Partially reported, including reporting the number of participants who:			
Withdrew	1	3	4
Withdrew consent	1	2	3
Lost to follow up (during treatment)	-	3	3
Lost to follow up (unclear whether during treatment)	1	1	2
Discontinued due to certain event(s)[3]	-	6	6
Completed study	3	8	11
Discontinued study	-	2	2
Completed different aspects of treatment protocol (reported separately)[4]	-	2	2
Completed treatment in trial overall (not by treatment group)	1	1	2
Not reported	5	15	20
Adherence over implementation period			
Fully or partially reported, including reporting[2]	5	22 (6)	27 (6)
Average measure of participant adherence [5]	2	20 (6)	22 (6)
Average measure of adherence on part of treatment provider	3	1	4
Treatment interruptions	-	2	2
Not reported	17	38	55
Overall reporting			
Some reporting x	19	62	81
Persistence reported only (fully or partially)	14	34	48
Adherence reported only (fully or partially)	2	11	13
Both persistence and adherence reported (fully or partially)	3	17	20
Not reported	3	4	7

[1] Reported the number of participants still taking treatment at end of treatment period, or the number who "completed" treatment, or who discontinued or withdrew from randomised intervention prematurely

[2] Trial publications may have reported on one more than one of the categories listed.

[3] Reported numbers discontinuing only for one reason (e.g. adverse events).

[4] Unable to discern how many participants received entire intervention.

[5] Figure in brackets indicates number of publications reporting adherence measure for overall trial or combinations of intervention groups, not by individual intervention group.

6.3.4. Ambiguities in trial reports

Commonly used terms in CONSORT flow diagrams alluding to nonadherence (or adherence) to treatment protocol such as "discontinued", "completed study protocol", "withdrew", "protocol deviations" and "loss to follow up" do not provide explicit information on completeness of treatment unless accompanied by clarification on timing or treatment actually received. For example, the term "withdrew" can indicate withdrawal from treatment only, withdrawal from further follow up or withdrawal of consent regarding inclusion of a patient's data in the study. In 13 (62%) of the 21 trials that included the term "withdrew" in the flow diagram, it was not possible to ascertain whether the participants who "withdrew" had actually initiated treatment before withdrawing. Similarly, the timing of withdrawal was not clear in 8 (38%) of the 21 trials that described participants who "withdrew consent" in the flow diagram.

Fifteen (18%) of the 85 trials with a longitudinal treatment which presented a flow diagram referred to the number of participants who "received" when they meant "initiated" treatment; this could potentially mislead readers as use of the word "received" may be incorrectly interpreted as receipt of the entire intervention.

Three trials with a long term intervention period referred in the CONSORT diagram (and 4 more referred in the text) to the number of patients who "completed" treatment when it would have been more accurate to report this figure as the number "still on treatment at end of treatment period/trial follow up", as completion of treatment may imply complete adherence to, as well as persistence with, randomised intervention throughout the treatment period.

6.3.5. Methods to check adherence

Less than half (22, 47%) of the 47 trials with participant-administered treatment mentioned checking adherence, and only 5 of these assessed reliability of the adherence data using a second method. The most common methods to ascertain participant adherence involved counting (or weighing) pills (11 trials) or participant self-reporting (10), even though these methods are not considered to be reliable (14). More accurate methods such as medication events monitoring systems (MEMS) or measurement of drug metabolite or marker in bodily fluids were not used, except in one trial (which used blood tests and laboratory assays).

As shown in Table 13, of the 69 trials that included a drug intervention, 17 (24.6%) included a definition of compliance (or non-compliance). All but three of these definitions were based on the percentage of prescribed medication received by the individual. Four trials simply described these as average percentages for the treatment groups separately (three trials) or for the entire trial population overall (one trial), while ten other trials specified a percentage cut-off rate indicating good compliance. The most common cut-offs were two thirds (three trials) or 80% (three trials) of treatment received, but other cut-offs included 50%, 60%, 75% and 100%. Three drug trials based their definition of compliance on factors other than simply the percentage of medication taken.

Trials of nonpharmacological interventions also reported on compliance, but the complex nature of the interventions and pragmatic trial designs meant that less straightforward and potentially ambiguous definitions of compliance were more commonly reported than in drug trials. Definitions of compliance on the part of the participant in nonpharmacological trials were much less common than in drug trials: only two (5.7%) of the 35 trials that involved nonpharmacological interventions defined

compliance on the part of the participant. One defined good compliance as attendance of at least five out of six possible therapy sessions, and the other defined adherence as having undertaken at least 80% of the required blood monitoring.

Nonadherence is not only an issue at the participant level, as treatment providers may also be unable or unwilling to adhere to treatment protocol. Twelve trials which involved administration of treatment by personnel other than just the participant mentioned an assessment of whether the treatment providers were adherent to the protocol. Most of these twelve trials simply discussed the fact that aspects of delivered treatment had been recorded (for example, using case report forms or program content checklists, videotaping treatment sessions, or recording information on a study web server) and thus it was determined that treatment providers had deviated from randomised treatment protocol in some manner. Two cluster trials discussed the rate of acceptance, participation or implementation of treatment guidelines by site staff.

Table 13 Definition of compliance in drug trials

	Number (% of 69 drug trials)	
Reported definition of compliance	**17 (24.6)**	
Based on percentage of medication received	14	
Reported percentage of medication consumed		4
By treatment group	3	
For trial overall	1	
Cut-off used to define good/poor compliance		10
50%	1	
60%	1	
2/3	3	
75%	1	
80%	3	
100%	1	
Based on discontinuation of medication	3	
Non-compliance defined as discontinuation of medication for any time period[1]		1
Percentage of time at risk that the participant spent *in receipt* of trial drug (<75%, 75-90%, >90%)[2]		1
Split into full/partial/total non- compliance based on timing of multiple courses of drugs[3]		1

[1] Documented discontinuation of trial drug for specific reasons rather than sporadic failure to take treatment.
[2] Based on evidence of collection of trial drug rather than actual consumption
[3] Dependent not only on the cooperation of the participant but also of the treatment provider, as the drugs were administered during outpatient visits.

6.3.6. Analysis populations

A benefit analysis was carried out in all trials. Nine trials did not define any benefit analysis populations, and two additional trials defined analysis populations but did not label them as ITT or PP.

Of the 91 trials that defined some benefit analysis population, 73 carried out analysis on one population set only (67 ITT, two modified ITT, two PP and two unlabelled populations), 15 analysed two populations (12 ITT/PP combinations, one ITT/modified ITT and two modified ITT/PP combinations) and three analysed three populations (one ITT/modified ITT/PP and two ITT/PP/PP compliant combinations).

Eight-seven trials included what was referred to as an ITT or "modified" ITT analysis. Of the 83 "ITT" analyses presented, 15 excluded patients if they did not initiate randomised treatment (and an additional three excluded treatment protocol violators), four included only eligible participants, and seven specifically stated that they only included those with a certain amount of follow up. In the six trial reports which stated that a "modified" ITT population would be analysed, specified requirements included receipt of treatment (five reports), availability of a certain amount of follow up data (four reports), complete follow up (one report) and eligibility of participants (one report).

Of the 19 trials that specified a PP population, only one did not specify any details of the composition of the population, while one specified that participants would be analysed "as treated". Fourteen of these 19 trials included a requirement of some degree of compliance with treatment protocol, but none provided any clinical justification for the threshold of received treatment used to define "compliance".

Table 14 Analysis populations

Benefit analysis population[1]	Number (=%)	
"ITT"	83	
	No details	47
	"Included all randomised individuals"	11
	"Analysed as randomised, irrespective of adherence"	10
	Other[2]	15
	Included requirement of "≥1 dose/received treatment"	12
	Included requirement of "at least one assessment/were followed up/ excluded those who died"	5
	Included only eligible patients	1
	Included only those with complete follow up	1
"Modified ITT"	6[3]	
	Included requirement of "≥1 dose/received treatment"	5
	Included requirement of "at least one assessment/ primary outcome data available"	4
	Included only eligible patients	1
	Included requirement of complete follow up	1
"PP"	19[3]	
	No details	1
	"As treated"	1
	Included compliance with treatment:	14
	≥1 dose	2
	complete adherence	6
	taking some proportion of the drug	6
	Included only eligible patients	3
	Included requirement of "compliance with protocol"	7
	Included requirement of data available/complete follow up	5
"PP compliant"	2	
	Included only those who used intervention, did not violate treatment protocol and had outcome data available; censored follow up when interruption in treatment	1
	Censored follow up when proportion of drug taken fell below 60%, or stopped trial medication and started non-trial medication to treat condition	1
Unlabelled	2	
	Included those who received ≥1 dose of assigned study drug	1
	Included those who completed all follow up	1
No details of population	9	

[1] Some trials defined more than one efficacy analysis population.
[2] Some of these trials fitted more than one of the "other" categories.
[3] Some of these trials fitted more than one of the listed categories.

6.3.7. Handling of departure from treatment protocol in statistical analyses

Of the 98 trials that reported some form of nonadherence to treatment protocol, 51 (52%) reported some form of analysis method to handle such nonadherence in the analysis of benefit outcomes (see Table 15). Forty-six trials adjusted for treatment protocol deviations by carrying out analysis based on PP analysis (total not shown in Table 15), by censoring or excluding participants who had violated the treatment protocol in some way, but half of these analyses were labelled as ITT or modified ITT analyses. Only one trial (164) aimed to address the bias introduced from potentially informative censoring of patients at the point of deviation from treatment protocol, by weighting their censoring by the inverse of their estimated probability of adhering, as advocated by Robins (150). Other analysis methods that dealt with departures from randomised treatment included treating discontinuation of treatment (or starting disallowed or rescue medication) as a treatment failure in analysis (three trials), AT analyses (analysing participants according to the actual treatment received regardless of randomisation allocation) (three trials) and analysing outcomes concerned with time to discontinuation of trial drug (four trials).

Of the 20 trials where adjustment for nonadherence was explicitly compared to other analyses, four trials reported that adjustment for nonadherence resulted in more extreme treatment effects; in the remaining 16 trials it was not reported to have made a substantial difference to conclusions.

Table 15 Statistical methods addressing nonadherence to treatment protocol

	Number (%[2])	Definition of analysis se	Number
Reported a statistical method addressing nonadherence to treatment protocol[1]	51 (52)		
Variant of PP			
Primary PP analysis described as "PP"			18
Included only those participants who received full randomised intervention			8
Included only those participants who received at least one dose of randomised intervention			2
Minimum degree of adherence required			5[3]
Included only those taking treatment at particular time during trial			1
Excluded participants if they started disallowed medication			1
Censored participants at time of stopping treatment			1
Primary PP analysis described as "ITT" or "modified ITT" analysis			23
Included only those participants who received full randomised intervention			1
Included only those participants who received at least one dose of randomised intervention			16
Included only those participants who received the single treatment			3
Excluded participants if they deviated from treatment protocol			3
Sensitivity analysis			12
Included only those participants who received full randomised intervention			4
Excluded participants if they received disallowed treatments			2
Minimum degree of adherence required			2[4]
Censored participants at the point of deviation from treatment protocol			4[5]
IPCW[6] method			1[6]
Subgroup analysis			2[7]
Unlabelled analysis			1[8]
AT analysis			3
Discontinuation of treatment analysed as treatment failure			3
Time to treatment discontinuation included as trial outcome			4
Did not report a statistical method to address nonadherence to treatment protocol	47 (48)		

[1] Nine trials carried out two methods of analysis, two trials carried out three methods and one trial carried out four methods.
[2] Number of trials reporting some form of nonadherence (98) is used as denominator for %.
[3] Adherence thresholds used were 60%, 2/3, 75%, 80%, and 90%.
[4] Adherence thresholds used were 2/3 and 5/6.
[5] Censoring times: time of starting disallowed intervention, when participants reported taking less than 2/3 of their medication in the past year, when received treatment out of trial, or censored following 6 month lag after receiving less than 80% of drug.
[6] Censored when received <80% of drug, weighted by the inverse probability of each participant's estimated adherence probability.
[7] Analysis split into two groups (according to whether participants had taken more or less than 50% of the prescribed medication) in one trial and into three groups (according to the proportion (>90%, 75-90% or <75%) of their time at risk that they were supplied with drug) in other trial.
[8] Included if received at least one dose of treatment.

6.3.8. Harms outcomes

Of the 69 trials that presented a harms analysis, 43 (62%) did not define the specific population set that was used in this analysis (see Table 16). Of the 26 trials that specifically defined a harms analysis population, the majority (18, 69%) specified that analysis was based on actual treatment received and included all patients who had received at least one administration of study agent, but only one study specifically stated that participants who received the alternative treatment rather than that allocated to them would be included in the alternative treatment group for this analysis. The remaining 43 trials that did not define a specific harms analysis population most commonly appeared to analyse the harms outcomes according to the specified benefit analysis population (31 ITT, two PP).

Table 16 Harms analysis populations in 69 trials that presented harms analyses

		Number (%[1])	
Harms analysis population specifically defined in methods		26 (38)	
Defined harms analysis population	Based on actual treatment received (i.e. including all patients who had received at least one administration of treatment)		18
	ITT		5
	All who started allocated treatment		2
	All who completed allocated treatment		1
Harms analysis population not specifically defined in methods		43 (62)	
Inferred harms analysis population	Stated "safety population" without further definition		1
	Apparently analysed as per efficacy outcomes		33
	ITT definition	31	
	PP definition	2	
	No details given of harms or benefit analysis population		9

[1] The number of trials with harms analyses presented (69) is used as the denominator for %.

6.3.9. Missing data

The distribution of the percentage of patients with missing primary outcome data (excluding 26 trials with a time to event primary outcome, where censoring was applied when follow up was terminated prematurely) can be seen in Table 17. In 12 trials, there did not appear to be any missing primary outcome data. In a further three trials there was mention of how the trialists dealt with missing data, but it was not clear what percentage of participants did not have primary outcome data available.

Table 17 Percentage of participants with missing primary outcome data (excluding 26 trials with time to event primary outcome)

	Number (=%)
Percentage of participants with missing primary outcome data	
None	12
0-5%	15
5-10%	16
10-20%	20
20%	8
Unclear but >0%	3

Of the 62 trials that reported missing primary outcome data, 29 (46.8%) trials did not attempt to account for the missing data and carried out complete case analyses only; simple imputation techniques were used in 26 trials (41.9%) of trials (either in the principal or sensitivity analyses) and more complicated techniques, such as maximum likelihood or multiple imputation, were implemented in 10 (16.1%) trials (see Table 18).

Of the methods implemented to account for missing data, the most common were last observation carried forward (LOCF) and assignment of the worst possible outcome to all participants with missing outcome data (for example, assuming treatment failure or non-response). Other methods included multiple imputation and maximum likelihood methods, pattern-mixture models, Markov chain Monte Carlo imputation, assuming no change from baseline and repeated measures analysis of covariance.

Nineteen trials described the profile of the missing outcome data either by comparing characteristics of those who did or did not have missing outcome data or by comparing rates of missing outcomes between treatment arms.

Table 18 Methods to adjust for missing outcome data

	Number (=% unless otherwise stated)	
Adjustment for missing primary outcome data		
Not applicable	38	
	No missing data	12
	Time to event outcome	26
Complete case analysis only[1]	29 (46.7)	
Principal analysis[1]	25 (40.3)[2,3]	
	LOCF	10
	Assumed worst possible outcome[4]	9
	Assumed no change from baseline	2
	Multiple imputation	3
	Maximum likelihood	2
	Markov chain Monte Carlo imputation	1
Sensitivity analysis[1]	9 (14.5)[3,5]	
	LOCF	2
	Assumed worst possible outcome[4]	3
	Assumed best possible outcome	1
	Pattern-mixture model	2
	Multiple imputation	3
	No details given	1
	Other[6]	1
Adjustment for missing secondary outcome data		
Principal analysis	19	
	LOCF	10
	Assumed worst possible outcome[2]	4
	No change from baseline	1
	Multiple imputation	1
	Maximum likelihood	1
	Markov chain Monte Carlo imputation	1
	Repeated measures analysis of covariance	1
Sensitivity analysis	2[7]	
	Assumed worst possible outcome	1
	Assumed best possible outcome	1
	Multiple imputation	1

[1] Number of trials with missing primary outcome data (where primary outcome is not time to event) (62) is used as denominator for %.
[2] Two trials used two methods to adjust for missing data in principal analyses of primary outcome.
[3] One trial adjusted for missing data in principal and sensitivity analyses of the primary outcome.
[4] For example, assumed to be a treatment failure/non-response.
[5] Four trials carried out two sensitivity analyses of the primary outcome.
[6] Estimated total number of events that would have occurred if follow up had been complete for all participants, and then assigned half of the extra events to each treatment group.
[7] Two sensitivity analyses of secondary outcomes were carried out in the same trial.

6.3.11. Protocol permitted changes to treatment

Protocol permitted changes to treatment were not considered as treatment deviations for this review. However, given that it may be necessary to factor out such treatment changes when considering certain causal research questions, it was of interest to note that the treatment protocols of 15 trials in this review permitted modifications to the dose of trial treatments or temporary or permanent cessation of treatment (in the event of certain adverse effects), which lead to necessary deviations from the optimal treatment dose. Six other trials introduced changes to the treatment protocols during the course of the trial. These included the introduction of additional concomitant treatments to both arms of the trial, dropping of a treatment arm (in a three arm trial), dropping of a lead-in treatment phase, replacing one treatment with another and modifications of the specified treatment dose as a result of external evidence. Two cancer trials permitted patients in the control arm to switch to the intervention treatment on disease progression or relapse, and one trial permitted participants to begin taking open-label treatment.

Only three of these 15 trial reports investigated the effect that the change, or permitted modifications, to treatment protocol had on the observed efficacy of the treatment. For example, in one trial (165), the protocol was amended part way through the study permitting a lower initial treatment dose (at the discretion of the investigator) because of safety concerns. A sensitivity analysis was carried out excluding those participants who received an initial dose lower than was originally permitted in treatment protocol.

6.4.Discussion

6.4.1. Reporting of adherence information

In this review of published trials, although the vast majority of RCTs were subject to at least one form of nonadherence to treatment protocol (most commonly incomplete treatment or non-receipt of allocated treatment), often even the most basic adherence information on initiation, completion and premature discontinuation of treatment was not presented in some trials. Perhaps most remarkable was the fact that forty-two percent of the publications did not explicitly state how many patients actually initiated their randomised treatment. The template for the CONSORT flow diagram suggests that, in the treatment allocation box, trialists should report the number of participants who "received" allocated intervention, and in the follow up box, they should then report the number who "discontinued" intervention. However, except in the case of trials with treatment given at a single time point (only 12% of the trials in this review), it would be less misleading and more accurate to ask trialists to report the number of participants who "initiated" rather than "received" intervention, as "initiated" is unambiguous but "received" may be interpreted either as initiation or as receipt of the entire study treatment. Indeed, in the 2010 CONSORT elaboration document, the table which details the information required in the flow diagram (table 3) states that the treatment allocation box should include the number of participants who "completed" treatment as allocated, rather than "initiated" or at least "received".

The absence of any explicit reporting of the number of participants who initiated intervention may lead readers to assume that all randomised participants at least started their randomised intervention. Indeed Vrijens (3) states that, in the context of clinical trials of prescribed medication, given that the first dose of a randomised medication is usually administered on site following informed consent, "it is often assumed that

initiation is implicit for all included patients". However more than one third (26, 38%) of the 69 drug trials in our review included participants who did not initiate their randomised intervention.

Other inadequacies in reporting (related to departures from treatment protocol) were evident in this review. For example, none of the 13 trial reports that specified a cut-off to define adequate adherence included an explanation for the choice of threshold. As discussed by Vrijens (20), the use of "ad hoc" threshold values, such as the "often-used but never pharmacometrically justified" adherence criterion of taking at least 80% of prescribed doses to define sufficient and insufficient exposure to drugs needed to achieve satisfactory therapeutic results, is unacceptable as such a threshold will depend on a wide range of underlying drug-, disease- and formulation-specific pharmacodynamics.

In the case of pragmatic trials with nonpharmacological interventions, especially those randomised at the cluster level with intervention administered by health care providers rather than by the patient, it was often difficult to determine which participants received the full intervention as per protocol. For example, some trial reports implied that treatment providers were either unwilling or unable to follow the treatment protocols fully, but it was not possible to ascertain the extent to which this affected the treatment of individual patients. Also, with some complex interventions, it was sometimes difficult to define when patients were "on" or "off" treatment, especially when use of, and the level of use, of the intervention was optional. Indeed, of the eight trials that failed to provide any information on treatment persistence or adherence, all had pragmatic designs and all but three were cluster trials with nonpharmacological interventions.

6.4.2. Incomplete reporting of exclusions from analysis

The reporting of methodological features of RCTs has historically been shown to be less than optimal, with numerous reviews providing evidence that authors of trial publications often fail to include details on the important features of design and analysis (154, 158, 159). In particular there is increasing evidence that trialists provide inadequate information on the exclusion of participants from analysis (153, 155, 157, 166, 167).

Schulz et al (166) investigated the association between methodological quality and evidence of biased treatment estimates in 250 RCTs (published between 1955 and 1992) from 33 meta-analyses from the Cochrane Database, and found that inadequate methodological approaches to trial design are associated with bias, particularly with respect to poor allocation concealment. They unexpectedly found, however, that those trials that reported excluding participants after randomisation did *not* yield exaggerated treatment effect estimates compared with trials that explicitly reported, or "gave the impression", of no exclusions (the vast majority of which were *not* explicit). They noted, however, that of those trials report that gave the *impression* that no exclusions had occurred, few *explicitly* stated that no exclusions had occurred. Thus they conjectured that this phenomenon may have be resulted simply because those trials demonstrating higher methodological quality (for example, using adequately concealed allocation) were similarly likely to demonstrate better reporting quality (by clearly reporting on the exclusion of randomised participants), whereas those involving poorer methodology were less likely to openly report excluding participants.

A similar review by Schulz (167) of obstetric and gynaecology trials published in 1990/1991 provided more evidence of this "exclusion paradox". Schulz categorised trials according to whether the trial publication explicitly stated that there were some or no exclusions or where there were "no apparent exclusions" (when they simply reported

analysing the same number of participants as were randomised, thus implying that no exclusions had occurred). They found that those trials with reported exclusions and those which explicitly reported no exclusions were more likely to have implemented adequate methods of allocation concealment than those with no apparent exclusions, and thus "paradoxically, trials that reported exclusions seemed generally of a higher methodological standard than those that had no apparent exclusions". Thus Schulz suggests that when there is a no mention of exclusions (or a lack of exclusions) in a trial publication, exclusions may nevertheless have taken place, leading to an "unsettling irony… some of the *more* biased trials might be mistakenly interpreted as unbiased, while many of the *less* biased trials may be interpreted as biased" (168).

Although the trials in both of Schulz's review are not current, and it is hoped that reporting of trials has improved in recent years following acceptance of the CONSORT statement (160, 169), the finding in our review that only 58% of studies explicitly report how many patients initiated randomised treatment suggests reporting could be improved further. Furthermore, although the number randomised and analysed (at least for the primary outcome) was reported in all of the trials in our review, given the findings by Schulz, without explicit mention of exclusions, this does not guarantee that the number analysed was equal to the total number initially randomised.

For example, the wording in the text of a few trial reports in our review regarding the number of randomised participants was slightly ambiguous and may be suggestive of the exclusion of certain participants from the trial report. This seemed especially likely when the description of the number randomised was conjoined with another descriptive characteristic of the participants, for example "883 HIV-infected, treatment-naive patients were randomised and analysed for efficacy" (170) or "A total of 579 people were randomised and completed the baseline questionnaires" (171).

6.4.3. Assessment of participant adherence in drug trials

Despite the fact that collection of participants' adherence data is now considered to be an important aspect of clinical trials and use of a combination of measures is recommended to improve accuracy of adherence data (14), less than half of the trials involving participant-administered treatment mentioned **checking** adherence, and less than one third of these assessed reliability of the adherence data. The most common methods to ascertain adherence involved counting pills (either directly or by weighing medication bottles) (11 trials) or questioning the participant (in person or through questionnaires or diaries) (11 trials), even though these methods are not considered to be particularly reliable (14). It is of interest to note that more accurate methods such as medication events monitoring systems (MEMS) or measurement of drug metabolite or marker in bodily fluids are not typically used (only one trial used blood tests and laboratory assays to assess adherence), even in the large scale and high quality trials which are published in the four journals included in this review.

Trial-specific definitions of compliance or adherence were common, especially in the case of drug trials with long term intervention periods, where compliance was usually defined as the proportion of prescribed drugs that were actually taken by the participant. One quarter of drug trials (17/69) in this review **defined** compliance, based most commonly on the percentage of medication received (14) but also on discontinuation of medication (3). Ten trials defined good or poor compliance based on cut-offs on the percentage of medication received, but none of these thresholds were justified.

6.4.4. Trials of nonpharmacological interventions

Trials of nonpharmacological interventions reported on participant compliance far less commonly than in drug trials, and the complex nature of the interventions and pragmatic trial designs meant that less straightforward and potentially ambiguous definitions of compliance were more commonly reported than in drug trials.

Checking compliance with treatment protocol on the part of treatment provider (rather than just the participant) is especially important in trials of nonpharmacological interventions. This is particularly true for those which are randomised at the level of the treatment provider (cluster trials), and this was evident in this review, as proportionally more of those randomised at a cluster level specifically mentioned that adherence of treatment providers was monitored compared to those randomised at the individual level (55.6% versus 26.9% respectively).

The compliance data reported regarding adherence by treatment providers to the trial protocol on administration of nonpharmacological interventions were not, however, always very clear or transparent, and it was not easy to determine the degree of impact that the reported nonadherence would have had on trial participants. Most of these twelve trials simply discussed the fact that aspects of delivered treatment had been recorded (for example, using case report forms or program content checklists, videotaping treatment sessions, or recording information on a study web server) and thus it was determined that treatment providers had deviated from randomised treatment protocol in some manner. Two cluster trials discussed the rate of acceptance, participation or implementation of treatment guidelines by site staff.

6.4.5. Statistical methods to deal with departures from treatment protocol

Although more than half (51/98, 52%) of the trials that were subject to nonadherence to treatment protocol implemented a statistical analysis method to deal with such nonadherence when analysing benefit outcomes, these were most commonly based on variations of PP analysis (46) and very few recognised, or sought to address, the potential for bias introduced when excluding or censoring patients at the point of deviation from treatment protocol. Only one trial (164) addressed the potential bias caused by censoring of patients at the point of deviation from treatment protocol, using the IPCW method (150), while another trial (172) provided justification for the decision to exclude participants who had not received the allocated intervention, citing Fergusson (173) and stating that the "omission [of such patients] would be equally distributed between groups, would be unrelated to treatment assignment, and would not bias outcome ascertainment". Trial publications commonly presented insufficient detail or explanation on the analysis sets used for benefit and harms outcomes. Fifty-seven percent of trials did not explain the composition of analysis sets for their benefit outcomes, and 62% of trials that presented harms analyses did not specifically define harms analysis populations. This common lack of discussion regarding whether, and if so why, patients were excluded from analysis and the potential biases which may result from such exclusions, indicates a low priority given to the issue of nonadherence to treatment protocol in the analysis of published trials.

6.4.6. "ITT" analysis to handle nonadherence to treatment protocol

Eighteen trials in this review carried out what were referred to as "ITT" analyses but the ITT analysis set excluded participants if they did not adhere to the treatment protocol in

some manner. This mislabelling of ITT analyses has been noted in previous reviews of ITT analyses in recent trial publications (174-177).

It is potentially interesting to compare the results of these four surveys of trial publications with those from our review, in order to investigate whether the reporting of trials and the use of ITT analysis has improved over time. Indeed, such comparison (summarised in Table 19) suggests that the use of the phrase "ITT" to describe analyses of RCTs is becoming more common (48% in 1999 (176), 71% in 2007 (174) and 83% in 2008), at least in these four leading medical journals, but that the correct definition of ITT analysis is being followed less commonly: 13% of "ITT" trials published in 1999 excluded participants who did not receive the randomised treatment (176) compared to 22% of such trials in our review ($p=0.086$, Pearson's χ_1^2 statistic = 2.95).

Table 19 Reviews of ITT analyses in RCTs

	Hollis (1999)		Kruse (2002)[1]		Gravel (2007)		Herman (2009)		Dodd	
Journals	BMJ, JAMA, Lancet, NEJM		Multiple English		10 (including BMJ, JAMA, Lancet, NEJM)		8 (surgical)		BMJ, JAMA, Lancet, NEJM	
Publication year	1997		1999		2002		2005-Aug 2008		2008	
Number of trial publications	249		100 (random selection)		403		274		100 (random selection)	
[number in BMJ, JAMA, Lancet, NEJM]	[249]				[283]				[100]	
ITT reported: number (% of trials)	119 (48%)		100 (by definition of search)		249 (62%) (ITT for primary analysis) (including 23 (9%) "modified ITT")		96 (35%)		83 (83%) (plus additional 6 "modified ITT")	
[number (% of trials) reporting ITT in BMJ, JAMA, Lancet, NEJM]	[119 (48%)]				[201 (71%)]				[83 (83%)]	
Assessment of appropriateness of ITT analysis: number (% of trials reporting ITT)										
	Analysis as randomised (solely with respect to deviation from treatment protocol)	104 (87%)	Analysis as randomised	42	Analysis as randomised (without consideration of eligibility)	97 (39%)	Analysis as randomised	45 (47%)	Analysis as randomised (solely with respect to deviation from treatment protocol)	65 (78%)
							No missing data	17 (18%)		
							Imputed missing data	28 (29%)		
	Explicitly stated	34 (29%)							Explicitly stated	21 (25%)
	Appeared to	65 (55%)							Appeared to	40 (48%)
	Stated no deviations	5 (4%)							Stated no deviations	4 (5%)
	Excluded participants from analysis due to deviation from treatment protocol [specifically non-initiation of treatment]	15 (13%) [12 (10%)]	Excluded participants from analysis	58	Excluded participants from analysis	152 (61%)	Excluded participants from analysis	51 (53%)	Excluded participants from analysis due to deviation from treatment protocol[3] [specifically non-initiation of trt[3]]	18 (22%) [15 (18%)]
			Missing follow up data	17	Missing follow up data	95 (39%)	Missing follow up data	44 (46%)		
			Treatment non-receipt	15	Treatment non-receipt	10 (4%)	Treatment non-receipt	6 (6%)		
			Ineligible	12	"Modified ITT"[2]	23 (9%)				
			Drop out	4	Disease exacerbation/ adverse effects	4 (2%)				
			Protocol violations	4						
			Unclear	13	Unclear	17 (7%)	Unclear	1 (1%)		
			Other	9	Other	3 (1%)				

	Hollis (1999)	Kruse (2002)[1]	Gravel (2007)	Herman (2009)	Dodd
Missing outcome data: number (% of ITT trials)					
Stated none missing	30 (25%)	42	97 (39%)	17 (18%)	8 (10%)
<10% missing	60 (50%)	47	100 (40%)		46 (55%)
>=10% missing	29 (24%)	11	52 (21%)		26 (31%)
Unclear (but >0%)					3 (4%)
Missing data techniques: number of trials (% of ITT trials with missing outcome data)					
Complete case analysis only	44 (49%)		89 (59%)	44 (46%)	21 (28%)
All available information[4]	29 (33%)				25 (33%)
Imputation technique	16 (18%)		36 (24%)	28 (29%)	29 (39%)
[of which conducted sensitivity analysis[5]]	[1 (exact methods unstated)]			[6]	[7]
LOCF	7		12	18	11
Explicit assignment of good/poor outcome	3		2	3	3
Implicit assumption of good/poor outcome (included in denominator)	4		18		8
Group average	1			4	
Complex[6]			1	2	10
Unspecified				1	
Unclear handling of missing data			27 (18%)	6 (6%) (only mentioned exclusion of participants who did not receive allocated treatment)	
Flow diagram included: number (% of trials)		41			96 (96%)
ITT method mentioned in one place only: number (% of ITT trials)					
Methods section			132 (33%)		81 (98%)
Results section			15 (3%)		2 (2%)
Abstract			18 (4%)		
Discussion			1 (0.4%)		
Effect of CONSORT		Journals adhering to CONSORT more likely to display flow diagram	Journals adhering to CONSORT more likely to perform ITT analysis		

[1] % of trials reporting ITT is equal to number of trials in all cells, as 100 trials reporting an ITT analysis were selected in this review

[2] Made exclusions due to ineligibility or non-receipt of treatment

[3] Note that this figure includes two trials that explicitly stated that participants who did not initiate randomised treatment would be excluded from the ITT analysis but which did not report any such deviations.

[4] For example, censoring participants at end of follow up.

[5] Note that some of these trials compared a single imputation technique against complete case analysis.

[6] Includes regression and multiple imputation, maximum likelihood, Markov chain Monte Carlo imputation, repeated measures analysis of covariance

6.4.7. Missing data

Missing outcome data may be linked to treatment nonadherence, and thus this review assessed the prevalence and statistical handling of missing primary outcome data. The majority (84%) of trials (excluding those with a time to event primary outcome) had some missing primary outcome data, but nearly half of these (46.8%) ignored these missing data in analysis and carried out complete case analyses only.

Wood et al (163) also reviewed how trialists report and adjust for missing outcome data in RCTs published in July-December 2001 in the same four medical journals as used in our review. Wood highlighted the need for multiple sensitivity analyses in the presence of missing outcome data, in order to assess robustness of the assumptions made in the principal analysis, and that sensitivity analyses should make different assumptions to the principal analysis. For example, extreme case analyses (whereby patients in the respective treatment groups are assumed to have the most and least favourable outcomes, and vice versa) may be performed to demonstrate extreme case scenarios; if conclusions from such analyses are not substantially different to complete case analyses, it may be concluded that the conclusions are not heavily affected by loss to follow up.

Of the 62 trials in our review that reported missing primary outcome data excluding those that censored participants at final follow up, 33 (53%) implemented methods to handle the missing data, including 9 (14.5%) that tested the assumptions of the principal analysis with sensitivity analyses or secondary analyses of the primary outcome. This compares to 32% and 21% of such trials, respectively, reported in the review by Wood et al, suggesting that although trialists seem to be increasingly likely to implement missing data methods, the practice of testing the robustness of the assumptions made in these analyses through sensitivity analyses is not becoming more common.

It is important to report the reason that outcome data are missing, in order to inform missing data assumptions (163); for example, only if data are missing completely at random will complete case analysis will be appropriate. Only one trial in this review explicitly stated the assumptions behind the missing data analysis, namely that the multiple imputation of missing data was carried out assuming the data were missing at random. It is also important to describe the missing data in each treatment group separately, as recommended in the CONSORT diagram. All studies did this, but reasons for discontinuation of treatment or missing outcome data were not given in 22 of the 96 trials that provided a CONSORT diagram; 13 of these stated only that participants were lost to follow up. Reporting reasons for missing data allows the reader to assess whether the missingness is likely to be random or not, as well as informing analysis assumptions. In particular, rather than simply reporting "withdrawal rates", it is important to differentiate withdrawal from treatment (failure to complete treatment protocol) from withdrawal from study (failure to complete follow up).

6.4.8. Harms analyses

A variety of analysis sets were used for the analysis of harms data in this cohort of trials, and this may be a consequence of the lack of consensus in the research literature on the most appropriate harms analysis population to be used in the event of departures from treatment protocol. The 2001 version of the CONSORT statement (160) supports the view of Lewis (75) that ITT is not appropriate for analysis of harms outcomes, stating that that "a pure ITT approach to the analysis of safety simply adds to the risk of failing to identify potential safety problems, and is therefore never advocated". Similarly, the ICH GCP guidelines (61) suggest that analysis of harms data should be according to treatment received (in other words, all participants who received at least one dose of a treatment should be included in that treatment group for harms analyses), and this was

the most common analysis set specifically defined for harms outcomes in this review. Furthermore, although only one quarter (18/69) of the trials that presented a harms analysis specified this analysis population for their harms outcomes, in fact over half (37/69) carried out analysis according to treatment received (including, for example, trials which specified an ITT analysis where no treatment switches occurred).

However the 2004 CONSORT extension for reporting harms (49) conversely states that ITT is usually preferred for both benefit and harms outcomes because it reflects the original trial design. The CONSORT 2010 statement has removed any reference to the appropriateness or otherwise of ITT analysis populations for harms outcomes.

The observed variation in the analysis populations chosen for harms outcomes in the presence of treatment protocol nonadherence (in this review of trials) suggests that explicit guidance is needed on how harms data from patients who deviate from treatment protocol should be analysed, for example what to do if a patient receives both or none of the trial treatments, or if they receive a treatment to which they were not randomised.

6.4.9. Trial report discussions relating to treatment nonadherence

Twelve reports specifically discussed the potential influence of treatment nonadherence on the observed treatment effect (but only seven of these trials actually carried out analyses that in some way handled departure from treatment protocol). Four trials discussed the fact that the observed results may be an underestimate of the true treatment effect because of non-acceptance of the intervention (two trials), crossover from placebo to active therapy (one trial) or discontinuation of active drug and commencement of rescue medication (one trial). Another trial concluded that decreasing compliance had not significantly influenced the results of the trial, as the

results of the sensitivity analysis assessing whether diminishing compliance had caused the treatment effect to wane over time did not differ greatly from those of the main analysis. Three trials reported contamination between treatment groups, and two of these commented that contamination from the intervention to control group may well have contributed to the discouraging observed treatment effects. Two of these trials also noted an unexpected improvement in the control arm, and they discussed the fact that those who had opted to take part in these behavioural intervention studies may well have been more motivated than the general target population, and that simply taking part in the trial may have improved their performance (because of increased awareness of the issues or increased time spent with research nurses). In addition, two cancer trials discussed the related problem of how to interpret ITT analysis in the face of substantial switching from placebo to active treatment, or from standard to new intervention, when patients' disease progressed, and they commented on the fact that this crossover was likely to have caused a diminished treatment effect on overall survival. Such treatment changes, rather than being caused by nonadherence to treatment protocol, were actually permitted by the protocol.

6.4.10. Limitations

This review is limited by a sample size of 100 trial publications taken from only four high impact general medical journals, and thus the generalisation of our results to other less widely read or more subject-specific journals may be limited. However it is therefore arguable that, for this reason, the findings evident from this review should be regarded as an estimate of the upper limit of the quality of reporting and analysis of nonadherence to treatment protocol.

As regards the critique of analysis methods, because trial authors were not contacted directly, it cannot be verified that the details reported in the trial publications were

complete and accurate accounts of how analysis proceeded. Additionally, it would have been of interest to ascertain, via discussion with authors, whether any ambiguity or absence of results on nonadherence analyses may simply have been due to poor (or restricted) reporting rather than poor practice. Furthermore, without access to the protocols or statistical analysis plans for the trials in this review, it was not possible to ascertain whether the analyses carried out to examine the effect of treatment protocol deviations were decided on prior to data collection or were post-hoc decisions. Two trials in this review admitted that the statistical method to deal with departures from treatment protocol were post hoc analyses.

The main limitation of this review was the use of a single reviewer for data extraction. However this reviewer was able to consult the opinion of a second reviewer whenever there was doubt as to appropriate classifications. This was necessary, however, in less than 5% of the articles reviewed, as systematic data extraction from all trial reports was undertaken using clear definitions for classification of results.

6.5.Recommendations

This chapter has highlighted numerous limitations relating to the measurement, reporting and adjustment for treatment adherence in published trials. In particular, trialists rarely employ reliable methods to measure adherence and fail to provide justification for thresholds used to determine good or poor compliance. Reporting on adherence to treatment is poor, particularly relating to treatment initiation and treatment adherence with long term treatment protocols. Causal methodology is rarely employed to handle treatment deviations; instead trialists typically employ simple naïve analysis methods. Investigations regarding the impact of missing outcome data are often basic, generally without regard for the impact of assumptions made in such analyses.

Thus the work carried out for this chapter has led to a number of recommendations. Firstly, recommendations were made to the CONSORT steering group to improve the wording regarding the reporting of treatment receipt in the CONSORT flow diagram. In particular, the suggestion was made that it would be more meaningful to ask trialists to report the number of participants who "initiated" rather than "received" intervention, as "received" may be interpreted either as initiation or as receipt of the entire study treatment. Furthermore, the error included in the 2010 CONSORT elaboration document (in table 3, which states that the treatment allocation box should include the number of participants who "completed" treatment as allocated, rather than "initiated" or at least "received") was brought to the attention of the steering committee, who subsequently acknowledged this error and stated that it would be corrected in the next revision of this document.

Secondly, specific recommendations were developed (see Table 8) for trialists regarding explicit reporting (in trial reports) of information relating to adherence to treatment protocol, providing a clear template for authors to ensure transparency and unambiguous interpretation of trial results.

Finally, in Table 20, more general recommendations are provided in order to improve trial conduct, reporting and analysis on issues relating to adherence to treatment protocol.

Table 20 Recommendations for trial conduct, reporting and analysis

1. Trialists should document in the trial protocol how they plan to measure and report nonadherence to treatment protocol, with explanation in protocol and publication of why the chosen adherence measure is clinically the most important and relevant.
2. The decision on whether, and how, to examine the effect of nonadherence to treatment protocol should be made prior to data collection, and details of the planned analyses should be documented in the protocol (and statistical analysis plan) with explanation for proposed methods. Trial publications should include a discussion of potential bias introduced by any such analyses, and efforts should be made in the statistical analysis to reduce any bias introduced by excluding participants from analysis.
3. Composition of analysis sets used for benefit and harms outcomes should be defined explicitly, rather than merely labelled as "ITT" or "PP".
4. The total numbers of participants who were randomised, excluded (with reasons) and analysed for each outcome should be reported as recommended in the CONSORT statement (41).
5. Potentially ambiguous phrases such as "protocol deviations", "completed study/treatment protocol", "withdrawal", "intention to treat", "modified intention to treat", and "per protocol" should be carefully defined in terms of the treatment protocol if used in trial reports.
6. Trial reports should clearly distinguish between withdrawal from treatment and withdrawal from study: i.e. clarify whether participants withdrew from treatment but agreed to continuing to provide follow up data or whether participants also withdrew from further follow up, and if the latter, whether or not they consented for the data collected up to the point of withdrawal to be included in the analysis.
7. Reasons for treatment changes and withdrawal from study should be recorded in order to inform causal analyses (adjusting for treatment deviations) and sensitivity analyses (assessing the impact of missing outcome data on trial conclusions).
8. Reporting of treatment receipt and completeness should reflect the duration of trial intervention(s):
    a. Trials with intervention given at a single time point should report the number of participants who received allocated intervention in each randomised group.
    b. Trials with short term intervention should report the number of participants initiating and completing allocated treatment as specified in the protocol in each randomised group.
    c. Trials with long term intervention should report the number of participants initiating and persisting with allocated treatment as specified in the protocol, along with a measure of participant and/or treatment provider adherence over the treatment period, in each randomised group.
9. Participant adherence should usually be assessed in at least a random subsample of individuals in trials involving participant-administered treatment using a reliable method or, if necessary, using more than one method to gauge reliability of the assessment.
10. Biological or medical explanation should be provided for any thresholds used to define adequate adherence, and these should be specified in the trial protocol.

6.6.Conclusions

The results of the review suggest that nonadherence with randomised treatment occurs in the vast majority of trials but is poorly considered and poorly reported in trial publications. Nonadherence information presented in trial reports can be ambiguous or scant, particularly relating to treatment initiation and completeness. Despite availability of appropriate adjustment methods (as summarised in the previous chapter), this review has demonstrated that trialists rarely consider approaches beyond simple biased methods such as PP or AT analysis to adjust for nonadherence, thus typically excluding or censoring participants who deviate from treatment protocol without discussion of the potential bias introduced in such analyses.

This suggests a need to explore potential barriers that prevent routine use of appropriate adjustment methods, with particular emphasis on the analysis of survival outcomes, given that these are most commonly associated with long term treatments and dynamic treatment regimens which are typically subject to treatment deviations. As such, subsequent chapters of this thesis will demonstrate the challenges and complications arising in the application of two contrasting causal techniques for survival analysis in a selection of the expository trials described in Chapter 4. Prior to these practical demonstrations of analysis methods, the next chapter aims to clarify the causal estimation process by considering the practical steps that are necessary to carry out such analyses, presenting two pictograms intended as research aids to simplify the process for non-specialist analysts.

7. Causal estimation process and considerations

7.1.Introduction

The results of the review of reporting and analysis of nonadherence in published trials, summarised in Chapter 5, demonstrate clearly that nonadherence to randomised treatment is a common problem affecting virtually all trials, but that appropriate methods of statistical analysis to adjust for this nonadherence are rarely applied. Instead naïve and statistically biased methods, such as PP or AT analyses, are frequently used.

There is therefore a pressing need to explore and address the barriers that prevent trial statisticians from applying appropriate methods to estimate efficacy of treatment when faced with treatment deviations.

Firstly, trialists and trial statisticians may simply be unaware, not only of the bias that may be introduced when methods that destroy the balance of randomisation are implemented, but also of alternative methods that more effectively prevent selection bias when accounting for departure from randomised treatment. Even when analysts are familiar with such methods, they are likely to face financial and time constraints at the point of trial analysis, as funding bodies may themselves be relatively unaware of the value of causal analysis methodology. Statistical analysis plans may already be substantial, covering analysis of numerous secondary outcomes as well as potentially addressing issues relating to missing data. Trial analysts and funding bodies may therefore question the value of spending time, resources and money on analyses beyond those that are usually carried out, and it may be difficult to persuade funders to cover additional costs that are required to ensure that compliance analyses can be carried out with sufficient rigour and attention. This suggests that there is a need to promote awareness of the usefulness of considering compliance methodology to complement the accepted ITT and PP analyses, both within trial and clinical communities as well as among funding and regulatory bodies.

Furthermore, the culture of acceptance of ITT as the standard method of analysis, along with a reluctance on the part of trialists to consider alternative methods of analysis out of fear of potentially introducing selection bias has meant that trials have typically been carried out with a strong focus on pragmatic aims, thus deterring statisticians from investigating or promoting alternatives methods of causal analysis. This may be particularly true for trials carried out in general medical practice, where the effectiveness of treatment is generally considered of greater importance than biological efficacy. However, as has been discussed in previous chapters, causal methodology is useful for demonstrating effectiveness of treatment, for example in alternative clinical

settings with different patient populations. Thus trials may benefit from additional analyses investigating the impact of adherence issues regardless of the ultimate trial aim. However, when such methods are introduced, their sheer complexity, along with a lack of experience of how to apply such methods in practice to what may be a complex compliance trial scenario, may hinder their use further. As such, in accordance with the remit of this PhD to facilitate use of appropriate but potentially complex and unfamiliar causal estimation techniques by non-specialist trial statisticians, this chapter presents a discussion of the issues that must be considered as part of the causal estimation process, not simply in terms of modelling and interpretation of results, but also with consideration for the necessary planning (of data collection and statistical analysis methods) to ensure transparency and applicability of chosen statistical methods.

A pair of pictograms are presented, intended as analysis aids, the first presenting a clear depiction of the necessary considerations that must be addressed prior to undertaking causal analysis of trial data and the second summarising the statistical issues and methodologies according to the particular trial traits. In subsequent chapters, these pictograms are put to use with demonstration of how theoretical methodologies may be applied, as an example for analysts who may be interested in carrying out similar adjustments for nonadherence.

7.2.Causal analysis considerations

Causal analysis estimation necessitates consideration, not only of the choice of statistical methodology, but also of the clinical requirements and data limitations. As such, before causal estimation is possible, it is necessary to consider the whole trial scenario relating to the causal estimands of interest.

The process of causal estimation necessarily involves a complicated, and potentially compromising, marriage between a number of factors. First, the causal research question of interest must be translated into appropriate causal contrasts (of treatment received) forming the basis for the causal modelling estimation. Secondly, it is necessary to decide on how to collect and convert available data on compliance (or treatment received) into appropriate formats to inform the causal contrasts of interest. Then it will be necessary to select the most appropriate method of statistical analysis, taking into consideration relevant compliance issues, necessary covariate adjustments and outcome type. Finally, the results from this analysis must be interpreted in light of the limitations presented by the data and methodology, with potential use of sensitivity analyses to investigate the robustness of conclusions to underlying assumptions.

Each of these key elements of the causal estimation process will be explored in this chapter.

7.3.Define research questions and causal contrasts of interest

First and foremost, (as introduced in Chapter 4) it is necessary to decide on a clear **research question of interest**, which will in turn allow identification of the causal estimand of interest. In order to estimate any effect beyond that of treatment assignment, it will be necessary to consider how the treatment patterns in the trial setting relate to this causal estimand. This will determine which of these deviations would be usefully factored out of analysis because they contravene the treatment path of interest, and which are inherently part of the treatment path that is of interest, such that their occurrence is not problematic for that particular causal estimation. As such, it is necessary to consider how deviations from treatment assignment are likely to

manifest themselves, both within the trial setting and the particular (potentially hypothetical) setting to which the trial results are to be applied.

The focus of this research question, and its subsequent implications on analysis, will vary according to the perspective of interested parties. For example, as discussed in Chapter 3, a motivated patient interested in the likely benefit of treatment if they comply with doctors' prescription of said treatment may be interested in the causal effect of treatment taken as prescribed while allowing for the potential for necessary changes to their prescription if they experience side effects or the treatment fails to be efficacious. Alternatively a policy maker may be interested in applying trial results to a general community setting where treatment changes such as those which occurred in the trial are likely but perhaps to different degrees to those observed in the trial. Differently again, funding decisions by regulatory bodies such as NICE often require full cost-effectiveness analysis which typically relate to the effect of treatment taken for life, particularly for drugs which impact on survival; as such, they require estimation of causal effects which factor out all changes from originally randomised treatment that do not reflect typical real-life availability of treatments.

As such, the treatment deviations to be factored out of analysis for a particular causal analysis will not necessarily contravene the treatment protocol. It may, for example, be necessary for recruitment or ethical reasons to permit trial patients to switch treatments on disease progression, but nevertheless the causal research question of interest may relate to the effect of treatment on survival in the absence of all switches.

On the other hand, deviations may also occur that are unexpected or undesired, for example when patients (or clinicians) have pre-existing ideas about the efficacy of the randomised treatments, and thus may request (or prescribe or administer) different treatments to those which have been allocated by randomisation. Patients may not

honour the agreement made at randomisation when they consented to receiving and complying with either of the potential randomised treatments, and may request to have their treatment changed or may simply cease taking their randomised treatments, possibly withdrawing from follow up altogether. Or clinicians external to the trial may overrule and allocate what they perceive to be the best treatment for their patient.

Although such treatment changes may also be expected to occur as part of clinical practice, and thus some may argue that the ITT analysis reflects a true assessment of the policy of assigning treatment, this will only hold if the patterns of deviation from treatment protocol that occur in practice are the same as those occurring in the trial. This may be unlikely: with the passage of time, new treatments will emerge as the "most promising", and thus alternative treatments switched to in the trial may well differ from those that are experimented with in years following the trial. Also the conclusions and dissemination of results from the trial may in fact alter the pattern of compliance with the drug: if the results of the trial are seen as positive by the clinical community or the media, patients' likelihood and willingness to persist with treatment may be increased and thus fewer deviations from treatment protocol may occur in practice than in the trial. Thus a simple ITT analysis may not answer *any* question of interest reflecting everyday clinical practice.

If instead one is able to estimate a true causal effect of treatment from a trial, this can then be used to answer a patient's enquiry as to what treatment effect they could expect if they were able to persevere with treatment. This causal estimate may also be used to estimate the treatment's effectiveness in clinical practice assuming various levels of acceptability by patients and clinical staff. Thus regardless of whether deviation from treatment protocol is expected in the trial setting or in clinical practice, there is a strong

argument for estimating a true causal effect of treatment, to accompany the necessary primary ITT analysis (91).

Thus, the definition of the exact efficacy question of interest should depend on the types of deviations from treatment protocol that are accepted as integral to the course of treatment or the disease, and which deviations would be usefully factored out as "noise". Once the particular clinical question to be answered has been defined in terms of treatment deviations to be factored out of analysis, it is then necessary to consider how the corresponding treatment effect of interest can be estimated, taking into account availability of the appropriate data in order to inform the relevant statistical methodology.

The causal analysis method will in some way transform data from the experimental and control patients' outcome data back to what would have been observed if they had followed exactly that course of treatment corresponding to the research question. Therefore, prior to undertaking any such analysis, it is necessary to firstly identify and specify these (potentially counterfactually) contrasting states of treatment.

Thus, as for PP and AT analyses, where it was necessary to define when patients were considered to be "on" and "off" treatment, these definitions must be considered when defining the causal contrasts between control and experimental conditions (relevant to the causal research question). These definitions of treatment receipt will vary according to whether the relevant measure of treatment receipt was constant or time-varying, and whether it was recorded using a binary (for example, all-or-nothing compliance or repeated binary measures over time) or quantitative measurement (such as dose, recorded as a single summary measure (for example, cumulative dose) or changing dose over time), or as a time to event variable (such as time to treatment withdrawal). It may also be necessary to consider wider forms of nonadherence, for example switching to

the alternative trial treatment; switching to, or supplementing randomised treatment with, non-trial treatments; premature withdrawal from all treatment or intermittent nonadherence.

7.3.1. Data collection

Thus, having identified the relevant forms of nonadherence which will impact on the estimation of the treatment outcome of interest, it is necessary to ensure that relevant data on treatment receipt have been recorded in order to appropriately inform this estimation. The precise collection of data should be determined by the trial interventions and type of nonadherence that need to be reported and adjusted for.

Gross (178) discusses the difficulties in accurate measurement of adherence information (with respect to preventive intervention trials, but with relevance to other forms of intervention) which stem from the ambiguities and subtleties surrounding compliance issues. In particular, prior to collection of any compliance data is the need to explicitly define the meaning of compliance relevant to the study. "Nonadherence" may be manifested in a number of ways relating to both treatment and follow up. For example, with a long term intervention, participants may fail initiate any part of their allocated intervention or attend any follow up; engage in some but not all of the intervention; start but then prematurely terminate participation or attend clinical follow up but not follow allocated intervention.

It may be possible to capture the most important features of treatment receipt using a binary factor; for example, as a simple all-or-nothing treatment receipt indicator in the case of a one-off treatment, or when treatment switches occur immediately following randomisation only. When partial compliance is of interest, treatment receipt may be recorded as a continuous measurement which potential changes over time; for example,

patients may vary their dose intake or may intermittently take prolonged drug holidays. Even with a one-off treatment, time to (rather than simply occurrence of) treatment receipt may be particularly relevant, when timing of treatment is related to prognosis as well as outcome (for example, in the case of surgery trials).

Compliance information may need to be recorded for both treatment arms, though the relevance of both control and experimental treatment deviations will again be determined by the causal estimand of interest (and in particular whether treatment deviations should be factored out from both or just one of the treatment arms). Similarly it may be necessary to supplement data on how well the patient adhere to their original randomised prescription with information on whether the patient sought alternative treatments or counter indications.

Thus, it is necessary to implement appropriate measurement techniques in order to capture information on the particular manifestations of nonadherence that are relevant to the clinical setting, in particular to the disease, treatment and patient population being studied, and the research questions of interest. Furthermore, the measures used to record compliance should be described in sufficient detail to allow assessment of the reliability of the measurements (76).

7.3.2. Validity of compliance measurements

Until relatively recently, trialists have necessarily measured patient compliance using techniques which provide only basic and unreliable compliance information, such as pill counts, treatment diaries and patient interviews. The emergence of methods such as internal monitoring of patients' drug levels and MEMS technology have facilitated the collection of more comprehensive compliance data, but even these methods are not without their disadvantages, most notably cost (a single MEMS device currently costs

approximately £50) or invasiveness, and may be equally unreliable. For example, it is known that patients will improve their treatment adherence in the days leading up to a clinic visit, thus inflating apparent levels of compliance measured at their follow up appointments (known as 'white coat compliance') (20). Measurement of compliance and assessment of its impact become even more complicated when patients take multiple potentially interacting treatments, especially if the poor adherence to one prescribed treatment impacts on the efficacy or toxicity of others.

The validity and suitability of each adherence measurement method must be assessed in light of the particular clinical setting being studied, as variations in the definition of compliance across research or medical settings may render even the most reliable or accurate techniques unusable in a particular trial.

Given that none of the available methods provide entirely unbiased or accurate accounts of compliance, a multi-method approach combining objective and self-reporting methods has been recommended to provide the most reliable information (5). For example, it may be noted that, in contrast to the indirect account of treatment timing rather than a direct measure of treatment ingestion provided by MEMS, measurement of chemical markers in the blood provides an indication of drug ingestion but not actual timing. Thus use of both electronic and chemical monitoring methods may be necessary to obtain optimal drug compliance information (15).

Pocock (71) discusses the need to question the validity of trial compliance data rather than accepting its reliability at face value, in particular with respect to whether the compliance data are sufficiently detailed and reliable. For example, it is important to assess whether individual patient compliance data have been measured using reliable methods along with reasons for non-compliance, and that these data have been collected using the same process for each randomised group. Furthermore, trialists

should endeavour to collect outcome data on non-compliers, to avoid the bias introduced when non-compliance is associated with missing outcome data. Ultimately it is necessary to decide whether the trial demonstrates sufficient quality in design and conduct to warrant any further complicated compliance analyses.

7.3.3. Reporting of compliance data

It is important to consider not only how the compliance data will be analysed, but also how it will be summarised and reported, in order to usefully and accurately convey the compliance profile within the trial. Indeed, even when trial aims do not include adjustment for non-compliance, it is important for trialists to report on compliance levels, in order that readers (and potentially meta-analysts) can assess the similarity of the trial setting to their own clinical setting. Thus, in the same way that the appropriate measurement technique will depend on the clinical setting being studied, so too will the precise reporting or definition of compliance or adherence.

When addressing the issue of how to summarise potentially complicated compliance data (for example, time series data available from MEMS), trialists should consider the consequences of potential forms of nonadherence on the course of disease. Compliance information recorded in the trial should be sufficiently detailed and accurate to allow reporting of relevant features of non-compliance which are likely to impact on the course of disease and associated AEs (for example, allowing estimation of the incidence of drug holidays when they are likely to cause disease exacerbation, drug resistance or rebound effects).

In the case of medication adherence, measurement and reporting of adherence data should differentiate between the three phases of drug taking, namely initiation, implementation and persistence, thus distinguishing between primary non-compliance

(when patients fail to initiate randomised treatment), non-compliance with the assigned drug regimen during the implementation phase and non-persistence (when patients prematurely discontinue randomised treatment). The tool used for collecting nonadherence information will impact not only on the accuracy and reliability, but also on the level of detail, provided by any reported measure.

Initiation rates are straightforward in their calculation and interpretation, but without MEMS data (potentially corroborated with chemical monitoring), unwillingness on the part of patients to admit failing to even start their randomised treatment may lead to exaggerated initiation rates. Non-persistence rates may be equally easy to calculate but unreliable, particularly as withdrawal from treatment is typically associated with complete withdrawal from the trial. Furthermore non-persistence is unlikely to be constant over trial follow up (for example, more patients are likely to discontinue treatment early on) and as such, non-persistence should be reported per unit time (for example, as an annual treatment withdrawal rate) (78). The choice of summary measures to describe adherence during the implementation stage must relate to those features of nonadherence that impact on patient health or treatment efficacy, but will be limited according to the level of detail provided by the chosen method of adherence measurement.

In failing to provide any precise information on dose timings, pill counts can inform only on very broad patterns of treatment taking, and are commonly summarised as the average proportion of prescribed doses taken by patients in each randomised group perhaps broken down into temporal intervals (for example, annual rates). However, given that adherence is likely to fluctuate over the course of the treatment implementation period, such a single summary measure is unlikely to adequately capture important differences between patients' prescribed and observed patterns of treatment

taking. For example, a high proportion of doses taken may mask features such as drug holidays or incorrect timing, which may have a greater impact on the course of disease or side effects.

The recognised flaws in using pill counts to assess adherence are compounded when these data are used to dichotomise patients into "good" or "poor" compliers defined by a certain threshold of drug taking, most commonly according to whether patients receive at least 80% of their prescribed intervention (or between 80 and 120%, allowing for overdose toxicity). Such categorisations not only disregard the continuous nature of the relationship between compliance and response, but are often based on arbitrary criteria without any pharmacological justification, despite the fact that the required amount of treatment exposure for therapeutic benefit will vary according to the particular disease or condition, drug and its formulation (20).

Availability of more detailed information on dose timings and quantities from MEMS allows estimation of more informative summary measures quantifying more precisely the fluctuations in patient adherence over time. Variations in dose quantities and timings may be informatively combined in a single summary measure known as therapeutic coverage, defined as the proportion of patients exposed to a (predefined) minimum clinically relevant level of drug exposure, required to provide a clinically beneficial therapeutic effect, for (at least) a specified percentage of time. It may be difficult to provide clinical justification for the chosen lower (and upper, if over-dosing on treatment can cause toxicity) limit of treatment exposure required to ensure validity of this measure. Other commonly reported summary measures from MEMS data are the proportion of doses taken at the correct daily dose or the correct interval, or the average number of drug holidays taken per patient within a certain time period (19).

In differentiating between temporary and permanent discontinuation of treatment, without use of MEMS, temporary deviations (for example, occasional missed doses or longer drug holidays) are more difficult to record accurately, as are temporary deviations (such as incorrect conditions or timings of treatment-taking). On the other hand, although permanent discontinuation of treatment is more obvious and concrete, it is typically associated with end of follow up, such that important details relating to the reasons for, or the patient's clinical state at the time of discontinuation may be inaccurate or incomplete (179).

The relative importance of collecting information on temporary versus permanent discontinuations depends not only on the impact of such deviations on the drug's action and the course of disease, but also on the potential for bias resulting from each different form of nonadherence. Vrijens (20) argues that nonadherence to the treatment schedule during the treatment implementation period (in other words, temporary deviations to treatment schedule), rather than being associated with the treatment or patients' prognosis, is more likely to be caused by patient forgetfulness or suboptimal organisation. Such deviations are therefore likely to occur in a random fashion rather than being associated with a certain treatment or particular types of patients, and as such are unlikely to introduce bias if unaccounted for.

In contrast, patients who prematurely and permanently discontinue treatment are likely to do so for reasons related to their condition, difficulties in taking treatment or apparent efficacy of treatment; as such, this subgroup are likely to differ systematically from those who persist with treatment and their exclusion from analysis will likely introduce bias. Thus, although challenging, it is vital to collect information on the reasons for, and patient outcomes following non-persistence.

Reporting of adherence must also relate to the types of treatment changes expected or encouraged in the trial setting. For example, when treatment switches or additions are likely, it may be relevant to record information not only on patients' compliance with their original randomised treatment but also with any alternative treatments received, as well as the timing of, and reasons for, such changes (68). It may also be of relevance to determine who or what was responsible for the decision or request to change treatment prescription (be that the patient, treating clinician or potentially the protocol itself) and whether this decision was made in a blinded fashion (179).

7.3.4. Planning requirements

The need for relevant compliance data collection highlights the importance of considering causal estimation during the planning stage of a trial, rather than simply at the point of statistical analysis. Given that compliance is typically a multifaceted feature of patient behaviour which is difficult to measure and quantify, considerations must be made during the planning stage of which data should be collected, and how, such that analysis adjusts for clinically relevant measures of treatment received. The method used to collect compliance data will determine the format of these data and how the data may be included in the model; thus it will be necessary to consider how to collect this information accurately (considering the potential for measurement errors) and unbiasedly (recognising the typical associated biases as discussed in Chapter 3). In particular, forethought of the likely missingness mechanisms may allow procedures to be employed in order to counter such biases. Furthermore, the complexities associated with recording compliance data mean that it may be useful to pilot any CRFs prior to trial recruitment, in order to ensure sufficient clarity for treatment providers, assessors and patients.

7.3.5. Statistical analysis plan

Consideration of the statistical methods that will be applied to adjust for nonadherence must also take place during a trial's design stage, not only to demonstrate transparency with respect to the planned analyses but also to ensure collection of all necessary information required to facilitate the chosen methods of analysis. This is especially important when considering how to adjust for nonadherence, as adherence is rarely a simple dichotomous measure and may fluctuate within individual over the course of the trial, providing opportunities for manipulation of the particular definition of "nonadherence" in a certain trial in order to produce the most favourable results, for example by excluding certain patients with particularly good or poor prognoses (35, 71).

As such, in order to avoid accusations of bias, a specific analysis plan should accompany every trial protocol, providing technical details of planned statistical analyses (61). Ideally, this plan should include definitions of "nonadherence" and whether, and if so how, the efficacy analysis will be adjusted for any nonadherence. These analyses should be linked to the research questions of interest, which then determine the corresponding forms of nonadherence which need to be factored out in order to investigate these questions. This may be a challenging exercise, given the difficulty in predicting all forms of participant or clinician nonadherence that will occur in a trial and therefore in defining precisely how particular patients' data will be analysed (which may explain why, despite the argument for upfront transparency, decisions regarding compliance analyses are often made post hoc) (71).

Cox (70) argues, however, that although it is necessary to provide a general plan of statistical analysis, it may be unrealistic to require analysts to stick rigidly to specific analysis plans, and that, following analyses carried out according to the original plan,

there may be justifiable reasons for making amendments to specific analyses. This will be demonstrated in practice in the next chapters of this thesis.

7.4.Use of compliance data in statistical methods

Collection of compliance information must reflect, not only the **clinically relevant aspect** of adherence defined by the causal question, but also how this compliance information will be used in the modelling process. Thus it is necessary to consider how the data that can be collected in practice (with all their likely limitations) can then be manipulated into useable forms for the modelling exercise.

The statistical methodology chosen to estimate particular causal estimand of interest will rely on availability of (compliance, covariate and outcome) data in the appropriate format. Collected data may therefore need to be transformed into a usable format for the chosen method of statistical analysis, in order to simplify modelling and its subsequent interpretation. For example, continuous dose data may be reduced to a single summary quantity, for example as the proportion of prescribed treatment received over the whole treatment period, or longitudinal treatment withdrawal data may be summarised as a time to event variable.

Similarly it is necessary to identify which **covariates** will likely be related to the occurrence of treatment changes and outcome, and whether these are likely to be recorded as baseline or time-varying covariates or confounders (which, as well as affecting future treatment, may furthermore be affected by prior treatment). It is vital therefore to consider what information clinicians (or patients) actually use when deciding to change treatment. Once it has been determined which of these factors are likely to be important, again it will be necessary to consider how to collect this

information to provide reliable (accurate and unbiased) data, again considering the likely mechanisms which will lead to missing data.

The chosen statistical methodology must of course also depend on the form of the outcome data of interest (typically either binary or continuous) and on the frequency and nature of the measurement of the outcome, which may be taken at a single time point or repeatedly over time, or may be recorded as a time to event variable. Furthermore, the missingness mechanism for each type of data included in the trial will impact on the modelling limitations and capabilities and its subsequent interpretation.

Available data will then need to be incorporated into appropriate models, taking into consideration modelling aims (research question and corresponding causal estimand of interest), inherent data limitations and statistical capabilities and limitations of the model. In particular, one should consider the assumptions implied by the model, and check the robustness of the results to these assumptions using sensitivity analyses (varying parameters or assumptions) or alternative models.

Therefore, prior to investing resources and time in the potentially challenging task of collecting compliance data, trialists should consider how these data are likely to be used in adjustment methods, in a manner which directly relates to the causal question of interest. It may be necessary to manipulate the collected raw compliance data into a format usable by the chosen model, such that the original compliance data recorded in patient CRFs can be transformed into a summary measure for inclusion in the model. However, the clinically relevant features of compliance must also be captured in these measures, for example it may be relevant to model time to treatment compliance rather than a simple summary measure of average compliance over the whole treatment period (180).

7.4.1. Manipulation of compliance data for modelling

Adherence data may take the form of (or may need to be transformed into) a constant or time-varying quantity, summarised as a binary, ordinal, continuous or time to event measure.

The form taken by the adherence data will determine the interpretation of the model. This may be most easily observed in the case of naïve models: for example, a time-fixed measure of compliance (or treatment received) may be included in a naïve model in place of the original randomised treatment group indicator. The simplest example of all-or-nothing compliance (for a one-off treatment, say) can be easily summarised as a binary factor; the corresponding naïve PP analysis would include only those patients whose randomised treatment indicator (Z) was equal to their received treatment (A), whereas an AT analyses would replace the randomised treatment indicator (Z) with the treatment received (A) for all patients. If more complicated data are collected, for example, when patient dose data are summarised into the proportion of total randomised drug actually taken by the patient, a dose-response model may be used to estimate the treatment effect, allowing for cumulative dose received. When treatment is time-varying, an AT analysis may feature a time-varying treatment covariate, for example a time-dependent treatment covariate in a standard Cox model.

7.4.2. Binary compliance

If treatment is a one-off event (for example, in the case of a single application or administration of treatment, such as a vaccine), compliance is naturally summarised as a binary factor and the associated causal effect may simply be defined as contrast between receiving experimental treatment versus standard (or no) treatment, as appropriate. This "all-or-nothing" compliance is the most straight forward form of compliance data to

analyse, as it provides a natural dichotomisation of patients into compliers and non-compliers and resulting binary contrasts to be included in the model.

Furthermore, analysis is simplified further if non-compliance only occurs in one, rather than both, treatment arms. For example, if trial design is such that control patients receive standard care as usual and are not given the option of beginning the novel experimental treatment (such that contamination is not possible) but those randomised to experimental treatment are able to refuse initiating this treatment, the only type of non-compliance that may be of interest is the simplest type of binary compliance categorization, all-or-nothing compliance in one treatment arm only, such that experimental patients either start and continue with allocated treatment or do not receive any of their allocated treatment.

However it is rare in practice for compliance to be a purely dichotomous feature. Instead, the typical multifaceted nature of treatment protocol compliance means that patients usually demonstrate various degrees of partial compliance, rather than full or no compliance. Furthermore, rather than being fixed over time, compliance may fluctuate according to patient's conditional or other external factors. Thus data collected on compliance is often continuous in nature, and may be time-dependent.

However, partial or time-varying compliance data are more complicated to analyse, requiring more identifiability assumptions. Complex compliance information requires definition of multiple potential outcomes and treatment contrasts, and in the case of PS methods, additional principal compliance categories (119). Thus, even when clinical scenarios are complex, compliance data may need to be categorised in order to simplify modelling.

For example, continuous or time-varying measures of compliance may need to be dichotomised in order to allow methods to be carried out which can accommodate only binary compliance categorisation. Whenever compliance data have been reduced, however, it is necessary to consider the sensitivity of results to the (potentially arbitrary) cut-off level used to define the dichotomy, as it may distort the resultant treatment effect by including disparate types of patients within each category (71).

7.4.3. Time-varying treatment

The more complex the prescribed treatment schedule, the more complicated the definition of regime "compliance" becomes and the more consideration is required regarding collection of data to inform compliance-adjusted analyses. These considerations must all be taken in light of the particular causal effect of interest.

When treatment is longitudinal in nature, the associated definition of causal treatment effect in the case of longitudinal treatment is therefore likely to demand greater consideration than for all-or-nothing compliance. The causal effect may simply take the form of a comparison between "continuous treatment with experimental treatment during trial treatment period" versus "continuous treatment with standard treatment during trial treatment period". The clinical scenario or research question of interest may, however, demand a more precisely defined causal effect estimate. For example, in a pragmatic trial reflecting a clinical setting where side effects or drug inefficacy may lead to alterations to prescribed treatment in practice, it may not be appropriate to factor out *all* changes to initial randomised treatment, as some treatment changes will necessarily occur in practice; for example, it may instead be of interest to estimate causal effects of treatment taken continuously but only until the point of toxicity or other counter-indications (87).

Thus information collected on compliance in a trial setting must relate directly to the causal effect of interest, in order to inform the associated causal analyses.

7.4.4. Sensitivity analysis

Given that it is not possible to estimate causal effects without invoking untestable assumptions, it follows that sensitivity analyses to investigate robustness of results to each analysis' particular assumptions should be a matter of course. Analyses may be repeated considering a wider range of assumptions, as demonstrated for example by Vansteelandt (116) and Dunn (181).

Furthermore, it may be necessary to check sensitivity of results when compliance data have necessarily been simplified; for example, it may be important to check whether conclusions alter when changes are made to (potentially arbitrary) dichotomising definitions of continuous measures of compliance, or if multiple aspects of exposure to treatment (for example, frequency and dose) have been combined into a single summary (116).

When undertaking causal analyses, it is also important to appreciate the particular nuances of the clinical setting, in order to ensure that these contrasts are clinically as well as statistically relevant. For example, White's analyses of the MRC hypertension trial (68, 89) demonstrate the need for an understanding of the subtleties of the particular disease or treatment being studied before one can confidently produce and interpret causal findings from an analysis. Effective communication between the statistical analysts and the clinical experts for a particular trial is therefore of paramount importance to ensure such nuances are not overlooked in the analysis, and to prevent statistical anomalies or biased interpretation of trial results.

7.4.5. Summary

Thus, despite the greater control afforded by a research setting, barriers to accurate collection of trial compliance data are many, often mirroring those arising in general practice. The methods typically used in trials have been discredited, as they are so often easy to falsify (pill counts), rely on unrealistic or biased recall by patients (patient interviews), or may be resisted by patients because of the daily reminder of their disease presented by the recording method (treatment diaries). Similarly health care providers who (in their opinion, justifiably) deviate from randomised treatment regimens when prescribing or administering patients' treatment may prefer not to disclose such treatment protocol deviations. An added dimension of complexity arises if therapeutic action of medication is provided only under certain conditions, for example when swallowed whole or chewed, with or without food. Assessing whether patients consistently followed such instructions when taking long term treatment is particularly difficult to ascertain.

The compliance measure actually used for analysis will depend on a number of factors. First, the clinical setting and research question will determine the measure of compliance that would be most clinically relevant. However, the complexity and detail of available data will depend on the compliance measurement method used to measure compliance during the course of the trial (for example, pill counts, treatment diaries or MEMS). Finally, the chosen statistical model may impose restrictions on the compliance measurement that can be included in the model (for example, the method may be restricted to handling a time-invariant (for example, single binary or continuous measure) rather than a time-varying measure of compliance).

In order that suitable chosen methods of analysis can be implemented in practice, it is therefore necessary for researchers to consider certain issues at the design stage of the

trial. Appropriate planning is required to ensure the necessary data are collected, not only regarding treatment receipt and outcome, but also on all potential confounding factors that may need to be accounted for. Furthermore, statistical analysis plans should be developed a priori to ensure that trialists are not accused of altering analysis techniques once outcome data have been collected and observed in order to obtain optimal results.

7.5.Pictogram

Two pictograms are presented intended as a research aid to simplify the causal estimation process, first by guiding researchers through the necessary considerations for each step of the causal analysis process and then by summarising which statistical method would be most appropriate depending on the trial scenario.

The first pictogram (Figure 2) can be used to clarify how relatively complex trial compliance scenarios can be converted into answerable research questions and estimable causal effects, while also highlighting the need to consider the limitations and assumptions underlying the chosen analysis. For example, trialists may feel overwhelmed when faced with the likelihood of numerous forms of deviation from randomised treatment which typically occur with prescription of long term medication in trials and in clinical practice (for example, treatment switches, additions, withdrawals (permanent and/or intermittent) and incorrect treatment administration) potentially in both treatment arms and involving trial and non-trial treatments. In such cases, rather than attempting to collect (potentially unreliable) data on all sources of nonadherence to randomised treatment, it will be simpler to identify which features of nonadherence will impact on the outcome, and then focus on how to obtain accurate data on these features alone, which can meaningfully be used to inform relevant analysis techniques.

Once data collection and manipulation issues have been considered, it is necessary for the researcher to then determine the most appropriate statistical methodology to apply.

Thus, the second pictogram (Figure 3) is also intended as an aid to the researcher, who must consider the requirements and implications of each option available. The statistical literature review summarised in Chapter 5 demonstrated that the available methods of adjustment varied in their underlying assumptions and data requirements, not only regarding the outcome but also the compliance and covariate information. To a non-specialist statistician or trialists, the choice may be somewhat confusing. As such, the pictogram provides a reminder of the statistical issues and appropriate methods to use when faced with a certain trial scenario.

As such, this pictogram aims to guide analysts in their thought processes regarding the underlying statistical complications of their particular trial, relating not only to the confounders which will may need to be allowed for in the chosen statistical method, but also regarding the choice of method, depending on the available data, assumptions and outcome type.

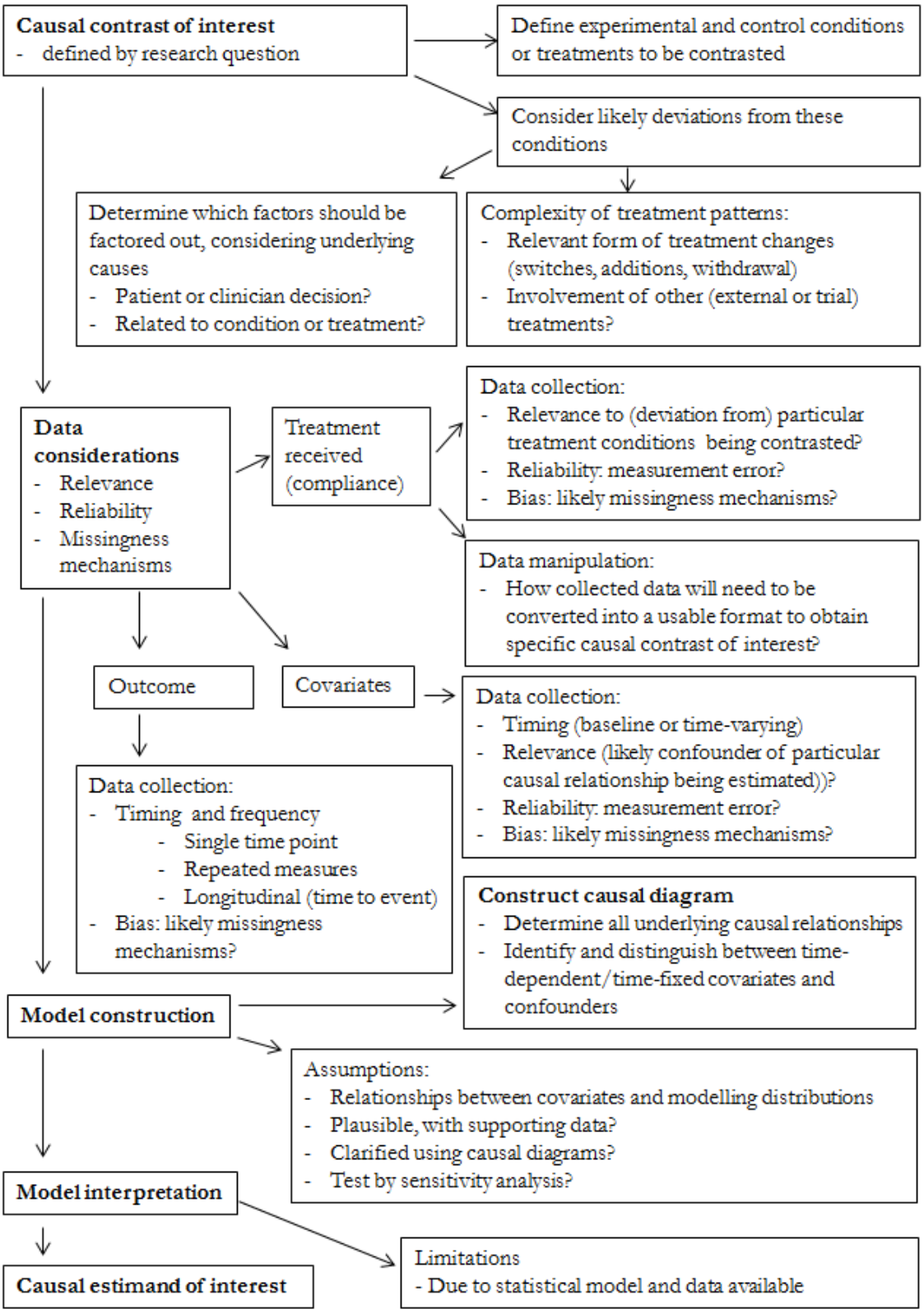


Figure 2 Causal estimation process

For example, the pictogram reminds the user that standard statistical methods can be used to adjust for confounding in the absence of any time-varying confounders, assuming all baseline confounders are known, measured and can be accounted for. However, in the absence of information on all confounders or if time-varying confounders are likely to affect both treatment received and outcome, alternative methods should be considered.

Alternatively, if a particularly appropriate method of analysis requires compliance or covariate data to take a certain format, the chosen method of analysis may determine the format necessary for the data being used in the model.

Furthermore the pictogram reminds the user that the interpretation of these methods varied depending on the underlying assumptions and causal estimation framework. Observational-based IPW methods assume that all confounders have been measured and adjusted for in the calculation of weights. The two forms of randomisation-based methodologies differ in their interpretation, which in turn relate to their relevance in answering a particular causal research question: PS (intervention-based) methods estimate treatment effects amongst the (latent) subgroup who would comply with whichever of the two treatments being contrasted they were assigned, whereas (efficacy-based) structural models estimate causal treatment effects for the entire population, thus assuming that all patients would be able to (or could be persuaded to) comply with their randomised assignment.

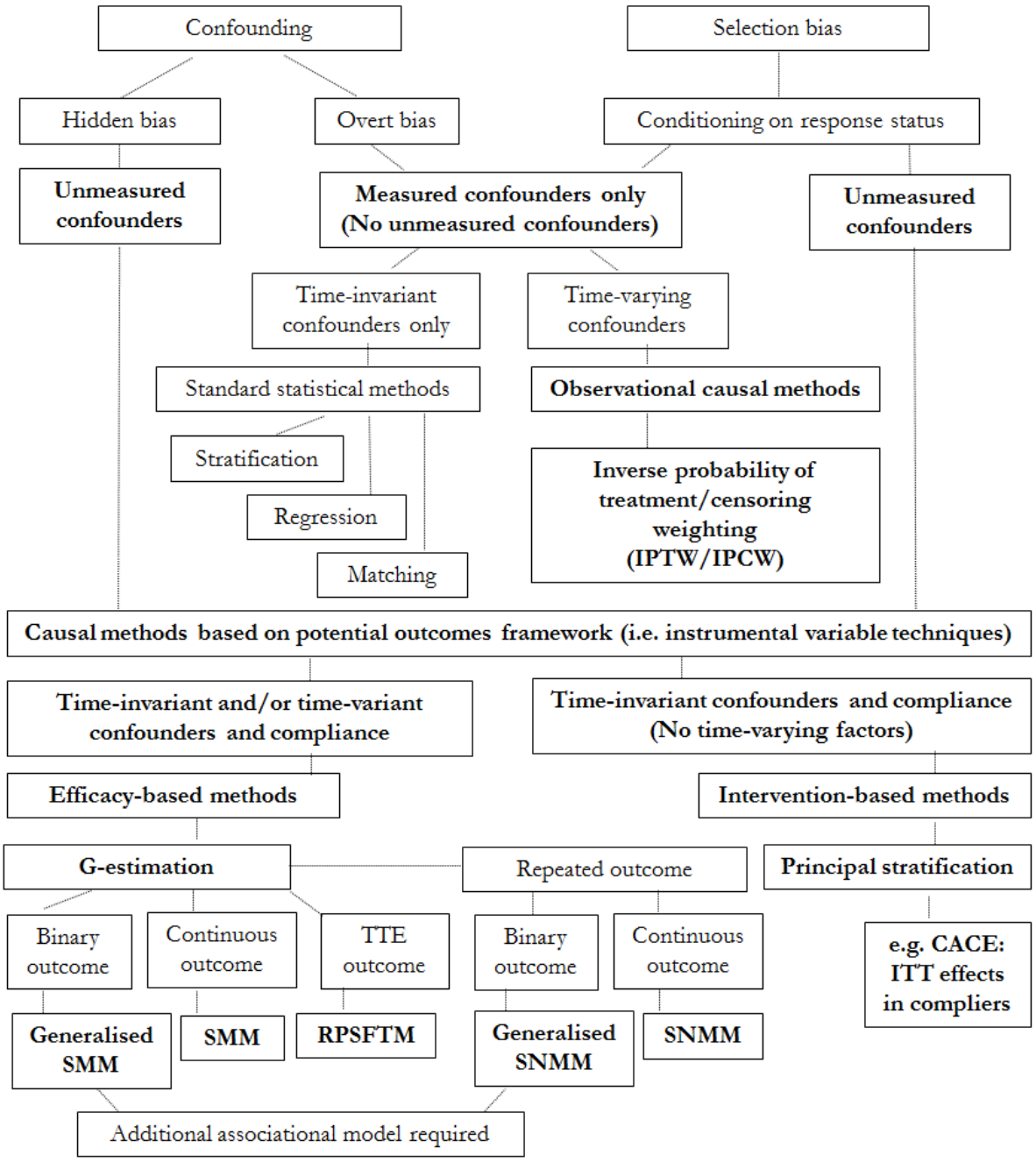


Figure 3 Statistical methods and considerations for causal analysis

7.6.Causal diagrams

A further aid to researchers when undertaking causal analyses may be to create a causal diagram relating to the causal scenario being considered. Causal diagrams help to qualitatively convey the nature of relationships between variables and the underlying implicit assumptions being made when making causal inferences regarding these relationships. In demanding explicit expression (and underlying assumptions) of relationships between causes, effects, confounders and selection variables, causal diagrams may help to identify potential forms of bias due to confounding and selection, as well as guiding plans on data collection (for example, regarding which variables will need to be included in analysis) and methods of analysis (182).

All statistical methods depend on underlying (often unverifiable) assumptions, and causal inferences from such models require even more assumptions than those required for associational inferences. Causal diagrams help to convey these additional assumptions in a manner that cannot be portrayed using standard statistical notation, thus highlighting whether these assumptions are reasonable, and the extent to which any variations in these assumptions may affect conclusions. Furthermore, causal graphs are useful in helping investigator to think about what causal relationships may be missing from the graph and whether the proposed analysis methods will appropriately adjust for these underlying relationships (182).

Thus, although causal diagrams are usually associated with more traditional methods (SEMs), they can be useful in guiding analysts, regardless of the exact method employed for causal estimation, by clarifying the likely causal links in their particular trial scenario as well as highlighting the underlying assumptions and relationships which must be appreciated when undertaking their analysis.

7.7.Conclusion

In order to address potential barriers for the obvious lack of wider appreciation by the trial community of the importance of adjusting for departure from randomised treatment in a manner that will not introduce selection or confounding bias, this chapter has presented recommendations and guidelines to trial analysts in order to simplify and facilitate wider use of available causal estimation techniques. In the subsequent chapters of this thesis, this aim will be further realised through the presentation of practical demonstrations of adjusting for treatment deviations, thereby illustrating not only some typical challenges arising during the causal estimation process but also how potential solutions to these problems may be developed.

8. Demonstration of novel applications of the RPSFTM

The review of statistical methods in Chapter 5 introduced a number of methods to adjust for treatment deviation in trials, but it became apparent from the review of published trials summarised in Chapter 6 that these methods are rarely employed, potentially because of their complexity or simply a general lack of awareness that such methods exist. Thus in Chapter 7, the range of available methods were presented pictorially with the aim of guiding researchers on the most appropriate methodology to use, along with a pictogram highlighting the pertinent practical issues relating to model requirements and interpretation. In this and the remaining chapters of this thesis, these recommendations are demonstrated in practice; in particular, the use of these pictograms are put into practice, demonstrating how a relatively complicated trial

scenario can be simplified into a feasible research question setting to facilitate analysis, as well as helping to highlight the limitations and assumptions associated with the application of the chosen analysis techniques.

Thus, in this and the following chapters, the aim of this thesis is achieved with novel application of two methodologies appropriate for survival outcomes in contrasting trial settings, demonstrating necessary considerations and practical solutions to the challenges arising when applying these methods, and providing an example to future researchers on implications and interpretation of important causal methods. The general format of these chapters is designed, not simply to aid the reader in understanding the analysis process required for the purposes of these particular trials, but also as to demonstrate to other researchers the level of thought and detail required prior to, and during, application of these methodologies.

8.1.Introduction

The review of statistical methods for adjusting for nonadherence to randomised treatment, summarised in Chapter 5, revealed the most appropriate methods to adjust for departure from randomised treatment when analysing survival outcomes, namely the RPSFTM and the IPCW modelling techniques. In this chapter, the first of these methods will be applied to three trials introduced in Chapter 4: the honey trial, SANAD arm A and SANAD arm B. These trials present issues and scenarios of increasing complexity with respect to adjustment for nonadherence. In particular, the SANAD trial provides numerous analysis issues and challenges, due to the nature of the trial design and treatment changes that occurred. These challenges are likely to be common in other chronic disease areas, where treatment changes occur for a number of reasons

and interest lies in achieving remission of symptoms. As such, this trial provides a useful illustration for discussion of the analysis issues and methods.

8.1.1. Model considerations

This chapter will demonstrate the RPSFTM and compare the results and conclusions with those of standard methods (using ITT and PP analysis populations).

As with any causal estimation, it is necessary to match the causal research question with the model treatment parameters; furthermore, given that the RPSFTM only reliably adjusts for one causal contrast (between experimental and control treatments), it is necessary to determine which single form of treatment deviation should be accounted for. As introduced in Chapter 5, a simple version of the RPSFTM can be presented as follows

$$U_{0i}(\beta) = T_{0i} + e^{\beta}T_{1i}$$

which demonstrates the first key property of the RPSFTM, namely that it assumes that there are only two exposures within the clinical trial. For example, T_{0i} may be defined as time spent on control intervention and T_{1i} as time spent on experimental intervention, such that e^{β} reflects the expansion (or contraction) in survival time attributable to treatment (130).

In practical terms, this means it is possible to adjust for only one form of treatment change, that which is most easily accommodated being direct switches between randomised treatments, such that each patient's follow up time is comprised solely of time spent on either control or experimental treatment (rather than alternative non-trial treatments or withdrawal from treatment altogether). This in turn implies that follow up is necessarily censored at the point of any other form of treatment change.

Therefore, as will be demonstrated in this chapter, it is necessary to define "on" and "off" treatment times (T_{0i} and T_{1i}) relating to the particular research question of interest.

Furthermore, as discussed in Chapter 5, given the need to recensor individuals to remove the dependence (on the U_0 scale) between censoring time and treatment received, it is necessary to define each individual's maximum censoring time.

These model requirements will be addressed in each application of the RPSFTM demonstrated in turn for each of the trials using the "strbee" code in Stata (134). Stata software is particularly user-friendly, making it ideal for this demonstration of the RPSFTM for non-specialist statisticians.

8.2.Honey trial

Departure from randomised treatment in the honey trial occurred in the form of withdrawal from randomised treatment, switches to the alternative treatment (honey or conventional treatment) or to a different course of treatment altogether (surgery, radiotherapy or antibiotics) or loss to follow up.

Although it is ethically necessary to allow patients to receive alternative or more powerful treatment if they experience side effects or wound deterioration, the aim of this trial was to ascertain the biological efficacy of honey treatment when compared to the best standard care. As such, it was of interest to estimate the causal effect of honey treatment compared with conventional dressings *in the absence of any treatment changes.*

However, given that the RPSFTM can only handle one treatment contrast, which here is taken to be the comparison between conventional and honey dressings, any time

spent receiving alternative treatment is necessarily discounted. Thus, in the model below,

$$U_{0i}(\beta) = T_{0i} + e^{\beta}T_{1i}$$

T_{0i} is time spent with conventional wound dressing, T_{1i} is time spent with honey applied to wound and any time spent in receipt of other forms of treatment is ignored. Follow up continued until a potential maximum of 24 weeks from randomisation (which thus defines the maximum censoring time).

The RPSFTM is compared here for the primary outcome (time to healing) and secondary outcome (time to 50% reduction in wound size) against the results obtained using three other analysis sets, namely an intention-to-treat (ITT) analysis set, a per protocol (PP) analysis set and an as treated (AT) analysis set.

In the ITT analysis, patients were analysed in the group to which they were originally randomised, and thus treatment switches were ignored. However, given that the trial comparison was intended for wounds healing by secondary intention (rather than due to direct invention on the wound), when patients were given more extreme forms of treatment (radiotherapy, surgery or antibiotics) or were lost to follow up, their follow up was censored (98). This censoring is potentially highly informative, given that the need for more radical treatment indicates a poor wound prognosis, and loss to follow up often occurred for reasons related to the satisfaction or motivation of the patient, which again are not likely to be independent of their prognosis. Death introduced a competing risk but for the sake of simplicity, this was ignored here and censored in the usual manner.

The PP analysis set only included patients who received the treatment to which they had been randomised, and censoring was introduced at the point of any deviation from

this treatment (thus excluding patients altogether when they switched to the alternative trial treatment at the start of the trial).

The AT analysis used a time-invariant binary measure of treatment received by categorising patients according to the treatment received from the start of the treatment period, and again censored follow up at the point of any deviation from the starting treatment. Thus AT analysis differed from the PP analysis set by including patients who switched treatment at randomisation in the group to which they switched (thus ignoring randomisation completely).

These variations of analysis were compared with the results from a RPSFTM which adjusted for treatment switches to the alternative trial treatment, but in which censoring was again necessary when patients received more extreme (non-trial) interventions or were lost to follow up. Thus the four analyses differed only in their handling of patients who switched to the alternative trial treatment; all other deviations from treatment protocol were necessarily censored.

8.2.1. Analysis

The `strbee` program in Stata requires that the user creates additional variables capturing information on whether (and if so, when) the patient switched between treatment arms. These variables are treatment arm specific (for example `honey_switch_t`, `honey_switch`, `conv_switch_t`, `conv_switch`) and equal to zero if patients did not experience treatment switches. For example, a patient who switched from honey to conventional treatment 20 days following randomisation would be assigned values equal to (20, 1, 0, 0) for these four variables respectively, whereas a patient randomised to receive honey who did not switch would be assigned a value of 0

for each of these variables. The `strbee` program can analyse only binary treatment comparisons and requires that treatments are defined as 0 (control) or 1 (experimental).

Prior to using the `strbee` program, it is necessary to specify the event time (`t_healing`) and indicator (`h_status`) variables using `stset`:

```
stset t_healing, failure(h_status)
```

The user then must indicate the switch timing and indicator variables (in that order) for the control and experimental arms separately within the brackets preceeded by `xo0()` and `xo1()` respectively. For example, the code with randomised treatment indicator `rand_trt` would simply be

```
strbee rand_trt, xo0(conv_switch_t conv_switch) xo1(honey_switch_t honey_switch)
```

Options include `hr` (to display results in terms of HR rather than AF), `kmgraph` to display KM curves and `trace` for a list of recensoring times.

8.2.2. Sensitivity analysis

The censoring of patients at the point of deviation from treatment protocol for reasons related to severe wound deterioration (when more radical clinical intervention was given in the form of antibiotics, radiotherapy or surgery, or when the patient was unable to attend clinic because the wound was so debilitating) is likely to be especially biased, given that the state of their wounds is not likely to be typical of all wounds in their randomised group at this time point; the prognosis for the healing of these wounds at the point of censoring is very poor. Thus a sensitivity analysis was carried out in the ITT, AT and RPSFTM analyses, reflecting the likelihood that such wounds would likely not have healed in the follow up time even if treatment and follow up had continued as per the protocol. Thus, when clinical intervention occurred due to concern over the

poor progress in healing process (for example, when a patient began taking antibiotics or was referred for radiotherapy or surgery) or when the patient was unable to attend clinic because the wound was so debilitating, instead of censoring follow up at these points of treatment change or loss to follow up, follow up was censored at the longest possible follow up (24 weeks) for the sensitivity analyses (in other words, assuming that the outcome event did not occur during this follow up period).

The results from these analyses, and the corresponding sensitivity analyses, are given in Table 22 to Table 25 below.

8.2.3. Results

105 patients were randomised to receive medical grade honey or conventional treatment for treatment of wounds. The number of treatment changes that occurred in each treatment group are presented in Table 21. There were only four (11.5%) switches from honey to conventional treatment (two immediately following randomisation) and six (11.3%) from conventional to honey treatment (all immediately following randomisation). Ten (19%) other patients from each treatment arm experienced another form of deviation from treatment (or follow up) protocol; these patients' follow up time was extended to the maximum (24 weeks) for the sensitivity analysis to reflect their poor prognosis at the time of treatment change or loss to follow up.

The results in Table 22 and Table 23 demonstrate that the RPSFTM did not have a large impact on trial conclusions, not surprisingly given that only a small proportion of treatment deviations were accounted for (four out of 14 (28.6%) in the honey arm and six out of 16 (37.5%) in the conventional arm). In fact, the impact of censoring at the maximum censoring time rather than at the point of receiving extreme treatment (in the sensitivity analysis) was generally similar to that of the adjustment for between-arm

treatment switches using the RPSFTM; both the RPSFTM and sensitivity analysis tended to increase the acceleration factor (AF = e^{β}), implying that these adjustments both increased the effect of honey on healing compared with standard dressings.

The AF is interpreted in terms of the rate at which time to healing is "used up" when honey is applied in comparison to standard conventional dressing. Thus the AF (95% CI) from the ITT equivalent of the SFTM (with no adjustment for treatment changes) suggest that healing is achieved sooner with honey, on average 1.27 times faster than with conventional dressing, though this relative effect may extend between 0.80 and 2.33 times; in other words, the result is not statistically significant. When the sensitivity analysis is applied (when censoring for alternative (extreme) treatment changes occurs at the maximum follow up time (24 weeks) but again without adjustment for any such treatment switches), healing time is "used up" even faster with honey (AF 1.31) and the upper confidence limit increases.

This difference in AF attributable to the sensitivity analysis (i.e. between the ITT version of the SFTM with and without the sensitivity analysis applied) is exactly the same as that observed when the RPSFTM correction is applied. Application of the sensitivity analysis in the RPSFTM adjustment further increases the AF (to 1.4 and 1.38 with and without recensoring respectively).

Recensoring in the main RPSFTM analyses did therefore not drastically influence results, as very few observed events were recensored (only two patients, both from the honey arm, had their event times censored). However, in the sensitivity analyses for both time to healing and 50% reduction, recensoring survival times reduced the estimated AF (1.38) compared to no recensoring (1.4); this is expected, given that follow up time was greatly extended for these patients (to the maximum of 24 weeks), increasing the probability that recensoring would alter this survival time.

Similar results are observed in Table 23 for the secondary outcome, time to 50% reduction in wound size.

As described in Chapter 5, the AF obtained from RPSFTM can be converted to HRs if one assumes a constant hazard ratio (i.e. according to a Weibull distribution). Thus, the corrected HRs obtained from the conversion of the RPSFTM AF are given in Table 24 and Table 25 below, to allow comparison with the results from the other (ITT, PP and AT) analysis sets. It can be seen that the effect of RPSFTM (and corresponding sensitivity analyses) is intermediary between the ITT and PP/AT analyses. This is intuitive, given that the RPSFTM in fact censors for many of the treatment deviations in a manner identical to that of the PP analysis, with only one third of the treatment switches being accounted for in the model. This demonstrates the main limitation of the RPSFTM in this context – an inability to handle more than one type of treatment change.

The results in Table 24 and Table 25 also demonstrate the effect of analysing according to treatment received, as the HRs from the PP and AT analyses are more extreme in favour of honey, even approaching statistical significance (0.085 for time to healing from AT analysis and 0.068 for time to 50% reduction in wound size from PP analysis). The bias in favour of honey is expected, given that the majority of switches between treatments occurred when clinicians transferred young fit patients to receive honey despite randomisation to control. Thus, by censoring their follow up in conventional arm (in PP analysis) or attributing their healing to honey (in AT analysis), the effect of honey was enhanced as demonstrated by the increase in HR.

Table 21 Compliance information

All values are number (% of randomised treatment group).

	Randomised treatment	
	Conventional treatment (n=53)	**Honey treatment (n=52)**
Withdrawal from trial		
Death	2 (3.8)	1 (1.9)
Lost to follow up	0	1 (1.9)
Wanted honey	1 (1.9)	0
Did not want to attend clinic	1 (1.9)	1 (1.9)
Dissatisfied with progress	0	1 (1.9)
Clinical intervention during the course of the trial due to severity of wound[1]		
Antibiotics	1 (1.9)	1 (1.9)
Radiotherapy	0	1 (1.9)
Too ill to attend clinic[4]	0	1 (1.9)
Surgery	3 (5.7)	1 (1.9)
Switched to other treatment during the course of treatment		
Pain	0	1 (1.9)
Deterioration of ulcer	0	1 (1.9)
Switched to other treatment immediately following randomisation		
Patient choice	1 (1.9)	1 (1.9)
Clinician choice	5 (9.4)	1 (1.9)
Withdrawal from trial immediately following randomisation		
Surgery	0	1 (1.9)
Radiotherapy	0	1 (1.9)
Disappointed with allocation	1 (1.9)	0
Transferred to other hospital	1 (1.9)	0

Table 22 RPSFTM for time to healing

		e^{β} (95% CI)
No RPSFTM adjustment (equivalent to ITT)		1.27 (0.80, 2.33)
Sensitivity analysis[1]		1.31 (0.80, 2.57)
RPSFTM adjustment		
	No recensoring	1.31 (0.75, 2.52)
	Recensoring[2]	1.31 (0.75, 3.29)
Sensitivity analysis[1]		
	No recensoring	1.40 (0.75, 3.07)
	Recensoring[3]	1.38 (0.75, 3.44)

[1] Sensitivity analysis censors patients with changes to alternative treatment at the maximum follow up time (24 weeks) rather than at the observed time of treatment change.
[2] Number of patients (events) recensored: conventional: 0 (0), honey: 9 (2)
[3] Number of patients (events) recensored: conventional: 1 (0), honey: 12 (2)

Table 23 RPSFTM for time to 50% reduction in wound size

		e^{β} (95% CI)
No RPSFTM adjustment (equivalent to ITT)		1.33 (0.75, 2.27)
Sensitivity analysis[1]		1.33 (0.68, 2.67)
RPSFTM adjustment		
	No recensoring	1.38 (0.67, 2.81)
	Recensoring[2]	1.38 (0.69, 2.81)
Sensitivity analysis[1]		
	No recensoring	1.42 (0.59, 3.50)
	Recensoring[3]	1.38 (0.62, 3.23)

[1] Sensitivity analysis censors patients with changes to alternative treatment at the maximum follow up time (24 weeks) rather than at the observed time of treatment change.
[2] Number of patients (events) recensored: conventional: 0 (0), honey: 9 (2)
[3] Number of patients (events) recensored: conventional: 1 (0), honey: 12 (2)

Table 24 Comparison of analyses for time to healing

	Median time (IQR) [range]		HR (95% CI)	p-value
	Conventional (n=51)	**Honey (n=50)**		
ITT	140 (91, -) [59, -]	100 (51, 142) [42, -]	1.30 (0.77, 2.19)	0.316
Sensitivity analysis	140 (91, -) [59, -]	100 (54, 162) [42, -]	1.32 (0.79, 2.23)	0.289
RPSFTM				
No recensoring	140 (98, -) [59, -]	100 (54, 142) [42, -]	1.37 (0.74, 2.56)	0.316
Recensoring	140 (98, -) [59, -]	100 (54, 142) [42, -]	1.27 (0.80, 2.03)	0.316
Sensitivity analysis				
No recensoring	147 (98, -) [59, -]	114 (54, 162) [42, -]	1.42 (0.74, 2.73)	0.289
Recensoring	145 (98, -) [59, -]	114 (54, 162) [42, -]	1.45 (0.73, 2.86)	0.289
PP	140 (84,-) [49, -]	91 (51, 142) [42, 163]	1.48 (0.86, 2.55)	0.161
AT	140 (84,-) [59, -]	100 (54, 132) [42, 163]	1.60 (0.94, 2.72)	0.085
Sensitivity analysis	154 (84, -) [59, -]	105 (55, 133) [42, -]	1.54 (0.91, 2.62)	0.111

Table 25 Comparison of analyses for time to 50% reduction in wound size

	Median time (IQR) [range]		HR (95% CI)	p-value
	Conventional (n=51)	Honey (n=50)		
ITT	46 (23, 61) [16, 159]	32 (20, 48) [16, 101]	1.29 (0.82, 2.03)	0.266
Sensitivity analysis	46 (28, 68) [16, -]	32 (20, 48) ([16, 120]	1.24 (0.79, 1.96)	0.355
RPSFTM				
Not re-censored	46 (23, 68) [18, 159]	32 (20, 48) [16, 101]	1.36 (0.80, 2.33)	0.260
Re-censored	46 (23, 68) [18, 159]	32 (20, 48) [16, 101]	1.31 (0.84, 2.06)	0.237
Sensitivity analysis				
Not re-censored	46 (28, 70) [20, -]	32 (20, 48) [16, 120]	1.30 (0.75, 2.24)	0.350
Re-censored	46 (28, 69.1) [19, -]	32 (20, 48) [16, 120]	1.25 (0.78, 1.99)	0.350
PP	46 (23, 68) [18, 165]	29 (20, 48) [15, 95]	1.54 (0.97, 2.43)	0.068
AT	47 (28, 70) [20, -]	29 (20, 48) [15, 105]	1.40 (0.88, 2.22)	0.157
Sensitivity analysis	46 (20, 61) [18, 159]	29 (20, 48) [15, 95]	1.45 (0.91, 2.32)	0.119

8.2.4. Conclusion

The application of the RPSFTM to these honey trial data suggest that adjusting for switches tends to increase the treatment effect in favour of honey to an intermediate degree to that observed from PP or AT analyses. However this analysis was limited due to very small numbers of direct switches between treatment arms, and the need to censor follow up at the point of any other change in treatment. Furthermore, as well as being unable to adjust for switches to non-trial treatments, the RPSFTM was subject to bias due to an inability to appropriately handle competing risk events such as death. As will be discussed in the next chapter, it is potentially possible to adjust for such competing events using the RPSFTM in conjunction with IPW methods. However, prior to considering this alternative method, the RPSFTM will now be demonstrated in the larger, more complex scenario present in the SANAD trial.

8.3. SANAD

Despite the overtly pragmatic nature of the SANAD trial, involving inevitable treatment changes when patients experienced UAEs or ISC, there was interest in the efficacy of treatments in the absence of such treatment changes. Prior to considering RPSFTM specifications, however, it is necessary to discuss a number of complications in this trial which make any causal analysis somewhat challenging.

8.3.1. Data issues

First the treatment receipt data available in SANAD relate only to the drugs and doses prescribed by clinicians rather than actual adherence of patients to these prescriptions.

The only compliance data compromises a single general question on an annual quality of life questionnaire asking patients to estimate how often they miss doses. The distribution of answers is given in Figure 4 (where NR refers to non-response); although less than 5% of patients who returned the questionnaire each year reported missing medication more than once a week, at least one third of patients did not return questionnaire each year.

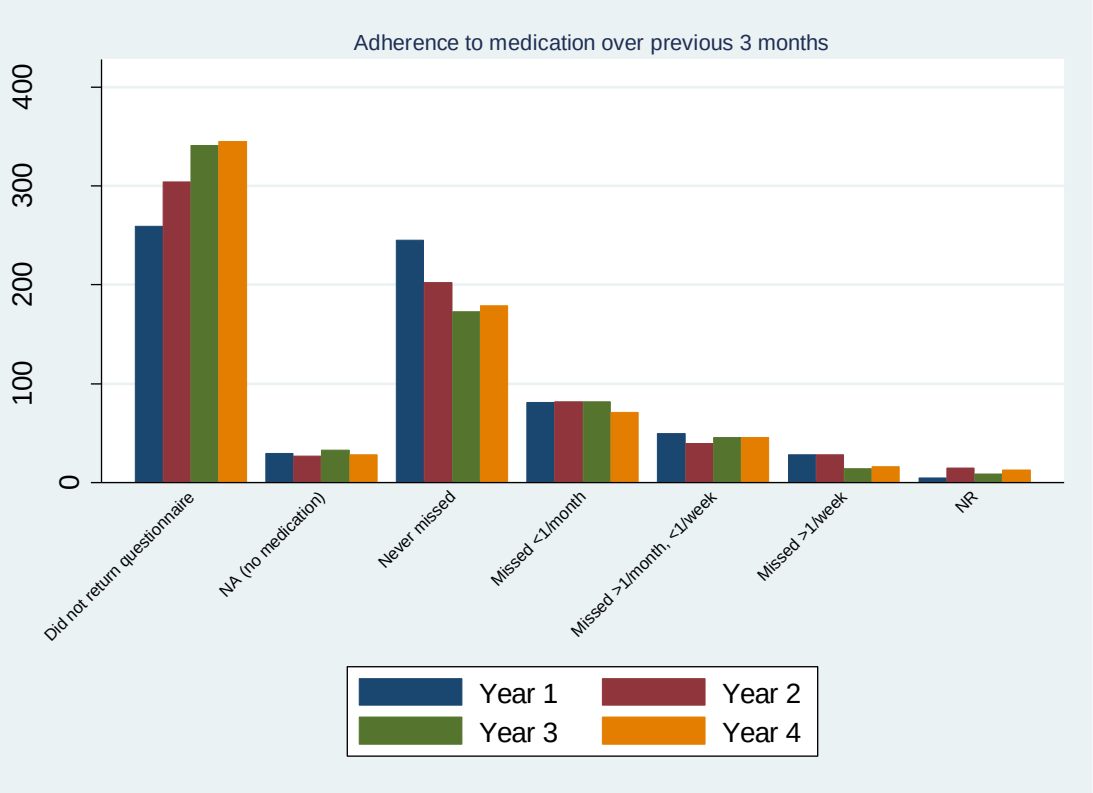


Figure 4 Adherence to medication from annual questionnaire

Thus, the data used in the analysis reflect *prescribed* drugs and doses, rather than actual treatment taken. Changes to randomised treatment prescribed by clinicians could take a variety of forms (including changes to prescribed dose, switching to other trial (or non-

trial) treatments and the addition of other trial (or non-trial) treatments) for reasons usually related to ISC or UAEs. Often patients experienced multiple treatment changes (up to three changes were recorded in the dataset).

Given the complicated treatment changes that occurred during patient follow up in SANAD, and the need to prepare data for a number of different types of analysis, as well as for comprehensive descriptions of the treatment changes that occurred, it was necessary to think carefully about how to code the data, in order to prevent time wasting with unnecessary data management prior to each analysis. However, these complications presented a challenge for data coding. First, it was necessary to consider carefully how best to code the treatment changes, to ensure a useful record of the time, type and reason for each treatment change in order to facilitate various potential subsequent analyses addressing different causal questions. Although the SANAD case report forms (CRFs) endeavoured to reliably capture information on treatment changes, inconsistencies in the interpretation of treatment information questions on the CRFs by different clinicians meant that the data on treatment changes needed to be extracted by hand from the dataset. This exercise demonstrated the need to appreciate the importance of careful consideration of data collection issues which must be undertaken during the trial planning stage, in order to ensure that the data required for any proposed statistical methods of adjustment are ultimately available and readily accessible.

It was necessary to classify changes to prescription, for each patient at each follow up visit, as either treatment switches (from randomised drug to alternative trial/non-trial

drug), additions or withdrawals. Furthermore, it was recorded when patients were still taking previous monotherapy at the time of randomisation.

Treatment changes in epilepsy rarely occur quickly; instead a withdrawal (from current treatment) and loading (increasing dose of new treatment) phase is required to prevent reactions to sudden treatment changes. Though SANAD CRFs recorded dates of (start and end of) treatment withdrawals and additions, it was necessary to decide on how to distinguish treatment *switches* (which necessarily involve an overlap period) from treatment *additions*. The overlap between starting a new AED and withdrawing from the randomised AED varied widely between patients (as demonstrated in Table 26, which summarises the withdrawal process and treatment overlap periods among SANAD A patients who switched between CBZ and LTG). Reasons for this depended on the patient's clinical condition; for example, the prescribing clinician may have wanted the patient to get established on their new drug before starting to withdraw from randomised drug.

Table 26 Withdrawal process details

	CBZ (n=20/21)	LTG (n=6/8)
Duration between starting/completing randomised AED	19.5 (0, 154)	15 (0, 21)
Duration between starting withdrawal from randomised AED and starting new AED	0 (-28, 42) [1]	0 (0, 7)
Overlap between starting new AED and finishing randomised AED	19.5 (-42, 155) [2]	13.5 (-1, 21)[3]

All values are median (range) number of days.

[1] Negative value means that patient started new drug before starting withdrawal from randomised drug: 3 patients with -28, -14 and -1

[2] Negative value means that patient finished taking randomised drug before starting new drug: 3 patients with -42, -7 and -2

[3] Negative value means that patient finished taking randomised drug before starting new drug: 1 patient with -1

The wide variety of treatment changes meant that it was necessary to code each treatment change according to the type (switch, addition, addition before switch (when a new treatment was added prior to the initiation of withdrawal of the randomised drug), withdrawal of randomised drug and subsequent reintroduction of randomised drug), timing (since randomisation), order (in the sequence of treatment changes experienced by that patient) and drug(s) involved in each treatment change.

The decision was made that, in the case of treatment switches or additions, the date of switch (or addition) was taken to be that on which the new treatment was added, regardless of how long the withdrawal period (in the case of treatment switches) or loading phase lasted.

The data available from GP records were not always complete, and some patients had extensive gaps between recorded visits. Prescribed dose data were missing on a number of patients, and it was unlikely that these data were missing completely at random: indeed, these patients were more likely to withdraw due to UAE and typically had shorter time to withdrawal than those with regular dose data.

Furthermore, following scrutiny of the data set and consultation with neurology experts, it became obvious that GPs often followed irregular prescribing patterns reflecting the uncertainty of the efficacy and safety of AEDs; for example, a new drug may have been recorded once only without an accompanying dose, demonstrating an underlying doubt on the part of the patient or clinician regarding whether the drug was suitable for the patient; given the lack of precise adherence information, it was not possible to determine in these cases whether the patient even started taking these additional drugs.

Thus, before carrying out the RPSFTM, it was necessary to decide how these vagaries would be handled. First, as in the case of White's analysis (89), it was decided that a lack of follow up or treatment data indicated no change in seizure or prescription. Given the evidence suggesting informative missingness, it would have been interesting to carry out a sensitivity analysis to investigate such effects. However, for this analysis, those with sparse follow up or treatment data (for example, 7 patients had their first treatment information recorded more than 365 days since randomisation, ranging from 428 to 919 days since randomisation) were assumed not to have experienced unrecorded treatment change or seizures between visits.

After consulting expert opinion, it was decided that a treatment switch (as opposed to addition) required less than one month (30 days) between the dates of the start of the withdrawal phase and the first prescription of alternative treatment. When short term drugs were prescribed (for example, in preparation for surgery, as surgery may induce seizures) it was decided not, for the pragmatic purposes of the trial, to consider these as treatment switches or additions.

8.3.2. Complications of the SANAD trial

The SANAD trial presented numerous other complications beyond that of the treatment data. First, both arm A and arm B involved multiple randomisation, with patients randomised to one of five treatment groups (in arm A) and three treatment groups (in arm B).

The solution to the problem of these multiple randomised groups was provided by considering the causal research question of greatest clinical interest, namely the

comparison between CBZ versus LTG. For patients with focal epilepsy (recruited in the main to arm A of SANAD), CBZ was the standard treatment most familiar to clinicians at the time of the trial, while LTG was the new drug of greatest interest to clinicians. Thus the CBZ arm of the trial presented a sensible baseline comparator for a two-way comparison with LTG for patients in arm A of SANAD. This decision was supported not only by clinical interest but also because the most common switches were between the CBZ and LTG arms; for example, more than half (31 (and 20) respectively) of the 60 (36) patients who changed treatment prior to achieving 12 month remission in the CBZ (LTG) treatment groups had switched to the alternative LTG (CBZ) treatment. As such, this treatment switching scenario appeared to lend itself quite well to the RPSTFM in practical as well as clinical respects.

Another complication in SANAD relates to the primary efficacy outcome, time to 12 month remission (T12mR). As will be discussed in section 8.4, although very important from a practical perspective, analysis of T12mR is a complicated non-standard summary of repeat events data which introduces problems when applying RPSFTM for two main reasons. First, analysis of time to 12 month remission using the RPSTFM is complicated by the fact that the event cannot happen prior to 12 months. The assumption of a constant acceleration factor is thus violated, given the necessary truncation of its effect prior to 12 months (without which the model will incorrectly predict remission times of less than 12 months). Secondly, a significant proportion of patients (in the region of 30%) are expected to achieve immediate 12 month remission (I12mR), such that there will be a peak of events at this time point. The RPSTFM is not able to recognise either

of these characteristics of the data (peak at 12 months and no events prior to this time) and is likely to incorrectly predict event times prior to 12 months.

Thus it was decided, as an initial investigation, the RPSFTM would be used to adjust the simpler secondary outcome, time to first seizure (TFS), for treatment switches occurring between those randomised to CBZ and LTG in arm A of SANAD, and would be compared against ITT and PP analyses.

Similarly to the RPSFTM analysis of the honey trial, the difficulty presented by the numerous types of treatment changes (withdrawal from randomised treatment, still taking other treatment at randomisation, switching to other (trial or non-trial) treatments or addition of other treatments) was addressed, although somewhat inadequately, by simply censoring those with any treatment changes other than switches to the alternative treatment arm (CBZ/LTG) at the point of the 'illegal' treatment change (including those who were still taking a drug at randomisation).

Finally, it was decided (as per the original analysis (99)) that those patients who were later found not to have epilepsy would be excluded from the analysis, the justification being that diagnostic methods have improved markedly since the SANAD began recruiting, and that receipt of AEDs by non-epileptics can be especially harmful.

Thus, all analysis sets excluded those patients found not to be epileptic, the PP analysis set censored patients at the point of any treatment change that occurred prior to first seizure, and the RPSFTM analyses censored at the point of any treatment changes that occurred prior to first seizure except switches between CBZ and LTG. (These treatment changes are referred to here as "illegalities", and include switches to other

trial/non-trial treatments, starting additional treatment (including starting LTG (or CBZ) without withdrawing from randomised CBZ (or LTG) treatment) or still taking an additional drug at randomisation). In order for a patient to be eligible for the analysis of the outcome TFS, any illegality must have occurred *after* the time of first seizure (or the end of follow up for first seizure, whichever occurred first). In order for a switch to be eligible for the analysis of the outcome TFS, the switch must have occurred *prior* to first seizure (or the end of follow up for first seizure, whichever occurred first). In other words, in order for a patient to be included in the RPSFTM analysis as a *treatment switch*, the following had to be true: time of legal switch to alternative (CBZ/LTG) < time of first seizure < time of any illegality. "Switch" therefore refers to switch from randomised treatment CBZ to LTG or from randomised treatment LTG to CBZ.

8.3.3. Model

As set out in section 8.1.1, it is necessary to define the "on" and "off" treatment times (T_{0i} and T_{1i}) relating to the particular research question of interest, as well as the maximum censoring time. Thus for this comparison, the model

$$U_{0i}(\beta) = T_{0i} + e^{\beta} T_{1i}$$

relates the underlying remission time U_{0i} (if patient had remained on CBZ throughout) to T_{0i} (the time spent on CBZ) and T_{1i} (the time spent on LTG). The maximum follow up time (necessary for recensoring) was calculated for each patient as the difference between their date of randomisation and known final date of follow up.

8.3.4. Results

Table 27 sets out how data and treatment changes were handled in each (ITT, PP and RPSFTM) analysis, with a summary of the number and timing of eligible and ineligible treatment switches for the RPSFTM analysis given in Table 28. The "eligible" switches between CBZ and LTG arms made up a small proportion of total treatment changes; of those treatment changes occurring prior to TFS, only 19 (43% of 44 treatment changes among patients randomised to CBZ) and 9 (26% of 35 treatment changes among LTG patients) were direct switches between treatment arms.

Table 27 Definition of ITT, PP and RPSFTM analysis sets

	Number of patients		ITT	PP	RPSFTM
	CBZ	**LTG**			
Originally randomised	**378**	**378**			
No dose data	24	20	Excluded	Excluded	Excluded
Not epileptic[1]	10	8	Excluded	Excluded	Excluded
Potential for inclusion in analysis	**346**	**352**			
Treatment changes[2]					
Switch to alternative CBZ/LTG[3]	19	9	Ignored	Censored[5]	Switch[6]
Switch to other AED	14	6	Censored[5]	Censored[5]	Censored[5]
Start additional AED	0	3	Censored[5]	Censored[5]	Censored[5]
Still on AED at randomisation[4]	11	17	Censored[5]	Censored[5]	Censored[5]

[1] Note that dose data are missing for two patients from each group who were found not to be epileptic
[2] Only relates to changes occurring prior to first seizure; those occurring after first seizure are ignored
[3] Switch from randomised CBZ to LTG, or from randomised LTG to CBZ, prior to any other treatment changes
[4] Those still on an AED at randomisation were censored at time 0.1 days.
[5] Those with any changes in treatment that occurred prior to first seizure (switches, additions or those already taking an AED at time of randomisation) were censored at the time of the first occurring treatment change
[6] Switches from randomised CBZ to LTG or vice versa were accounted for in this analysis.

Table 28 Summary of the number and timing of eligible and ineligible switches

	CBZ arm (n=346)			LTG arm (n=352)		
	Number (%)	Time from randomisation to switch		Number (%)	Time from randomisation to switch	
		Median (IQR)	Range		Median (IQR)	Range
Eligible switch[1]	19 (5.5)	84 (25, 112)	(1, 213)	8 (2.3)	56 (21.5, 110.5)	(14, 794)
Ineligible switch[2]	40 (11.6)	180 (49.5, 370)	(3, 1498)	48 (13.6)	283.5 (87, 479)	(4, 2043)

[1] Switch occurred prior to first seizure and any illegality occurred after first seizure.
[2] Switch occurred after first seizure, or switch and illegality occurred prior to first seizure.

Table 28 demonstrates that similar proportion of each treatment group experienced a switch to the alternative group (CBZ/LTG) (CBZ arm: 59 (17%); LTG arm: 56 (16%)) but many more of these switches in the CBZ group were eligible (19/59, 34%) compared to the LTG group (8/56, 14%). This was primarily because, in the LTG group, switches tended to occur due to ISC (and therefore *after* the first seizure) because clinicians were less familiar with this novel drug and therefore less adept at prescribing the correct dose to prevent seizures. In contrast, clinicians were more familiar with CBZ and therefore more likely to prescribe higher initial doses to ensure good seizure control, potentially at the expense of tolerability, thus increasing the likelihood of switches due to UAEs *before* the first seizure.

The results (for the relative effect of LTG vs CBZ) are given in Table 29. Comparison of the treatment effects obtained using these different analysis sets demonstrates that PP, ITT and RPSFTM analyses increasingly suggest that TFS is earlier with LTG than with CBZ (acceleration factors 1.87, 1.92 and 2.00 and hazard ratios of 1.20, 1.21 and 1.22 respectively), as was concluded in the original HTA report. Thus adjustment for

direct switches between treatments using the RPSFTM implies a slightly greater advantage of LTG over CBZ than that demonstrated by ITT or PP analyses.

Table 29 ITT, PP and RPSFTM analyses for SANAD A

		e^{β} (95% CI)	HR (95% CI) LTG:CBZ	p-value
ITT		1.92 (1.03, 3.41)	1.21 (1.01, 1.44)	0.034
RPSFTM				
	No recensoring	2.00 (1.03, 3.94)	1.23 (1.02, 1.48)	0.034
	Recensoring[1]	2.00 (1.02, 3.70)	1.22 (1.01, 1.46)	0.037
PP		1.87 (1.001, 3.33)	1.20 (1.005, 1.44)	0.042

[1] Number of patients (events) recensored: CBZ: 27 (2), LTG: 82 (16)

The considerable difference between AF and HR suggest that, if approximating the data using a Weibull distribution, the shape parameter would not be equal to 1 (in other words, the TFS data do not follow an exponential distribution). This would occur, for example, if the hazard rate was initially steep and then flattened out, perhaps due to an initial peak in the number of events due to early seizures in a subset of patients for whom treatment was inefficacious, followed by a more stable period of seizures among remaining trial participants.

8.3.5. Limitations of time to first seizure CBZ/LTG RPSTFM

There are a number of limitations of this analysis. Firstly, this analysis compares only two of the five treatment arms. Although more treatment contrasts could in theory be included in the RPSTFM, estimation is likely to be unreliable due to model instability (94).

Because only those treatment changes occurring up to the time of first seizure are considered in this analysis, analysis of TFS by definition considers only those changes due to UAE and thus does not relate to problems related to seizure control. Analysis of TFS is further disadvantaged because it is likely to be influenced by initial dose: standard AEDs are more familiar to the prescribing clinicians, and hence are likely to be introduced at higher doses than those drugs which are newer and less familiar. Given that seizure control is directly correlated with dose, this familiarity and subsequent higher dosing with the standard drugs is expected to bias the treatment comparison of TFS in their favour.

In addition, as the RPSTFM can only reliably adjust for one type of treatment change, if patients experience changes other than pure treatment switches between the comparator drugs (for example, those who add on another AED or those who withdraw from their randomised drug without switching to the alternative AED of interest), their follow up is necessarily artificially censored at this point. This censoring is likely to introduce bias, as those who undergo treatment changes may well have a particularly poor prognosis (for example, in the case of adding further AEDs) or good prognosis (if, say, they have decided to stop taking randomised treatment because they have not experienced seizures for some time) at the point of their treatment change.

Thus, the most notable limitation of this analysis is that the majority of treatment changes are necessarily discounted in the RPSFTM, because of the need to censor any patients experiencing (prior to their first seizure) any treatment changes other than direct switches between CBZ and LTG. Of the potential 59 switches that occurred from randomised drug CBZ to LTG, and 56 that occurred from randomised drug LTG

to CBZ, only 19 and 8 respectively were included in the RPSFTM analysis, either because of some ineligibility that occurred prior to the first seizure (for example, when the patient started an additional drug or switched to a different drug) or because the switch (between CBZ and LTG) occurred after the first seizure. Indeed, it is of interest to note that these direct switches (after the first seizure) occurred more often in the randomised LTG group, given that patients tended to switch from LTG to CBZ due to ISC (i.e. after first seizure) whereas patients were more likely to switch from CBZ to LTG because of UAEs (i.e. potentially before first seizure). The omission of these treatment changes from analysis not only reduces the power of the RPSFTM analysis, but is also likely to introduce selection bias, given that censoring at the time of such treatment changes is likely to be informative (related to the patients' prognosis).

Thus, these RPSFTM results do not relate to patients who experienced treatment changes other than switches between CBZ and LTG, and the choice of outcome (TFS) means that only switches primarily due to UAE (rather than ISC) were accounted for. It is therefore of interest to therefore consider a more relevant outcome reflecting long term remission of seizures, as demonstrated in the next analysis.

8.4.SANAD B

The third version of the RPSFTM relates to adjustment for treatment changes when analysing the primary outcome of the SANAD trial, time to 12 month remission (T12mR). T12mR is of greater interest to patients and clinicians alike than TFS, not only because of its significance for regaining the right to a driver's licence but also because it is less dependent on initial dose (as it requires a *sustained* effect of treatment

on preventing seizures). However, as will become evident, it is a complicated summary of repeated events data that introduces complications for analysis.

Given that the primary explanatory effect of interest for SANAD investigators was the causal effect of treatment in patients with generalised, rather than focal, epilepsy, this final evaluation of causal effects using the RPSFTM is applied among those randomised to arm B, rather than arm A, of the SANAD trial. At the time of recruitment, clinicians recorded whether patients had generalised or focal epilepsy (or if unclear at the time, the patient was recorded as having "unclassified" epilepsy). In the case of focal epilepsy, the drug of choice is CBZ, and such patients should have been recruited into arm A of SANAD (where the standard comparator treatment was CBZ). VPS is the drug of choice for generalised epilepsy, and such patients were intended to be recruited into arm B (where the standard treatment was VPS). However this did not always occur: some clinicians entered patients into arm B despite there being focal elements to their epilepsy, where VPS is definitely not a recommended treatment. Given that this inclusion of patients with focal type epilepsy is likely to blur the true difference between the treatments, it was decided that these patients should be excluded from the explanatory analysis. Although it may have been more statistically valid to avoid excluding patients and instead include an interaction term for the type of epilepsy in the causal model, this would have added to the complexity of the model; furthermore, this exclusion affected only 39 (9%) of those randomised to arm B.

The choice of treatment comparison was once again guided by clinical factors. The standard treatment for patients with generalised epilepsy at the time of SANAD was VPS, and, as for arm A, the newest most promising alternative drug was LTG; as such,

treatment switches most commonly occurred between these treatment arms, and it was considered of greatest clinical interest to compare efficacy of LTG versus VPS.

Thus in this analysis, only those arm B patients with generalised (or unclassified) epilepsy were evaluated to determine the causal effect of LTG versus VPS on the primary outcome T12mR. Although treatment switches occurred relatively commonly between these two treatment arms, other treatment changes also occurred. Thus it was necessary (as for the previous two RPSFTM applications) to consider how to handle these other treatment changes in the analysis, particularly in relation to the causal contrast from the model as defined by the research question. A reminder of the RPSFTM

$$U_{0i}(\beta) = T_{0i} + e^{\beta} T_{1i}$$

highlights the need to determine the "on" and "off" treatment states, which relate directly to this research question (and thus, in turn, to the handling of changes other than direct switches between treatments).

8.4.1. Clinical question of interest

In the arm A analysis presented earlier in this chapter, the inputted times (T_{0i} and T_{1i}) related only to time spent on CBZ or LTG monotherapy (respectively); thus censoring was introduced at the point of any changes to treatment prescription excluding direct switches between these treatments, in an attempt to estimate a purely explanatory relative treatment effect (by adjusting for "true" switches between randomised treatments).

However, in reality, the main research question of interest for SANAD clinicians is not purely explanatory, but rather exists on the explanatory/pragmatic continuum: although interest lies in the biological effects of certain treatments, it is acknowledged that AED treatment is given in a pragmatic setting, where patients may experiment with doses or new drugs before finally settling on a treatment of choice, or where patients may begin randomised treatment while still taking a previous monotherapy, or where switches between treatments may involve an extensive overlapping period to allow for necessary loading or withdrawal phases.

Thus, a more realistic analysis would adjust for situations where patients end up on the alternative treatment, perhaps first having tried other treatments or combinations, rather than automatically censoring such patients at the point at which they try other treatments. This pragmatic take on the causal analysis allows for the fact that patients may have been offered treatments and may have tried them for a short period before deciding whether to take them as long term treatment, and assumes that these short dalliances with other treatments do not overly influence the effect on sustained freedom from seizures.

Thus, one may make a comparison between two pragmatic treatment regimes, namely "VPS plus any treatment changes that occur (except those involving LTG)" versus "LTG plus any treatment changes that occur (except those involving VPS)". In such an analysis, no artificial censoring would be necessary, as any treatment changes not involving the alternative treatment of interest are considered to be part of the pragmatic treatment regimens being compared.

In the case of the SANAD trial, the question of interest for clinicians may well lie somewhere in between these "true" and "pragmatic" treatment switch scenarios. Discussion with neurology consultants suggests that interest lies primarily in adjusting for switches which result in the patient "ending up on" the alternative treatment; thus they are willing to accept the pragmatic situation that patients may well try other treatments or combinations of treatment prior to ending up on the alternative treatment of interest. However, if a patient then ceases to take this alternative monotherapy (for example, if they subsequently add on alternative treatments or withdraw from all treatment), it is more clinically appropriate to censor the patient's follow up at this point of treatment change. Thus all such treatment changes are ignored *prior* to the occurrence of a direct switch (between the randomised treatments of interest), but *following* a direct switch, any such treatment changes are considered to blur the comparison and thus would need to be censored.

8.4.2. Causal contrasts

The chosen definition of switch (either true, pragmatic or clinically relevant) translates directly to the causal research question and causal contrast, defined by T_{0i} and T_{1i}, in the RPSFTM.

First consider the pragmatic question "what are the relative effects of LTG (with any treatment changes that ensue with this treatment policy, excluding adding or switching to VPS) and VPS (with any treatment changes that ensue with this treatment policy, excluding adding or switching to LTG)?"

For such a pragmatic analysis, T_{0i} would be calculated as the total time spent on VPS (including time following any other additions or switches to subsequent treatments, but excluding addition with, or switching to, LTG); similarly T_{1i} would be calculated as the total time spent on LTG (including any other additions or switches to subsequent treatments, but excluding addition with, or switching to, VPS). Thus, in evaluating the effect of "ending up on" the alternative treatment, this analysis considers any other changes to treatment as a pragmatic continuation of randomised (or switched) treatment, thus removing the need to censor follow up at the point of these alternative treatment changes.

Note, however, that this pragmatic assessment of treatment switches is somewhat inconsistent, as the addition of VPS to (randomised) LTG would be recorded as T_{0i} and thus is considered the opposite (in terms of treatment contrasts) to when LTG is added to (randomised) VPS (recorded as T_{1i}), despite these scenarios being identical in terms of treatment received.

In practical terms, this approach meant that any changes to randomised drug (VPS or LTG) such as withdrawal, addition and switches of AEDs were ignored, except those which involve the alternative drug of interest (LTG or VPS): any addition or switch to alternative (LTG or VPS) was considered to be a "switch". Subsequent changes following such a "switch" were also ignored, as they too were considered to be part of the pragmatic treatment on the alternative treatment (as shown in Figure 5, where different coloured lines indicate different treatments).

In contrast, the explanatory approach to this analysis adjusted only for "true" switches (direct switches between LTG and VPS) which occurred prior to any other treatment changes (as demonstrated by Figure 6 below), thus answering the research question "what are the relative treatment effects of LTG versus VPS?" Thus in this scenario (as for arm A comparison), follow up is censored at the point of any treatment changes (including additions of the alternative drug (LTG or VPS)) other than direct switch to alternative (LTG or VPS).

Finally, of greatest *clinical* interest would be adjustment for those pragmatic switches in which the patient ends up on alternative treatment but any *subsequent* deviation from that treatment would then be censored, thus assessing the research question of interest "what is the effect of LTG versus VPS, when any treatment changes *prior* to the treatment switch are considered to be part of the treatment experience on each randomised drug?" (shown in Figure 7).

This analysis, although clinically sensible, is neither consistent nor statistically valid (even beyond that of the "pragmatic" analysis mentioned above), as a certain treatment change would be handled differently depending on whether it occurred before or after the treatment switch of interest. Thus this example demonstrates how careful discussion between statistician and clinician is necessary to reach a decision on how to analyse the data in a manner which is both statistically and clinically appropriate.

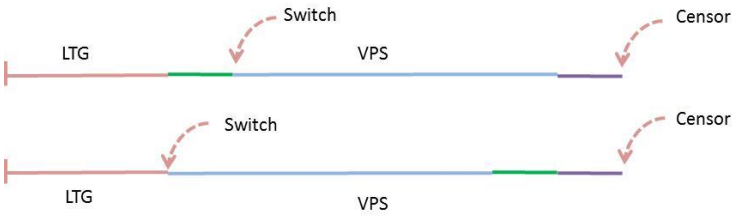


Figure 5 Pragmatic scenario

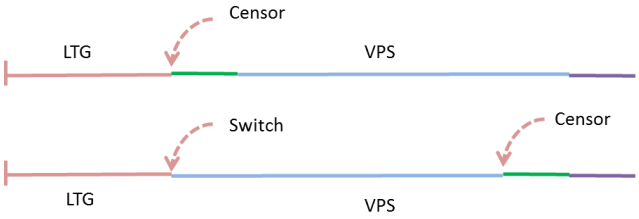


Figure 6 Explanatory scenario

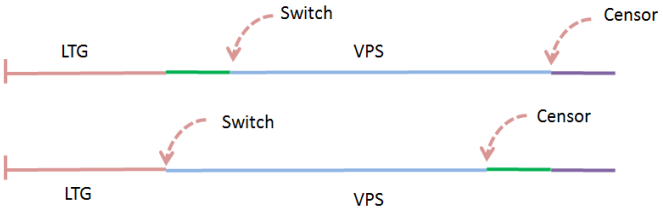


Figure 7 Clinically relevant scenario

8.4.3. Methods

Thus, given that the intermediary (though most clinically relevant) scenario was not statistically valid, the two remaining versions of the analysis of time to 12 month remission were carried out, allowing for switching between LTG and VPS; the first adjusted for "true" switches, thus censoring at the time of any other treatment changes, while the second "pragmatic" switches analysis considered other treatment changes to be part of the pragmatic treatment regimens being compared and thus no artificial censoring was introduced. Each patient's maximum censoring was calculated as the

difference between their date of randomisation and the known universal final date of trial follow up.

Individuals' treatment times T_{0i} and T_{1i} were calculated according to the adjustment scenario: when considering the effect of "true" switches, T_{0i} (T_{1i}) were calculated as the times patients spent receiving VPS (LTG) prior to censoring due to any "illegalities". For the "pragmatic" switches adjustment, T_{0i} (T_{1i}) were calculated as the times patients spent receiving VPS (LTG) including any time spent receiving additional treatments to, or switches from, these drugs (except switches to or additions of the alternative LTG (VPS) drug).

8.4.4. Results

The frequency of these switches in the VPS and LTG randomised arms is described in Table 30, with further detail given in Table 31 and Table 32 regarding the relationship between these pragmatic and true switches, and differences between the (censored) remission times between the true and pragmatic switches in each treatment arm (due to the differential censoring patterns implicit in the definitions of these two types of switch). It can be seen that only 60% of the 59 (and 53% of the 34) "pragmatic" switches that occurred in the LTG (VPS) arm were eligible to be analysed as "true" switches; furthermore 2 (5) of these were censored earlier for the "true" switch analysis than for the "pragmatic" switch analysis, because of another treatment change that occurred following the "true" switch. Thus these tables describe the effect of artificial censoring (due to disallowed events referred to as "illegalities") necessary for the "true" switch comparison.

Table 30 True and pragmatic switches between VPS and LTG

Values indicate number (% of treatment group).

	LTG (n=218)	VPS (n=217)	Total (n=435)
Pragmatic switch	59 (27.1)	34 (15.7)	93 (21.4)
True switch	36 (16.5)	18 (8.3)	54 (12.4)

Table 33 (Table 34) provides detail on the timing and reason for true (pragmatic) treatment changes, broken down according to whether or not patients achieved remission. These tables demonstrate that true switches were twice as common (and pragmatic switches were nearly twice as common) in LTG compared with VPS patients; true (and pragmatic) switches tended to occur earlier in the VPS arm, and these true switches were most likely to be triggered by ISC in the LTG arm.

Table 31 Summary of switches for LTG arm

LTG arm (n=218)	Total	Remission time[1]					
		Remission time (true) = remission time (pragmatic)[2]	Remission/follow up time (true) < remission time (pragmatic) i.e. censored early because of other treatment change (not switch)				
			Total		Remission time (true) < (pragmatic) switch time[3] i.e. treatment change occurred prior to pragmatic switch	Remission time (true) = (pragmatic) switch time[4] i.e. censoring at pragmatic switch time (as pragmatic "switch" involved adding alternative AED)	Remission time (true) > (true and pragmatic) switch time[5] i.e. treatment change occurred after pragmatic switch
No pragmatic switch	159	132	27		0	0	0
			Withdrew	18			
			Added AED	3			
			Switched AED	6			
Pragmatic switch	59	34	25		9	14	2
No true switch	23	0	23		9	14	0
					Withdrew	4			
					Added AED	5	Added AED	14 (14 VPS)	
True switch	36	34	2		0	0	2
							Switched AED	2

[1] Note that "remission time" refers to length of follow up for remission (i.e. minimum of time to remission or censoring)
[2] No censoring event occurs before remission time
[3] Censoring event occurs prior to pragmatic switch
[4] Pragmatic switch is censoring event (i.e. involves addition of, rather than switch to, VPS)
[5] Censoring event occurs after (true and pragmatic) switch

Table 32 Summary of switches for VPS arm

VPS arm (n=217)	Total	Remission time[1]				
		Remission time (true) = remission time (pragmatic)[2]	Remission time (true) < remission time (pragmatic) i.e. censored early because of treatment change after switch			
			Total	Remission time (true) < (pragmatic) switch time[3]	Remission time (true) = (pragmatic) switch time[4]	Remission time (true) > (true and pragmatic) switch time[5]
No pragmatic switch	183	154	29 Withdrew 18 Added AED 6 (VPS, CNZ, ETH, LVT, TPM) 2 Switched AED 5 (3 VPS, VPS/TPM, ETH)	0	0	0
Pragmatic switch	34	13	21	8	8	5
No true switch	16	0	16	8 Withdrew 7 Switched VPS 1 (LVT)	8 Added VPS 8 (8 VPS)	0
True switch	18	13	5	0	0	5 Switched AED 2 Added AED 2 Reintroduced randomised AED 1

Table 33 Summary of (time to) remission or censoring according to true switch status

	LTG				VPS			
	Number (% within **treatment** or treatment/switch group)	Time to event (median, range)	Reason for (first) treatment change		Number (% within switch group)	Time to event (median, range)	Reason for (first) treatment change	
True switch	**36 (16.5)**	**174.5 (1, 1245)**			**18 (8.3)**	**136 (14, 734)**		
Remission	27 (75.0)		ISC	16	10 (55.6)		ISC	2
			UAE	7			UAE	5
			ISC/UAE	4			ISC/UAE	2
							Pregnancy	1
Time to switch		182 (20, 790)				135.5 (14, 570)		
Time from switch to remission		373 (244, 1066)				326.5 (162, 1704)		
Time to remission		617 (365, 1819)				475 (365, 1934)		
Censoring	9 (25.0)		ISC	5	8 (44.4)		ISC	3
			UAE	3			UAE	1
			Other	1			ISC/UAE	2
							Pregnancy	1
							No seizures	1
Time to switch		32 (1, 1245)				158.5 (56, 734)		
Time from switch to censoring		449 (55, 1900)				294 (3, 508)		
Time from switch to censoring due to other treatment change*	(n=2)	55, 145			(n=5)	335 (40, 508)		
Time to censoring		1387 (76, 1932)				442 (168, 1063)		
Time to censoring due to other treatment change*	(n=2)	76, 150			(n=5)	429 (168, 610)		

	LTG				VPS			
	Number (% within **treatment** or treatment/switch group)	Time to event (median, range)	Reason for (first) treatment change		Number (% within switch group)	Time to event (median, range)	Reason for (first) treatment change	
No switch	**182 (83.5)**				**199 (91.7)**			
Remission	101 (55.5)	379 (365, 2192)			132 (66.3)	365 (365, 1718)		
Censoring	81 (44.5)				67 (33.7)			
Time to censoring		344 (5, 1735)				225 (1, 1840)		
Time to censoring due to other treatment change*	(n=50)	188 (5, 1280)	ISC	27	(n=45)	147 (1, 1036)	ISC	15
			UAE	13			UAE	17
			ISC/UAE	3			ISC/UAE	5
			Pregnancy	1			Pregnancy	6
			No seizures	4			Patient decision	1
			Non-compliance	1			Non-compliance	1
			Unknown	1				

* i.e. includes only patients whose censoring time was brought forward because of other treatment change.

Table 34 Summary of (time to) remission or censoring according to pragmatic switch status

	LTG			VPS		
	Number (% within **treatment** or treatment/switch group)	Time to event (median, range)	Reason for (first) treatment change	Number (% within switch group)	Time to event (median, range)	Reason for (first) treatment change
Pragmatic switch	**60 (27.1)**	**229 (1, 1523)**		**34 (15.7)**	**164.5 (14, 1036)**	
Remission	40 (67.8)			22 (64.7)		
Time to switch		218 (5, 958)			142 (14, 570)	
Time from switch to remission		418.5 (169, 1070)			394 (162, 1704)	
Time from randomisation to remission		686.5 (365, 1819)			724 (365, 1934)	
Censoring	19 (32.2)			12 (35.3)		
Time to switch		252 (1, 1523)			225 (39, 1036)	
Time from switch to censoring		960 (151, 1900)			545.5 (3, 1700)	
Time from randomisation to censoring		1539 (678, 2010)			1087.5 (328, 1802)	
No switch	**159 (72.9)**			**183 (84.3)**		
Remission	117 (73.6)	380 (365, 2192)		147 (80.3)	365 (365, 1753)	
Censoring	42 (26.4)	739 (71, 1899)		36 (19.7)	644 (13, 1932)	

* i.e. includes only patients whose censoring time was brought forward because of other treatment change.

Table 35 demonstrates the degree to which artificial censoring in PP analysis and the "true" switches RPSFTM curtail analysis times, as the median (and IQR) of survival times drop considerably in the PP analysis (due to censoring at the time of each patient's initial treatment change) and to a lesser degree, as expected, in the "true" switches RPSFTM (given that all treatment changes except the "true" switches are censored).

Table 35 Summary of T12mR

	Treatment			
	LTG (n=218)		VPS (n=217)	
	No. switches	Median (IQR)	No. switches	Median (IQR)
PP	-	488 (365,960)	-	370 (365,677)
ITT	-	570 (365, 1162)	-	438 (365, 906)
RPSFTM (true switches)	36	540 (365, 912)	18	371 (365, 733)
RPSFTM (pragmatic switches)	59	570 (365, 1162)	34	438 (365, 906)

Table 36 ITT, PP and RPSFTM analyses for SANAD B VPS vs LTG

	e^{β} (95% CI)	HR (95% CI) (LTG:VPS)	p-value
ITT	0.88 (0.74, 0.9993)	0.76 (0.63, 0.93)	0.008
RPSFTM			
True switches	0.85 (0.69, 0.9994)	0.70 (0.54, 0.91)	0.007*
Pragmatic switches	0.81 (0.58, 0.9993)	0.63 (0.44, 0.89)	0.008
PP	0.90 (0.78, 0.9995)	0.74 (0.59, 0.93)	0.009

* P-value differs from ITT due to censoring at the time of other treatment changes in the "true" switch analysis.

It is interesting to note (in Table 36), first that the AF (e^{β}) and HR differ within the same analyses, indicating again that the shape parameter is unlikely to be close to 1; this is because the hazard is initially steep (with approximately 30% of patients achieving I12mR) and then flattens off. Also, it is noted that the AF for the PP analysis is closer

to 1 than the AF for the ITT analysis, whereas the HR for the ITT analysis is closer to 1 than that for the PP analysis. This apparent anomaly would arise if the baseline hazard functions have different shapes in the PP and ITT analyses, thus providing an explanation of why the relationship between HR and AF differ between analyses.

Adjustment for pragmatic (and true) switches in this trial causes the estimated AF (and HR) to fall further away from one compared to ITT and PP analyses. The AF is interpreted in terms of rate of "using up" survival time; hence these results suggest that T12mR is "used up" increasingly slowly (with LTG compared to VPS) from PP to ITT to true switch adjustment and finally to pragmatic switch adjustment. In other words, as more switches are accounted for, the benefit of VPS becomes increasingly apparent.

Thus, in summary, although it is important to consider clinical opinion when deciding on how to approach causal estimation, the clinically relevant causal scenario in this trial did not provide a valid basis on which to perform statistical assessment, thus requiring a compromise in approach. Although the "pragmatic" switch scenario requires less artificial censoring (and hence is less biased than the PP or "true" switch equivalent), it addresses a less clear cut clinical question, distinguishing only between randomised (VPS and LTG) treatments, rather than accounting for any the impact of any other trial (or non-trial) treatments. The "true" switch scenario is more precise but results in a large amount of censoring (due to other treatment changes occurring before legitimate treatment switches or end of event follow up).

8.5.Discussion

8.5.1. Honey trial

When interpreting results from the RPSFTM, it is necessary to consider the suitability of the assumptions underlying use of the acceleration factor in the RPSFTM; first, it is assumed that the effect of treatment is immediate and there is no carryover effect from previous treatment received. This is likely to be true in the honey trial; although it may be possible that treatment effects extend for a short while beyond the time of treatment, and that treatments take some time before becoming clinically effective, the extent of delayed or extended treatment effects is not likely to be substantial in this context. The model also assumes that the causal effect is attributable directly to the actual treatment received, rather than to any effect caused by switching treatments. However, in this clinical area, it is not likely that the act of switching treatments will cause a biological reaction beyond the actual effect of treatment. The constant treatment effect implied by a single acceleration factor may however be contravened in this setting; although switches do not occur solely at a fixed clinically defined point (such as disease progression), they are most likely to occur when prognosis with randomised treatment appears poor. As such, the effect of treatment received at randomisation may differ from that of treatment received at the point of switch (when wounds are likely to have deteriorated).

Thus, in theory, the ability to assess the impact of time-dependent treatment changes on the time to the event of interest without needing to make any assumptions about the relationship between a patient's baseline prognosis and their actual treatment would suggest that the RPSFTM would be especially useful in the honey trial, as the treatment switches in this trial were often informative and related to prognosis. Patients who

discontinued treatment because they were disappointed not to have received honey were unlikely to have been experiencing healing success with their conventional treatment. Conversely patients randomised to conventional treatment who were switched to receive honey by external clinicians were often younger, fitter patients who were expected to do well with honey, and thus these switches were related to prognosis also.

However, the restriction of the RPSFTM to a single causal contrast (comparing conventional and honey dressings only) meant that, in the same way as for the other analysis sets, all treatment changes other than direct switches between randomised treatment arms needed to be censored in the RPSFTM analysis, even though such censoring is likely to be informative and may therefore introduce selection bias. For example, censoring patient follow up when wound deterioration necessitated more extreme treatment options (such as surgery, radiotherapy or antibiotics) introduces bias, as these wounds are not typical of all the wounds in the trial at the point of censoring.

The need for such censoring would be removed if the time spent on alternative treatments (to conventional or honey dressings) could be incorporated into the causal contrasts (T_{0i} and T_{1i}), but this would then change the interpretation of the causal research question being addressed by the RPSFTM. For example, if the time T_{0i} instead included all the time spent on *all* alternative treatments (including the more extreme options such as surgery), this (at least in theory) would remove the need to censor follow up at these alternative treatment changes; however the interpretation of the causal estimand would change, as the RPSFTM would instead compare honey to *all* other (minor and major extremes of) treatment, ranging from conventional wound dressings to surgery, which is not likely to be a clinically sensible contrast.

Furthermore, given that the trial staff were often unable for practical reasons to continue following up patients who experienced treatment changes other than simple switches to the alternative trial treatment, corresponding healing times were typically unavailable following such switches; as such, the lack of follow up data due to practical reasons meant that it would not have been possible to account for the causal effect of such deviations from treatment protocol even if this alternative causal question (or indeed a more complex RPSFTM model involving more than one causal contrast) was practicable.

The extent of this bias is demonstrated by the use of a sensitivity analysis censoring patients at the longest possible follow up time when they received extreme forms of treatment due to severity of their condition, reflecting the fact that they would otherwise have been unlikely to experience healing during the trial follow up. This sensitivity analysis had a similar impact on the treatment effect estimates as the RPSFTM adjustment, demonstrating the impact of this informative censoring is similar to that introduced by switches between randomised treatments.

8.5.2. SANAD

Similar considerations are necessary for the evaluation of the use of RPSFTM in SANAD as for the honey trial, in particular concerning the underlying assumptions regarding the effect of treatment as specified by the model. For example, the simple RPSFTM implies that the impact of treatment is immediate and constant, without any carryover effect or inherent effect attributable to (the act of) switching.

The validity of these assumptions again depends on the clinical scenario. Thus, in SANAD B, it may be more appropriate to introduce lagged coefficients to allow for the

likely reduced effect of treatment during the titration phase, or a coefficient that decreases with time from the start of the withdrawal phase (as discussed for the MRC trial by White (89). It may be necessary to allow for the overlapping phases of treatment withdrawal (from randomised drug) and loading (of alternative drug) when patients are receiving two treatments (even though they are technically experiencing a treatment switch). Also the very act of switching treatment, regardless of the actual treatments involved, could potentially be a trigger for seizures, such that the switch is having a causal effect independently of treatment. To investigate whether this was the case, it would be necessary to test the significance of an indicator variable for treatment switches in the causal model. Furthermore, it is important to consider the impact of recensoring, in particular whether recensoring is likely to influence overall estimated treatment effects (if, for example, there is a treatment-time interaction, such that recensoring of late events causes a distortion in the estimated overall treatment effect).

However the greatest concern regarding the use of the RPSFTM in the context of SANAD B relates to the complexities introduced by the T12mR outcome. In particular, as introduced briefly earlier, the RPSFTM fails to recognise two key features of T12mR.

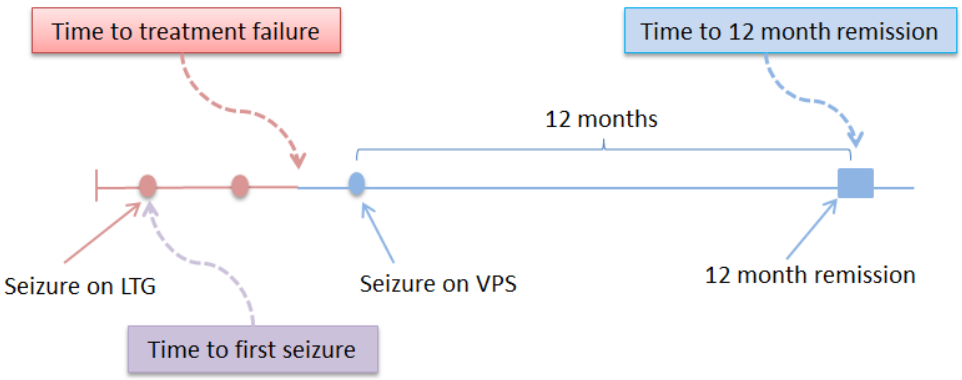


Figure 8 SANAD outcomes

First, by definition, 12 month remission cannot occur before 12 months, representing a discontinuity when modelling the effect of treatment. Secondly, a substantial proportion of patients are expected achieve I12mR, resulting in a large peak of events occurring at 12 months (for example, in SANAD B, 28.4% (41%) of LTG (VPS) patients achieved I12mR). In order to accurately represent these data, the estimated AF from the RPSFTM would therefore need to be able to predict a very high event rate at 12 months with truncation of this treatment effect immediately prior to 12 months, which is not possible given the assumption of constant treatment effect.

One solution to this would be to simply bring the event time forward by 12 months and instead analyse the time to the *start* of 12 months remission, equivalent to $t - 12$. This simple transformation of event times would mean that the event could then occur at any time from randomisation onwards. However, this may introduce problems from a philosophical point of view, as one is technically analysing an outcome that occurs in the future: at the start of a period of 12 months remission, one does not actually know that, during this period, the patient will be seizure free.

The RPSTFM would be used to adjust for the effect of treatment switches on time to start of achieving *delayed* 12 month remission (in other words, time to start of achieving 12 month remission, conditional on not having achieved it by 12 months), estimated in only those patients who did not achieve I12mR. However, this model would also be complicated by the shift along the time axis ($t - 12$), given that the first 12 months of follow up not only regarding outcome but also *treatment* would no longer be considered in the model; in other words, the clock would start at 12 months for the time-dependent treatment covariate as well as outcome.

It would not be appropriate to simply disregard the treatment received during the first 12 months following randomisation, as this period importantly reflects the short- to medium-term tolerability of AEDs, during which treatment changes are common: for example, in the LTG arm, 45/59 (76.3%) of treatment switches (to VPS) occurred within the first 365 days, of which 40 occurred in patients with remission (or censoring time) beyond 1 year, whereas in the VPS arm, 25/34 (73.5%) of treatment switches (to LTG) occurred within the first 365 days, of which 18 occurred in patients with remission/censoring time beyond 1 year.

It may at first seem possible to address this problem by including a summary measure of compliance until first seizure (or until 12 months, if the patient achieved I12mR) as a time-fixed rather than time-varying covariate (for example, by including a baseline variable "proportion of first 12 months spent on randomised treatment") and then use a time-varying binary treatment covariate from 12 months onwards. However, this would introduce bias, given that this post-randomisation summary compliance variable is very likely to be related to prognosis. Instead, a time-varying summary measure of treatment receipt (such as the proportion of the previous 12 months that the patient was on randomised treatment) could be included in the RPSFTM as a time-varying indicator of treatment, thus capturing treatment information from the time of randomization. However, this introduces practical problems, as the RPSFTM software (in Stata) currently can only utilise binary treatment factors (although there are plans to extend the program to accommodate other forms of treatment variable in the near future).

Furthermore, bringing forward the analysis time by 12 months would introduce another complication, namely that those with an I12mR (as well as those patients who were

censored with less than 12 months of follow up) would be excluded from the analysis (as their $t-12$ analysis time would be less than or equal to 0). Given that approximately 30% of newly diagnosed epilepsy patients achieve I12mR, this would result in a considerable loss of information and the model results would provide no information on this important class of patients. Although it would be possible to add on a small fraction of time (say 0.5 days) to these "0" event times to ensure that these patients are included in the analysis, this would reintroduce the problem concerning the peak in the number of events at this time, which cannot be accurately represented by the RPSFTM.

An alternative solution would be to consider a mixture modelling approach, excluding those patients with I12mR from the RPSTFM and instead including them (along with all of the other patients) in an appropriate (causal) version of logistic regression to assess the effect of treatment on achieving I12mR.

However, causal analysis of the simplified binary outcome I12mR is not straightforward, as compliance remains time-dependent despite the simplification to a binary outcome; therefore it is not appropriate, for example, to use the method of Sommer and Zeger (91), which instead assumes all-or-nothing compliance. Instead the method must in some way account for different treatments received within the first 12 months of follow up (if immediate remission was achieved) or until the time of the first seizure (if a seizure occurred within 12 months of randomisation); treatment information following the first seizure (or 12 months, whichever occurs first) is irrelevant as it has no causal impact on achieving (immediate 12 month) remission.

There is, however, no obviously appropriate method to assess binary outcome (I12mR) with a time-dependent treatment covariate that reflects treatment receipt only up to

time of time seizure. For example, the general structural mean model (GSMM) proposed by Vansteelandt (146) accommodates time-dependent compliance with binary outcomes, but this method assumes no switches from the control to experimental arm, which is invalid in this setting. Thus, although it would have been interesting to consider mixture modelling, given that the treatment switch variable is time dependent, it is likely to be overly complicated to implement such a method for these data.

8.5.3. Conclusion

The exploration of the RPSFTM in this chapter has highlighted its limitations when applied to complex trial scenarios, most notably the problems associated with multi-way randomisation and numerous forms of treatment deviation, given that the model is likely to become unstable with inclusion of more than one treatment covariate. Furthermore, the assumptions relating to the estimated AF do not match the characteristics of the remission outcome. Given that these complicating features are likely to be typical of trials in clinical areas subject to deviations from long term treatment of chronic disease, it would be useful to consider alternative methods of analysis for adjusting for such deviations when analysing survival outcomes.

In particular, for the SANAD trial, the main problems with the application of RPSFTM related to the irregular distribution of event times at 12 months (the discontinuity of events before 12 months coupled with a peak of events at 12 months) and bias introduced by censoring of treatment changes other than direct switches between randomised treatments. As such, in light of these limitations in applying RPSFTM in these two trials, the next chapter will consider the use of an alternative approach to causal estimation for this complicated trial scenario, based on observational

methodology of inverse probability weighting rather than the counterfactual framework employed by the RPSFTM.

9. Novel application of IPCW methodology

9.1.Introduction

As seen in the previous chapter, the RPSFTM suffers from a number of major limitations when applied to analysis of T12mR in the complex SANAD trial scenario. In this chapter, an alternative method suitable for survival outcomes, but based on observational techniques rather than randomisation, will be considered, namely the IPCW method.

The IPCW has the advantage over the RPSFTM in being more able to handle a variety of types of treatment changes, rather than simply dealing with a single type of treatment change. We will see how this added scope means that the IPCW is able to differentiate

between different reasons for treatment change, and more readily answer clinical questions of interest.

The IPCW overcomes problems of a standard time-dependent treatment Cox model by adjusting for TVCs externally to the outcome model through their impact on treatment probability weights. In this chapter we will consider the fundamental issues of IPCW, the assumption of NUC in the presence of numerous potential confounders, of which some are complicated TVCs changing on a daily basis for some patients.

9.2.Methods

The IPCW method recognises the fact that any change from randomised treatment will likely lead to distorted event times which do not accurately reflect the experience of patients without such treatment changes. Thus, in the context of a randomised controlled trial with a survival outcome, the IPCW can be used to adjust for changes from randomised treatment by artificially censoring patients at the point of the first of any (relevant) treatment change. The potential bias introduced by this censoring is removed if one is able to make the assumption of NUCs, in other words, all variables which predict both the occurrence of treatment change and of the outcome event of interest, in this case 12 month remission, have been accurately recorded and are available for analysis. The time to treatment change is typically modelled separately for each treatment arm, as the effect of each confounder may differ according to treatment. The probability of remaining uncensored (i.e. not experiencing a treatment change) at each time point and for each patient is estimated from this treatment-specific model, dependent on their values of the variables included in the model. The inverse of these probabilities are then used to weight the analysis of the time to outcome event, which

must also be adjusted for all the baseline covariates. The reasoning behind this methodology is that, if one is able to determine all factors that jointly predict treatment change and outcome, weighting up the analysis of time to outcome (by the probability of remaining uncensored), while censoring at the point of any treatment change, assigns appropriate weight to individuals who remain *uncensored* to account for loss of those previously censored who *share* their same characteristics. Given the assumption that all predictors of the *outcome* have been included in the model, this weighting provides a pseudo-population that mimics the same likelihood of occurrence of outcome events that would have been observed if none of the patients' follow up had been censored due to treatment changes.

The IPCW analysis procedure therefore involves three main stages, to be described here, before considering their practical application in the SANAD arm B comparison of LTG versus VPS comparison on T12mR.

9.2.1. Step 1: Fitting the weight determining (WD) model

First, it is necessary to identify the important baseline and time-varying covariates that predict both switching and outcome. If the pool of potential confounders is large, it may be necessary to use some form of **variable selection process**, to avoid the possibility of overfitting the model (leading to model instability due to a low event per variable (EPV) ratio) and including variables that are correlated with each other.

With time-varying covariates, it may be helpful to set up **panel data**, for example (as in SANAD) if the covariates are updated on a daily basis for some patients (for example, for those with multiple daily seizures) and it would be too complicated to adjust for covariates changing on a daily basis. These panel data are created by partitioning each

patient's follow up data into intervals of fixed length, with TVCs which are valid at the start of each interval (e.g. seizure counts for month n equal the total up to the end of month $n-1$).

The probability of remaining uncensored (in other words, *not* experiencing treatment change) at each follow up time point, given these covariates, is then estimated for each patient using the **weight determining (WD) model**, either a Cox model for time to (first) treatment change or pooled logistic regression (PLR) for the discretised interval data with **treatment change** as the **dependent variable**.

It may be more practical to use a discrete time model (based on logistic regression) rather than a model assuming continuous event follow up (survival model) to estimate the probability of remaining uncensored, not only when the TVCs are necessarily discretised into intervals (and therefore naturally lend themselves to discrete time analysis) but also because weight determination analysis with TVCs result in weights that vary within individual over time. Unless the statistical software can allow for within-individual time-varying weights in Cox regression, normal standard errors (SEs) from the Cox model (ignoring the patient identifier in the analysis) will be inaccurate and it will be necessary to implement bootstrapping (BS) for valid confidence interval (CI) estimation.

Pooled logistic regression

Therefore, as an alternative to the Cox model (using BS CIs), it may be helpful to instead consider applying the variety of logistic regression known as pooled logistic regression (PLR), in which follow up time is split into uniform short time intervals such that LR can be performed independently in each interval (183). If follow up time is

discretised into sufficiently short time periods such that the probability of event within each interval remains low (less than 10%), logistic regression (LR) approximates Cox model well, so that the estimated HR within each interval can be estimated from the odds ratio (OR) obtained from LR (184). Each observation interval (of equal length) is in essence a mini follow up study in which the current risk factor measurements are used to predict the probability (p) of event of interest in the interval among patients at risk of the event at the start of the interval; thus the probability of *not* experiencing the event ($1-p$) can be calculated for each interval. The cumulative probability of *not* experiencing the event by each time point (mimicking the estimated survival probabilities at each follow up time obtained from the Cox model) are obtained by simply multiplying the ($1-p$) estimated for the current and all previous intervals, leading to the term *pooled* logistic regression (PLR).

Therefore, if PLR is used, the cumulative probability of remaining on assigned treatment is calculated by multiplying a patient's predicted probability of remaining uncensored in the current time interval by all their previous time intervals' estimated probabilities of remaining uncensored. These probabilities are then used to determine, for each patient i at each time point k, their probability of remaining under randomised treatment (in other words, their probability of *not* experiencing a treatment change and therefore remaining uncensored).

Note that, regardless of choice of WD model, it is advisable to calculate the probabilities of censoring separately for each randomised treatment group, as the influence of covariates on the probability of treatment changes are likely to differ according to treatment. As such, separate WD models should be created for each of the randomised treatment groups separately.

Bootstrapped CIs

The incorporation of weights has been shown to reduce SEs (185) (186). Thus bias results when standard methods as used for CI estimation as they regard the weights as fixed, whereas non-standard methods (such as bootstrapping) can explicitly allow for the estimation of the weights, reflecting more accurately the associated reduction in variability. Thus, it may be advantageous to use bootstrapping for CI estimation regardless of the choice (between Cox or PLR) of WD model, in order to appropriately account for the reduction in SE estimates as a result of the weight estimation process.

Splines necessary for PLR

When using PLR in place of a Cox model, one needs to consider how best to model the underlying baseline event rate. In order to mimic the underlying hazard function of Cox regression, while avoiding the need for a separate intercept term for each interval in PLR, spline variables can be used to allow the underlying risk of the event to vary from interval to interval (187). (Although penalised splines are another popular choice that would have been suitable in this context, these were not considered here, as they require specialised software and are thus less generally applicable for wider use by non-experts.)

The simplest example of a spline function is a linear spline function, such that each section between each pair of knots consists of a linear function. For example, with three knots at positions a, b and c, the linear spline function would be defined as

$$f(X) = \beta_0 + \beta_1 X + \beta_2 (X - a)_+ + \beta_3 (X - b)_+ + \beta_4 (X - c)_+$$

where $(u)_+ = \begin{cases} u, u > 0, \\ 0, u \leq 0. \end{cases}$

In practice, linear splines are overly simple and are not smooth; instead, cubic polynomials are more suited to modelling curving distributions. In order to ensure that such splines are smooth at the joining positions (i.e. at each knot), the first and second derivatives of these functions are forced to agree at each knot position. Furthermore, in order to overcome the potential problem of irregular behaviour of such splines in the tails of the distribution (i.e. preceding the first knot and following the last knot) due to sparse data, *restricted cubic splines* which constrain the function to be linear in these tails, are recommended, with the additional advantage of requiring estimation of only $(k-1)$ parameters (instead of the usual $k+3$ parameters required for unrestricted cubic splines) (187). A restricted cubic spline with k knots $t_1, t_2, .., t_k$ is defined by $f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_{k-1}$, where $X_1 = X$ and for $j = 1,2 \ldots, k-1$,

$$X_{j+1} = (X - t_j)_+^3 - (X - t_{k-1})_+^3 \frac{(t_k - t_j)}{(t_k - t_{k-1})} + (X - t_{k-2})_+^3 \frac{(t_{k-1} - t_j)}{(t_{k-1} - t_{k-2})}$$

In order words, for restricted cubic splines, variables are created equal to the positive part of $(x-k_1)^3$, $(x-k_2)^3$, $(x-k_3)^3$ but with restrictions imposed to ensure linearity in the tails (i.e. for $x < k_1$ and $x > k_3$).

Placing knots at fixed percentiles of each covariate is recommended, thus ensuring sufficient data points within each interval and preventing undue influence from outliers on the choice of knot location. For example, for $k = 3$, knots may be placed at the 10th, 50th and 90th percentiles; for $k = 4$, knots may be placed at the 5th, 35th, 65th and 95th percentiles; and for $k = 5$, knots may be placed at the 5th, 27.5th, 50th, 72.5th and 95th percentiles. Durrleman (188) demonstrated that results are generally insensitive to the location of knots unless they are placed in particularly non-uniform locations (for

example at 25th, 75th, 90th, 95th and 99th percentiles). However, if the hazard is likely to change shape or be particularly changeable at a certain time, one or more knots may need to be placed near these change points (187).

The number of knots can impact on the EPV ratio, as a spline with k knots will require $(k-1)$ parameters, in addition to the coefficient for the linear time variable (which necessarily accompanies the spline variable) (thus adding k knots in all). When choosing the number of knots to be used when fitting a spline, one should aim for balance between allowing sufficient flexibility without overfitting the model to the data, which leads to loss of precision (189). Stone suggests that between three and five knots will usually be adequate, depending on number of events (in the case of logistic or Cox regression).

The number of knots to use for the time spline variables must therefore be considered in light of the number of other variables included in the WD model and relative to the number of treatment change events, in order to avoid overfitting the model. If data are available on numerous potential confounders, some of which may be correlated, it may be necessary to impose some form of selection procedure in order to cut down the number of variables in the model.

These spline variables are created in Stata using the command `spbase`:

```
centile m if treat_num==0, centile(5(45)95)
spbase m if treat_num==0, gen(m_spline_LTG) knots(1, 9, 34)

centile m if treat_num==1, centile(5(45)95)
spbase m if treat_num==1, gen(m_spline_VPS) knots(1, 9, 30)

* combine splines into one variable
forvalues i = 1(1)3 {
      gen m_spline_`i' = m_spline_LTG`i' if treat_num == 0
      replace m_spline_`i' = m_spline_VPS`i' if treat_num == 1
}
```

```
forvalues i = 1(1)3 {
      rename m_spline_LTG`i' original_m_spline_LTG`i'
      rename m_spline_VPS`i' original_m_spline_VPS`i'
}
```

Selection of variables

Some advocate not using any selection procedure to determine which covariates to include as confounders in the models of time to treatment change and time to event, but instead to include all covariates that are considered to be clinically important (151). The justification for this approach is to ensure that no important confounders are excluded on the basis of the chosen selection procedure, especially if the selection of variables depends on their statistical significance in predicting these events, as there may be insufficient power in the dataset to demonstrate such significance.

However there is a danger that including too many covariates which are not correlated with outcome may introduce more uncertainty into the analysis results (specifically very variable weights), rather than the intended opposite of reducing variability. Thus, although it may therefore be tempting simply to adjust for *all* known baseline and time-varying covariates, the inclusion of variables that do not predict both outcome and remission will lead to spurious results. Furthermore, there is a danger of overfitting the model if too many variables are included, leading to model instability (manifested by very large coefficients and SEs) due to low EPV ratios (190).

Therefore, in aiming for the most parsimonious model that still explains treatment changes, it may be deemed necessary to use a variable selection process to determine which of the many measured covariates should be adjusted for in this model. The selection of variables for inclusion in a model is usually determined by p-values of potential variables in the model. There are, however, a number of reasons why this would not be an appropriate selection method for this model.

Although both time-varying covariates and baseline variables need to be included in the WD model, the WO model only includes baseline variables that predict remission, in order to increase power. Confounding due to TVCs has been eliminated via weights, simulating a balanced RCT, and explicitly adjusting for the TVCs in the model would introduce bias. As such, there are no p-values generated for the TVCs in the WO model.

Furthermore, as discussed above, the SEs (and hence the p-values) from the weighted PLR model of time to remission and from the Cox model (in particular) are invalid; thus bootstrapped SEs must be constructed. If the selection procedure relied on p-values, it would be necessary to bootstrap the entire selection procedure, adding to the complexity of the selection procedure beyond what would be feasible.

Therefore it is advisable to use an alternative method, which assesses the change in treatment effect HR from WO Cox model (or odds ratio (OR) if using PLR) when each potentially confounding variable is included (if using forward selection) or excluded (if using backward selection) from the IPCW model, using a relatively small threshold (suggested maximum of 10%) to determine whether each variable should be included. For example, when using backward selection, if exclusion of the variable (from the WD model) causes the treatment effect HR (in the WO model) to change by more than 10% (also known as the CHEST (CHange in ESTimate) criterion), the variable would be retained (191).

Practical considerations for this selection process include the choice of **threshold for inclusion of covariates**, choice of whether to use **forward or backward selection** procedures and the inherent reliability of the selection procedure, in terms of **EPV** ratios.

9.2.2. Step 3: Creating stabilised weights

The WD model is then used to provide time-dependent stabilised weights, which are created for each patient i at each time point k during which they remain under randomised treatment, derived from the inverse of their estimated probabilities (of remaining uncensored). As described in Chapter 5, two separate WD models are in fact required in order to generate these stabilised weights: the **first WD model adjusts for baseline** characteristics only, whereas the **second WD model adjusts for both baseline and time-dependent covariates**. Stabilised weights for each time interval are calculated as the ratio of the cumulative probability of remaining uncensored (due to treatment change) from randomisation until (the end of) that interval given only baseline covariates (in other words, using estimated probabilities from the first WD model) divided by the corresponding probability given both baseline and time-dependent covariates (using estimated probabilities from the second WD model).

For example, the numerator of the censoring weights W_k at time point k (where subscript i denoting individual is dropped for simplicity) could be estimated, using a PLR model applied separately for each treatment arm (Z) to determine the likelihood of remaining uncensored at time k dependent on baseline covariates X_0 : $\Pr[C_k = 0|\bar{C}_{k-1} = 0, X_0 = x_0, Z = z]$. Similarly, the denominator of the censoring weights W_k would be estimated, again separately for each treatment arm, using a PLR model to determine the likelihood of remaining uncensored at time k dependent on baseline *and* time-varying covariates valid at the start of time period k (i.e. $\bar{X}_{k-1}$): $\Pr[C_k = 0|\bar{C}_{k-1} = 0, \bar{X}_{k-1} = \bar{x}_{k-1}, Z = z]$.

These estimated probabilities are then used to produce the following stabilised weights

$$W_k = \frac{\Pr[C_k = 0 | \bar{C}_{k-1} = 0, X_0 = x_0, Z = z]}{\Pr[C_k = 0 | \bar{C}_{k-1} = 0, \bar{X}_k = \bar{x}_k, Z = z]}$$

Therefore each individual only contributes observations for the time points during which they remain under their assigned treatment protocol (after which $C_k = 1$).

Once these weights have been created, it is necessary to check their distribution to identify any extreme weights, which may occur for a number of reasons. Firstly, if the majority of patients in a given subgroup (defined by the cross-classification of the baseline and time-varying covariates included in the WD model) change treatment, the remaining patients will be assigned very large weights to account for the huge loss of information from all other patients in this subgroup. Extreme weights will also result if an incorrect functional form for covariates is specified in the WD model, as predictions of probabilities are then based on an incorrect relationship between each covariate and the outcome.

9.2.3. Step 3: Fitting the weighted outcome (WO) model

Once any problems regarding weights have been addressed, the stabilised weights are applied to the WO model of time to outcome event, regressed on randomised treatment group and baseline covariates *only* and artificially censoring patients when they deviate from their assigned treatment. Thus this **weighted outcome (WO) model** (Cox model or PLR) is fit to the randomised group data, applying the time-dependent patient- and interval-specific stabilised weights to the outcome data, with patients who change treatment being assigned a weight equal to 0 (i.e. inducing censoring) in all intervals after treatment change.

As for the WD model, when the probability of the outcome event in each time interval is relatively small, the cumulative probability of remaining uncensored (obtained by multiplying each person's estimated probabilities for each successive month of follow up) estimated using PLR will be approximately equal to the probability obtained using a Cox model (183).

The randomised treatment indicator and baseline covariates included in the WD model are included as covariates in the WO model, but the time-dependent covariates are not included, given that their influence has been removed via the weighting process; indeed adjusting for TVCs in the WO model would introduce bias as would any standard analysis regressing on TVCs. Note that, if the analysis plan specified that no baseline variables were to be adjusted for in analysis, the first "baseline only" WD model (providing the numerator of the stabilised weights) would include no covariates; in other words, for such an unadjusted WO model, it would be necessary to simply apply the *unstabilised* weights = $\frac{1}{\Pr[C_{k+1}=0|\bar{C}_k=0,\bar{X}_k=\bar{x}_k,Z=z]}$.

9.3.Application of IPCW to SANAD B

SANAD is an example of a trial with dynamic treatment regimens, in which, following random assignment of initial treatment, a patient's prescribed treatment over the course of the trial may vary depending on the patient's evolving prognosis (118). When defining the causal research question for such a trial, it may not be clinically sensible to ignore the underlying reasons for these changes, for example, as for a simple comparison of "always treated" versus "no treatment" contrasts, which adjusts for *all* observed treatment changes. Instead a more precise causal contrast may be more

appropriate, given that some treatment changes are inevitable and therefore it may not be relevant to factor them out of causal estimation.

The ability to account for *all* treatment changes in the IPCW model (rather than simply treatment switches, as in the RPSFTM) provides added scope to extend investigation to consider the underlying *reasons* for treatment changes, and thus more adequately address the clinical question of interest.

Thus, it was necessary to consider how the treatment effect estimated by the IPCW model is directly linked to causal question of interest through the adjustment of relevant competing forms of treatment change.

9.3.1. Reasons for treatment changes

When considering the reasons for treatment change in SANAD, in contrast to the RPSTFM, which was only able to adjust for treatment switches between randomised treatments, the IPCW is able to adjust for all varieties of treatment changes, of which there are many. Adjustment for all types of treatment changes answers a slightly different research question from that which was answered by the RPSTFM, by providing an estimate of the relative treatment effect between treatments that would occur in the absence of any departure from randomised AED.

Furthermore, in contrast to treatment changes which occur prior to the TFS (which are primarily due to unacceptable adverse events (UAEs) associated with randomised drug), treatment changes prior to a period of 12 month remission may occur for a number of competing reasons, most commonly UAEs or ISC. These events are inversely related via dose: as dose increases, seizure control is likely to improve but adverse effects may well increase. Withdrawal due to UAE was largely limited to early post-randomisation

period, and withdrawal due to ISC was generally later because upward titration of dose took place before clinicians decided to withdraw patients due to ISC.

Although there is interest in the causal efficacy of treatment generally (obtained by adjusting for *all* treatment changes), clinicians are primarily interested in adjusting for switches that occurred due to ISC rather than because of UAEs. The argument follows that it may be possible to reduce the incidence of treatment changes due to ISC by persevering with the treatment and applying more appropriate dosing (incrementally increasing dose until seizure control is achieved) whereas the issue of UAEs is less easily solved without appropriately ameliorating drugs. Thus, given that it is unlikely that genetic testing will develop sufficiently to allow identification of patients who will be able to tolerate such AEs, adjustment for this cause alone isn't a relevant clinical question. Instead the causal treatment effect estimate of primary interest to clinicians is that which adjusts for switches occurring due to ISC alone (private discussion with Marson, Smith 2012).

Competing risks approach

Given the medical interest in the competing reasons for treatment change, it may seem natural to consider adjustment for multiple causes of treatment change using a standard competing risks approach (as demonstrated by (151)), whereby each patient's follow up is censored at the time of the first treatment change, with separate weights being calculated for each of the (first occurring) different competing reasons for treatment change. These weights can then be applied in separate analyses (to adjust for each type of treatment change independently of the others) or combined into overall weights by simple multiplication (under the assumption that the covariates in the model remove any association between these weights) (151).

Indeed, in the analysis of time to treatment withdrawal, it is relevant to consider competing reasons for treatment change, as the effect of ISC and UAE may compete, cancelling out the effect of a certain treatment on outcome (192). However when considering *adjustment* for treatment change, the question of whether to adjust for treatment changes occurring for different reasons instead directly relates to the relevance of these treatment changes to the causal research question. Thus, rather than applying standard competing risks methodology to the context of IPCW modelling (151), it was deemed more appropriate to approach this investigation into the effect of different reasons for treatment change using a sequence of IPCW analyses, determined according to the relative clinical importance of each reason.

This alternative approach was possible primarily due to the luxury provided by this data set, such that follow up generally continues beyond the point of each treatment change; therefore, it was not automatically necessary to censor patient follow up at the point of the first treatment change. As such, if a particular form of treatment change was not relevant to the causal research question being investigated, the treatment change could simply be ignored without introducing any informative censoring; for example, if considering what the causal effect of treatment would be in the absence of treatment changes due to ISC alone, none of the other treatment changes would be adjusted for in the weighting or outcome model.

Sequential analyses according to reasons for treatment changes

Thus, an alternative, more appropriate approach to adjusting for different reasons for treatment changes in this context was to consider a sequence of analyses, adding different reasons for treatment changes sequentially according to clinical importance. Therefore, as adjustment for reasons related to ISC was of primary interest in SANAD,

this was the first reason for treatment change included in the hierarchy. Then, assuming that it would be easier to persuade a patient to remain on treatment when they are considering changing treatment for reasons of personal choice (for example, due to non-compliance or potential to become pregnant), this was the second reason added to the hierarchy, with the final analysis including all treatment changes (in other words, including UAEs also, which is considered the least easily or ethically altered treatment change).

Analysis mimicking ITT was also possible, in other words, one that adjusts for no treatment changes at all and thus requires no weight estimation. Furthermore, the possibility of carrying out an additional analysis, similar to an ITT analysis but which allows for the possibility of informative censoring due to LTFU, was considered. The probability of such censoring can be modelled in a similar way to the probability of treatment change, using the appropriate baseline and time-varying covariates in a PLR or Cox model; the inverse probability of *censoring* weights can be multiplied by the inverse probability of *treatment change* weights in order to obtain an *overall* weight to be used in the model of time to 12 month remission. (The validity of simply multiplying these weights to obtain an overall weight holds if one is able to assume that inclusion of the baseline and time-varying covariates in the model removes the dependence between these two events, censoring and treatment change.) This analysis would provide an estimate of the expected treatment effect if no informative censoring had occurred.

In the context of SANAD, however, although it would have been interesting to investigate the effect of censoring (of time to remission) that occurred as a result of LTFU, non-epilepsy-related death or withdrawal of consent (collectively referred to as LTFU), such an analysis was not possible, as there were very few patients who were

censored prior to remission or treatment change. Thus it was not feasible to reliably model separate weights for the probability of treatment change versus censoring. However, as there were very few censoring events, any associated bias due to this censoring is unlikely to unduly influence the overall results. Instead, therefore, given that the ITT analysis assumes that censoring is non-informative, this assumption was applied to *all* censored observations throughout the hierarchy of analyses (even when patients were LTFU, withdrew consent or died for reasons not related to epilepsy), implying that such weights (to reduce bias due to early censoring) were not necessary.

Thus the sequential structure to the investigation of the effect of covariates on treatment change and remission took the following format, with interpretation of the corresponding causal research question given in Table 37.

Table 37 Sequence of IPCW analyses, relating to reason for treatment changes

	Artificial censoring	**Interpretation**
ITT (no treatment changes)	No artificial censoring	Observed treatment effect demonstrating **effectiveness of treatment assignment**
ISC alone	Artificial censoring at the time of treatment changes occurring due to ISC	Treatment effect estimate that would have been observed if it had been possible to **prevent treatment changes occurring due to ISC** (of primary interest to clinicians)
ISC/personal choice	Artificial censoring at the time of treatment changes occurring due to ISC or personal choice	Treatment effect estimate that would have been observed if it had been possible to **prevent treatment changes occurring due to ISC or for reasons of personal choice**
All (ISC/personal choice/UAEs)	Artificial censoring at the time of (the first of) any treatment change	Treatment effect estimate that would have been observed if it had been possible to **prevent all treatment changes occurring**

In practice, this artificial censoring affects analysis only if the (relevant) treatment change occurred prior to the remission time. If the treatment change occurred before the patient achieved remission (or before their follow up for remission was censored), their remission time T_i would be brought forward to the time of their (first) treatment

change F_i and the corresponding remission indicator R_i (denoting whether or not the patient had experienced remission ($R_i = 1$) or had simply been censored ($R_i = 0$)) would be changed to indicate censoring ($R_i = 0$). If the treatment change occurred post-remission, however, the remission time remains unchanged. Furthermore, outcome analysis is affected only if the reason for the patient's *first* treatment change corresponds with the causal question; for example, when adjusting only for treatment changes due to ISC or personal choice, a patient's remission time would *not* be artificially censored if their first treatment change occurred because of UAE.

9.3.2. Causal diagram

In applying this method to the SANAD B (VPS versus LTG) data, there were a number of factors requiring careful consideration. Prior to making decisions regarding analysis, it was helpful to sketch a causal diagram, such as the one in Figure 9, to clarify the causal scenario and summarise the interplay between (initial and updated) treatments, prognosis (in terms of baseline and time-varying factors related to the patients' condition) and remission. In this simplified diagram, Z denotes randomisation, A_0 initial prescribed (i.e. randomised) AED, A_1 first change in treatment determined by patient's state at time 1 (denoted by time-varying prognostic factors X_1 such as UAE or seizure occurrence), U underlying baseline prognosis and Y refers to outcome (time to remission).

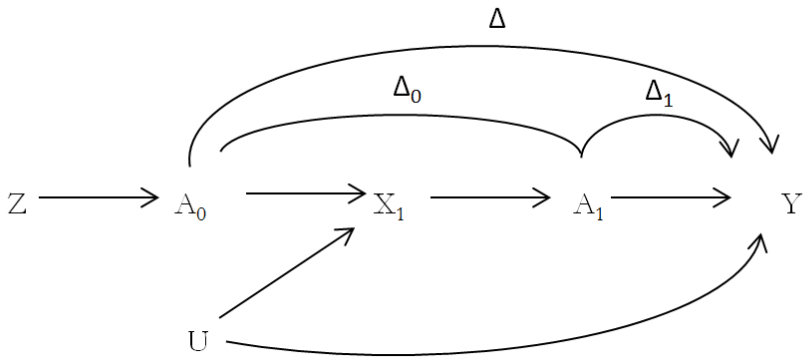


Figure 9 Simplified causal diagram for SANAD T12mR

A causal diagram is particularly helpful when applying IPCW methodology to complex scenarios such as SANAD. First, the diagram may help when deciding which causal effect is of interest, for example, perhaps that corresponding to the effect of initial treatment (only up to the point of a certain clinical event, such as UAEs) (Δ_1) or to the whole treatment experience (Δ). Furthermore, the diagrams prompt one to identify all baseline and time-varying factors that are likely to confound the relationship between treatment received and outcome, such as dose, seizure control and AE occurrence. Indeed this was one of the main challenges when applying the IPCW methodology in SANAD, as data were collected on a large number of such potential confounders.

Thus one of the main complication of this modelling exercise was in determining which of the numerous (time-dependent and baseline) variables should be included in the WD models. In SANAD there were a number of important baseline characteristics that conceivably influence the occurrence of both treatment changes and remission, including eight categorical variables (gender, febrile seizure history, first degree relative with epilepsy, treatment history, neurological insult, EEG result at randomisation, seizure type and epilepsy type) and three continuous variables (age, time interval between first ever seizure and randomisation, and total number of tonic clonic seizures

prior to randomisation). Additionally the three time-varying covariates (TVCs) (prescribed dose of randomised drug, and the number of AEs and seizures since randomisation) were all conceivably influential on the occurrence of changes from randomised treatment and on time to remission.

Prior to deciding which of these covariates should be included in the WD and WO models, by way of a selection procedure, it was necessary to decide how to model the TVCs in a manner that was clinically meaningful and statistically viable; this was extra challenging in SANAD given the highly skewed nature of these variables due to extreme values of these covariates experienced by a minority of patients.

9.3.3. Time varying covariates

The first obvious TVC was (cumulative) seizure count (since randomisation), as seizure occurrence impacts on the likelihood of treatment change as well as remission. Although seizure count is in fact structurally related to the remission outcome, its adjustment in this setting is nevertheless valid (and indeed essential), given that (as a TVC) it is included only in the treatment change (WD), rather than the remission (WO), model.

Data on seizures were collected in line with the pragmatic nature of trial; given that it would have been too difficult to collect date of each individual seizure, patient CRFs simply collected dates of first and last seizures and total seizure count since last visit. This imprecise record of seizure times, coupled with the potential for multiple daily seizures for some patients (such that their value of this TVC changed on a daily basis) suggested that instead of using individual estimated seizure times to calculate daily

updated records of cumulative seizure counts, a sensible compromise would instead be to use **interval data**.

In order to calculate such interval data, the date of individual seizures were first estimated based on the assumption of regular intervals between seizures (by subtracting date of first seizure from date of last seizure, and dividing by the total number of seizures to get an estimate of the interval between each seizure, and then working out approximate dates of seizures based on this interval). These estimated seizure dates were then used to create cumulative interval-based counts of seizures, leading to a step function that changed on an interval (for example, monthly or weekly) basis.

The second obvious confounder, being related to both remission and treatment changes, was the occurrence of AEs. Unlike seizure dates, the date of onset of each type of AE was recorded precisely in patient CRFs, along with information on whether the patient was hospitalised and the length of their stay in hospital. In order to account for the severity of the AE, the occurrence of any AE which resulted in hospitalisation was weighted by the number of nights spent in hospital. Hence the cumulative AE variable increased by a value of one on the date that each AE occurs, except when the AE required hospitalisation, in which case the variable increased by a value equal to the number of nights spent in hospital.

The third potential confounder was dose of randomised treatment, which was recorded at each clinic visit (and was assumed to be unchanged and adhered to between visits).

When these TVCs were discretised into interval values, the value assigned to each interval was that which was valid at the *start* of each interval, such that this value was then assumed to stay constant for that entire time period (to ensure that the values were

not affected by any event occurring within that interval). For example, the cumulative AE (or seizure) counts for time interval n equaled the total up to the end of interval $n-1$. Similarly, the interval value for dose at a particular interval was the last recorded dose from within the previous interval.

Width of time interval

Prior to modelling using these TVCs, it was therefore necessary to decide on the optimal length of time interval, in order to strike the balance between greater accuracy (which increases as interval length decreases) and computational intensity (which increases with interval length). Taking into account the frequency and duration of follow up information in this analysis, it appeared sensible to use fortnightly intervals; however, in order to check whether this choice of interval length was overly influential on the results, repeat analyses using weekly and monthly intervals for these time-varying cumulative counts were also planned.

9.3.4. Covariate issues relating to stability of model

The next feature of the model that required investigation related to the impact of covariate profiles on the stability of the model, firstly with regard to the highly skewed distribution of all of the continuous baseline and time-varying covariates. It was necessary to consider how reduce the influence of these outlying values in the regression modelling. Royston (193) advocates truncating continuous variables at their 1% and 99% centiles, in order to prevent unstable modelling due to overly influential extreme values; thus this **truncation** was applied to these variables.

The next issue regarding covariate distributions related to the small number of patients in some categories of the baseline categorical variables; for example, an uneven split of

patients among treatment history and seizure types meant that some categories contained very patients or events, which would lead to unstable modelling. As such, clinical guidance was sought on how to collapse categories within these variables to increase statistical stability whilst preserving clinical relevance. For example, the two "previously treated" categories of the treatment history variable (current monotherapy (M) and relapse following previous remission and withdrawal from therapy (R), in which there was no variation in the occurrence of events for certain treatment/history combinations (as shown in Table 38) such that these patients were in effect excluded from PLR analysis) were combined so that patients were instead simply classified as either previously treated or untreated.

Additionally, given its high correlation with the binary variable epilepsy type (generalised or unclassified), it was decided that the seizure type variable (with five different categories, some of which contained very few patients) would not be considered for potential inclusion in the model. In reducing the number of potential (categories of) variables available for inclusion in the model, these considerations also helped to address the problem of low EPV ratios.

Table 38 Number of patients with treatment withdrawal out of total number of months of follow up (out of total number of patients)

Treatment change due to	Treatment group	Monotherapy	Relapse after remission	Untreated
ISC	LTG	5/268 (5/13)	1/184 (1/6)	44/2941 (44/178)
	VPS	4/190 (4/14)	0/110 (0/7)	22/2881 (22/175)
ISC/choice	LTG	5/168 (5/13)	2/167 (2/6)	46/2926 (46/178)
	VPS	4/190 (4/14)	0/110 (0/7)	25/2885 (25/175)
All treatment changes	LTG	6/242 (6/13)	3/90 (3/5)	56/2502 (56/166)
	VPS	4/178 (4/13)	2/79 (2/7)	50/2355 (50/169)

9.3.5. Low power

As already mentioned, there were too few patients who experienced LTFU (prior to their first treatment change or remission outcome) to allow independent investigation into the effect of LTFU on remission. However, as each analysis proceeded in the series of treatment change scenarios, it similarly became apparent that there were low numbers of treatment changes in each treatment group on which to base the treatment-specific WD models for the "restricted" treatment change scenarios. Thus, as can be seen in Table 39, treatment changes were especially low in the VPS arm, in particular those occurring due to ISC alone (26) or ISC/choice (29). Thus the corresponding analyses had little power to determine which confounders were influential on the treatment effect (HRs or ORs from Cox or PLR WO models, respectively) for these scenarios.

Table 39 Number of events

	Withdrawal		Remission	
	VPS	**LTG**	**VPS**	**LTG**
None (ITT)	-	-	154	143
ISC only	26	50	141	110
ISC/choice	29	53	139	108
All (ISC/choice/UAEs)	56	64	114	90

To put this into context, allowing for $(k-1)$ dummy variables required for a categorical variable with k categories, there were 20 variables being considered for each FS and BE selection process: three time-varying covariates (dose, cumulative seizure count, cumulative AE count), three baseline continuous variables (age, time between first ever seizure and randomisation, total number of tonic clonic seizures ever before randomisation), four binary variables (sex, neurological insult, febrile seizures, first degree relative with epilepsy) and three categorical variables with more than two

categories (treatment history: untreated, taking (non-trial) monotherapy at randomisation, relapsed after previously achieving remission following treatment; EEG result: epileptic abnormality, nonspecific abnormality, normal, not done; seizure type: generalised tonic-clonic seizures only, absence seizures, myoclonic or absence seizures with tonic-clonic seizures, tonic-clonic seizures uncertain if focal or generalised, focal seizures with or without tonic-clonic seizures, and other unclassified).

Thus, given the rule of thumb that a stable model requires at least 10 events per variable considered in a model selection process (180), the validity of using a FS or BE selection processes was called into question, especially in the restricted withdrawal scenarios (ISC or ISC/choice) in the VPS arm.

There were a number of options to prevent model overfitting and instability: either reducing the number of variables considered for selection; carrying out only FS (as the full models (including all variables) in BE would exhibit extreme overfitting); or only fitting models for the "all treatment changes" scenario. The last of these would be the least favourable option, as clinicians are primarily interested in the "ISC" scenario. It was decided, therefore, to cut down the initial pool of potential variables before FS or BE were carried out, by applying a cut-off of at least 1% change in HR; in other words, when the inclusion of the variable in the WD model caused the treatment effect estimated in the WO model to vary by more than 1% compared to the estimated treatment effect when the variable was excluded (and therefore the WD contained only the time spline variables). The likelihood of overfitting would also be checked by explicit reference to the EPV ratio for each model, relating to the maximum model size for each (considering in particular the full initial model for BE).

This consideration of EPV ratio relates directly to the next issue that required consideration for this modelling exercise.

9.3.6. Knots for spline variables

The next major consideration for analysis regarded the number and positioning of the knots required for spline variables in the PLR analyses. These spline variables are intended to mimic the underlying probability of treatment change (in the WD model) or remission (in the WO model) over time.

In order to determine the optimal number and position of knots for the spline variables, a comparison of PLR results using 3, 4 and 5 knots were made against those obtained from a Cox model. As this analysis progressed, there were a number of key observations that are worthy of note relating to the choice of number of knots and their positioning.

Initially it was decided that the splines would be generated using the number (and position) of spline knot positions as recommended by Stone (189); hence, for $k = 3$, knots were be placed at the 10^{th}, 50^{th} and 90^{th} percentiles of the observed follow up time intervals (weeks, fortnights or months); for $k = 4$, knots were placed at the 5^{th}, 35^{th}, 65^{th} and 95^{th} percentiles; and for $k = 5$, knots were placed at the 5^{th}, 27.5^{th}, 50^{th}, 72.5^{th} and 95^{th} percentiles.

First, it is important for users to appreciate that ***different*** **spline variables** must be created for the **WD and WO models**, according to the (chosen) percentiles of the follow up times for the treatment change and outcome events respectively. The spline variables for the WD model should be treatment-specific, based on the centiles of observed treatment change times in each treatment group separately, whereas the WO

spline variables are positioned according to centiles of the observed outcome times for the overall trial.

If, instead, the *same* spline variables are used (for the WD and WO models), the PLR model becomes unreliable and unstable. This instability results because the use of the treatment-specific splines (generated for the WD models) in the WO (remission) models in turn means that separate baseline odds are estimated for each treatment group in the WO models, implying a complex underlying time-treatment interaction (as opposed to the overall baseline hazard estimated by the Cox model). For example, in Table 45 of Appendix A, it can be seen that, if the (treatment-specific) splines generated for the WD model are applied to the WO, the PLR results become highly sensitive to the number of knots, demonstrating highly variable treatment estimates across time intervals and treatment change scenarios compared to the corresponding Cox model estimates. Given that this occurs even under the ITT scenario and without weights (in other words, when no TVCs were adjusted for), this model instability is obviously due to mis-modelling of the main time effect; otherwise, with correct specification of the time effect, the PLR model (including no baseline or TVCs) would have more closely mirrored the Cox model.

Thus it was necessary to create separate spline variables for the WD and WO models, the former being treatment-specific and the latter overall for the whole trial. When considering the positioning of knots for the WO model, in the particular case of SANAD, it was important to recognise that the remission outcome could not (by definition) occur prior to 12 months. However, at 12 months, there is then a peak in the number of events occurring, due to a high proportion (approximately 30%) of patients achieving I12mR. Thus a major non-linearity (a spike) of the underlying hazard occurs

at this time point, presenting a challenge for the spline variable to adequately fit the data.

Investigation into the empirical distribution of the probability of treatment change and remission over time demonstrated the obvious nonlinearity in remission at 12 months (for example, see plots in Figure 10 of these empirical probabilities when all treatment changes are accounted for).

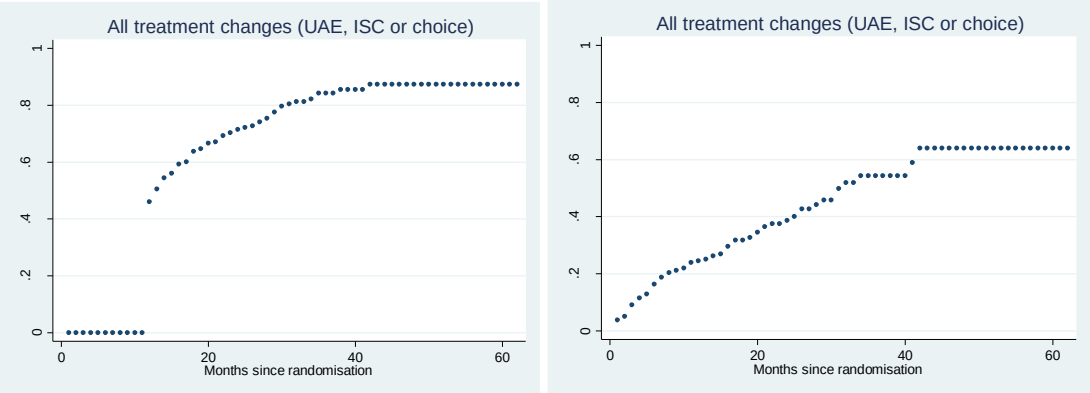


Figure 10 Empirical remission and treatment change probabilities, adjusting for all treatment changes

It was therefore necessary to consider how best to construct the (WO) spline variable around the 12 month point in order to ensure the model fit the data well at this point, for example by fixing one or more knots around 12 months and dropping observation times less than 12 months when calculating percentiles. Note that in the case of the latter, the effect of TVCs and baseline covariates on the probability of withdrawal prior to 12 months would still be modelled in the WD (treatment change) models, but their (cumulative) effect up to 12 months would only be included as weights in the WO (remission) model from 12 months onwards.

An investigation into model fit (assessed by stability of treatment effect and baseline hazard (odds) fit between Cox (and pooled) LR models) was therefore undertaken to

assess the effect of the number and position (particularly in relation to the 12 month time point) of splines on the simple (covariate- and weight-free) remission model, using overall spline variables for each of the following four scenarios:

1) All follow up times were modelled, with centiles (for knot positions) calculated from all follow up times.
2) All follow up times were modelled, with centiles (for knot positions) calculated for all follow up times and 1 (for 3 knots scenario) or 2 (for 4-5 knots scenario) knots fixed at (or clustered around) 12 months.
3) All follow up times were modelled, with centiles (for knot positions) calculated only for follow up times ≥ 12 months and spline variables for time < 12 months were forced to be 0.
4) Only follow up time ≥12 months were modelled, with centiles (for knot positions) calculated only for follow up times ≥ 12 months.

Remission models were assessed using pseudo log likelihood (pll) and AIC measures (where a lower AIC demonstrates better model fit) and by comparing treatment effects across scenarios, as well as baseline hazards and odds plots from Cox and PLR models respectively (see Table 46 to Table 48 in Appendix B). Assessment of the effect of number of knots on these unweighted remission models demonstrated a problem of non-convergence most particularly for scenarios 1-3. As would be expected, given that there is no chance of the event occurring prior to 12 months, the relative stability of results from scenario 4 suggested that it would be sensible to exclude follow up prior to 12 months (for centile estimation as well as outcome modelling).

Indeed, the baseline hazards and odds plots from the comparison between the Cox and pooled LR models (with 3 to 5 knots) demonstrate an (expected) high hazard at 12

months followed by a rapid decline and subsequent little variation over time. As such, it was decided that this pattern would be best fit by **dropping any time at which there are no events occurring** (which automatically included times less than 12 months) and **fixing one or more knots around 12 months**. Given that the outcome is rare, it would also be advantageous to fix knot positions at **centiles of the distribution of observed *event* times**, rather than all observation times. Note that dropping times at which there are no events occurring does not affect modelling (as event probability at such times are perfectly predicted by the model, given that there is no variation in outcome at this point, and are therefore automatically dropped at this point); the only effect would be to reduce the range of data used to construct the splines, leading to greater stability in the model.

Applying these restrictions reduced the problems of non-convergence, but there was still evidence of overfitting in those scenarios with fewer observations (with longer time intervals) and fewer events (more restrictive treatment change scenarios). However the results were not definitive regarding the most appropriate choice for the number of knots to use for each scenario; although non-convergence was an issue as the number of knots increased in scenarios with fewer observations or events, the AIC (and loglikelihood) measures tended to favour models with more knots. As such, it was decided that the **variable selection procedure** would be carried out using both 3 and 5 knots for each scenario in the proposed analysis plan, with due attention being paid to the stability of the estimated coefficients and their SEs, as well as to the EPV ratio (as an indicator of the reliability of the model).

Given that a maximum threshold of 10% is recommended when applying a selection procedure according to the CHEST criterion (191), a threshold of 5% was chosen for

the main selection procedure, with sensitivity analyses using 2 and 10% to assess whether the conclusions of the analysis varies according to threshold. Furthermore, it was decided that both forward selection (FS) and backward elimination (BE) would be employed, in order to provide insight into the impact of the selection procedure on the choice of covariates, with consideration for the stability of each model assessed using the maximum EPV ratios for each selection process (for example, the EPV corresponding to the full initial model when undertaking BE).

9.3.7. Analysis plan

Given the uncertainty surrounding these modelling issues, an analysis plan was developed to facilitate an organised and structured approach in order to investigate the relative impact of each of the variable factors on the model results. For each of the treatment change analyses (namely adjusting for no treatment changes (ITT); treatment changes due to ISC alone; treatment changes due to ISC or personal choice; and finally all treatment changes) there were four factors to consider: model type (Cox regression (with bootstrapping) or PLR (with and without bootstrapping)), time intervals for discretised TVCs (weekly, fortnightly or monthly), selection procedure (FS or BE) and variable selection threshold (2%, 5% or 10%). Thus the IPCW modelling procedure was applied using a factorial design, allowing the impact of each of the following model features on estimated treatment effects to be assessed, described in Table 40.

The analysis plan therefore consisted of a comparison of three versions of the IPCW model for the SANAD comparison: Cox regression (with bootstrapping), PLR (with bootstrapping) and PLR (without bootstrapping). Comparison of PLR with and without bootstrapping would assess the validity of the SEs produced by the model in

comparison to bootstrapped SEs. Comparison of PLR and Cox regression (both with bootstrapping) would assess how similar the models are, both in terms of the variables that are selected for inclusion in the model and in the estimated treatment effect estimates (bearing in mind, however, that the models will not be directly comparable as the logistic regression model estimates ORs, and the Cox model estimates HRs). In particular it was of interest to determine whether the use of smaller time intervals (for example, weekly or fortnightly intervals compared with monthly intervals) would lead to more similar results between the logistic and Cox regression, as would be expected (given that event rates decrease with decreasing time intervals, and therefore the estimated ORs should more closely match the corresponding HR estimates).

Table 40 Analysis plan

	Levels of analysis
Reason for treatment change	ITT (no treatment changes) ISC alone ISC/personal choice All (ISC/personal choice/UAEs)
Model	Cox regression (with bootstrapping) PLR (with bootstrapping) PLR (without bootstrapping)
Time interval	Weekly Fortnightly Monthly
Selection procedure	Forward selection (FS) Backward elimination (BE)
Selection threshold	2% 5% 10%

Bootstrapping was carried out using by drawing (200 repeat) samples at the patient level (in other words, selecting the entire record for that patient) rather than at an individual

record level (in other words, restricted to a particular time interval for that patient) to ensure correct allowance for clustering**.**

In linking with the causal question of interest, the estimation sample for the IPCW modelling excluded patients with focal epilepsy (as for the RPSFTM) as these patients were not intended to receive VPS or LTG treatment; the estimation population thus consisted only of those with generalised or unclassified epilepsy.

Research questions

This analysis plan sought to provide answers to the following pertinent questions:

1. Which baseline and time-varying covariates are identified as **confounders** between treatment changes and remission?
2. Regression models
    a. How does **bootstrapping** compare with normal CI estimation?
    b. Are the same variables selected using **Cox and PLR** (with three and five knots)?
    c. Do **treatment effect estimates** differ greatly between regression models?
3. Time intervals (for logistic regression)
    a. How does the **time interval** affect selection of variables and subsequent treatment effects?
    b. Are treatment effects from **Cox and PLR models more similar** with smaller (weekly) intervals?
4. Reasons for treatment change

a. What are the general differences in the selection of variables and subsequent treatment effects when **different treatment changes** are adjusted for?
b. What is the **overall treatment effect** adjusting for specified treatment changes?

9.3.8. Data preparation

In order to carry out this modelling, the data needed to be prepared appropriately, which was not a trivial exercise. In particular, determination of the cumulative seizure counts, whether at monthly, fortnightly or weekly intervals, required careful programming in Stata.

Seizure counts

First, it was necessary to create a total seizure count for each patient across multiple visits

```
bysort trialid: gen cum_total_seizures = sum(totsez_num)
```

It was then necessary to generate individual seizure dates (which were assumed to lie equally between the first and last date of seizure recorded for each visit)

```
forvalues i = 1(1)1000 {
      gen date_sez_raw_`i' = .
      gen date_sez_`i' = .
      }
```

in order to convert these into monthly (or fortnightly or weekly) interval counts. For example, monthly counts were calculated using the following code:

```
forvalues i = 1(1)30 {
            gen cum_sez_m`i' = .
            }
```

```
forvalues i = 1(1)30 {
      local j 1
            while `j' <=1000 {
                  bysort trial: replace cum_sez_m`i' = `j' if
((date_sez_`j' -  rand) > 30*(`i'-1)) & ((date_sez_`j' -  rand) <=
30*(`i')) & _n==1
                  local j = `j'+1
            }
}

forvalues i = 1(1)30 {
      forvalues j = 1(1)1000 {
            bysort trial: replace cum_sez_m`i' = `j' +
cum_total_seizures[_n-1] if ((date_sez_`j' -  rand) > 30*(`i'-1)) &
((date_sez_`j' -  rand) <= 30*`i') & _n > 1
      }
}
replace cum_sez_m1 = 0 if cum_sez_m1 == .

forvalues i = 2(1)30 {
      local j = `i'-1
      sort trial visit
      bysort trial: replace cum_sez_m`i' = cum_sez_m`j' if
cum_sez_m`i' == .
      local i = `i'+1
}
```

Adverse event counts

AE counts were weighted according to the number of nights spent in hospital

```
gen AE_num = 1 if hosp =="N"
replace AE_num = nights if hosp=="Y"

gen date_AE = datons
gen month_AE_fraction = (date_AE - rand)/30
gen month_AE_fornextmonth = ceil(month_AE_fraction)

* cumulative AE at the end of one month counts towards next month
gen month_AE = month_AE_fornextmonth + 1
note  month_AE:  total  AEs  up  to  start  of  previous  month  (ie
corrected)

sort trialno month_AE date_AE
* keep only one entry per month
bysort trialno month_AE: keep if _n==_N
```

Dose data

Dose data were recorded whenever a patients' prescription (treatment or dose) changed, and these data needed to be converted into monthly (or fortnightly or weekly) values.

```
forvalues i = 1(1)63 {
      gen dose_prelim_m`i' = .
}
* dose is carried forward from anytime in the previous month
forvalues i = 1(1)63 {
      replace dose_prelim_m`i' = dose if ((date_pres_fup -  rand) >
30*(`i'-1)) & ((date_pres_fup -  rand) <= 30*`i')
}

* if more than one entry for a given month, use the last entry
forvalues i = 1(1)63 {
      replace dose_prelim_m`i' = . if dose_prelim_m`i'[_n+1]~=. &
trial==trial[_n+1]
}
* create single line per patient
forvalues i = 1(1)63 {
      bysort trial: egen dose_m`i' = total(dose_prelim_m`i'),
missing
}
bysort trial: keep if _n==1
drop dose_prelim*
forvalues i = 2(1)63 {
      local j = `i'-1
      replace dose_m`i' = dose_m`j' if dose_m`i' == .
}
forvalues i = 1(1)63 {
      count if dose_m`i' == .
}

local i = 23
while `i' <= 23 & `i'>0 {
      local j = `i'+1
      replace dose_m`i' = dose_m`j' if dose_m`i' == .
      local i = `i' - 1
      }

reshape long dose_m, i(trialno) j(m)
```

Merging time-varying covariates

It was then necessary to merge the three time-varying covariates appropriately according to time interval.

```
gen month_AE_sez = m if include_AE==1
replace month_AE_sez = m if include_sez==1
sort trial month_AE_sez

* check only one entry per month for AE/sez
sort trial m
assert m > m[_n-1] if trial==trial[_n-1] & m~=. & m[_n-1]~=.

sort trial m
assert m > m[_n-1] if trial==trial[_n-1] & m ~=. & m[_n-1]~=.
```

```
sort trial month_AE_sez include_sez
gen cum_AE_month_withsezdata = cum_AEnum
bysort trial: replace cum_AE_month_withsezdata = 0 if
cum_AE_month_withsezdata==. & _n==1

* if no cumulative AE total for sez month, copy from previous
cumulative AE total
bysort trial: replace cum_AE_month_withsezdata =
cum_AE_month_withsezdata[_n-1] if month_AE_sez < month_AE_sez[_n+1]
& include_sez==1 & cum_AE_month_withsezdata==. & _n>1
* if cumulative AE total for sez month (ie more than one entry for
that month - one relating to AEs, one relating to seizures), copy
from following cumulative AE total (from that month)
bysort trial: replace cum_AE_month_withsezdata =
cum_AE_month_withsezdata[_n+1] if month_AE_sez == month_AE_sez[_n+1]
& include_sez==1 & cum_AE_month_withsezdata==. & _n>1

sort trial month_AE_sez include_AE
gen cum_sez_month_withAEdata = cum_sez_m
bysort trial: replace cum_sez_month_withAEdata = 0 if
cum_sez_month_withAEdata==. & _n==1
bysort trial: replace cum_sez_month_withAEdata =
cum_sez_month_withAEdata[_n-1] if month_AE_sez < month_AE_sez[_n+1]
& include_AE==1 & cum_sez_month_withAEdata==. & _n>1
bysort trial: replace cum_sez_month_withAEdata =
cum_sez_month_withAEdata[_n+1] if month_AE_sez == month_AE_sez[_n+1]
& include_AE==1 & cum_sez_month_withAEdata==. & _n>1

* fill down cum_AEnum values now that we have full months up to 62
months
sort trial m
by trial: replace cum_AEnum = 0 if _n==1
by trial: replace cum_AEnum = cum_AEnum[_n-1] if cum_AEnum==. & _n>1
```

It was also necessary to create time-varying treatment change and remission (indicator and time) variables corresponding to each analysis in the sequence of analyses investigating reasons for treatment change.

```
* create reason for dropout or withdrawal
gen w_reason = 0 if wclass=="C"|wclass=="D"|wclass=="L"|wclass=="RE"
* note that there is one patient who was censored because of "RE" at
< 365 days (trialno B1092009 withtime 255, wcens 0 but remtime 365,
rcens 1)
* ? consider sensitivity analysis treating this patient as having
event (withdrawal of treatment) at 225 days
replace w_reason = 1 if
wclass=="ISC"|wclass=="DISC"|wclass=="ISCUAE"
replace w_reason = 2 if wclass=="PAE"|wclass=="PRE"|wclass=="UAE"
replace w_reason = 3 if
wclass=="NC"|wclass=="O"|wclass=="P"|wclass=="PD"|wclass=="U"
replace w_reason = 4 if wclass=="NA"
label variable w_reason "Reason for withdrawal from drug or study"
```

```
label define w_reason 0 "Ltfu, death or consent withdrawn" 1 "ISC" 2
"UAE" 3 "Personal reasons" 4 "NA"
label values w_reason w_reason

* Changing for any reason: w_reason = 1/2/3; or censoring: w_reason
= 0; hence wr0123

gen rem_with_wr0123_min_date = min(withdate, remdate,
totsez_censdate)
label variable rem_with_wr0123_min_date "Minimum of trt withdrawal
(for any reason, including end of fup and lftu (both have wcens==0))
or remission dates"
```

For example, time of treatment change was brought forward to the time of remission *and* censored, if the time of treatment change was greater than the remission time; alternatively it was necessary to bring the remission time forward to the time of treatment change *and* to censor (if the time of treatment change was less than the remission time).

```
gen min_cens_withdate = min(totsez_censdate, withdate)
gen rcens_wcens_wr0123 = rcens
replace rcens_wcens_wr0123 = 0 if rem_with_wr0123_min_date < remdate
label variable rcens_wcens_wr0123 "Remission indicator censored at
first treatment change (for any reason)"

* create rem_with_min_month/biweek/bimonth: so that any event during
the present month/biweek/bimonth counts towards the end of the
previous month/biweek/bimonth
* note that the (n-1)th values of tvcs are aligned with nth values
of outcome (see bottom of p413 in Fewell) as we have brought the (n-
1)th values forward to the nth:
* need cumulative sum up to m`i' to count towards next m (`i' + 1)
* sort trial m
* bysort trial: gen cum_sez_m = final_cum_sez_m[_n-1] if _n>1
* bysort trial: replace cum_sez_m = 0 if _n==1
gen rem_with_wr0123_min_month = floor((rem_with_wr0123_min_date -
rand_date)/30)
* gen rem_with_wr0123_min_biweek = floor((rem_with_wr0123_min_date -
rand_date)/14)
* gen rem_with_wr0123_min_week = ceil((rem_with_wr0123_min_date -
rand_date)/7)
```

These times and indicators variables varied according to the particular reason for treatment change; for example, changes occurring due to ISC would be adjusted in this manner for all analyses except the ITT analysis, whereas changes due to UAEs were only adjusted for in the final overall adjustment for all treatment changes.

Once these analysis-specific variables were created, they were converted into **time-varying event indicator variables** in order to allow them to be linked with the TVCs.

The ITT version of the models (for both survival and PLR) required no WD model, as there were by definition no treatment changes to account for, and thus there were no weights applied to the WO (remission) model; as such, TVCs were not required for ITT analysis (as they enter only through their influence in the WD model).

9.3.9. Selection process

The forward and backward selection processes required manual programming in Stata, which increased in complexity as the number of variables were added to (or included in) the model (in FS and BE respectively). For example, the simplest stage in analysis (determining the "significance" of individual baseline variables in initial models for the ITT outcome, in other words adjusting for no treatment changes and thus requiring no weights) was programmed as follows:

```
* trt only model:
* fit logistic model for remission outcome
xi:   logistic   rcens_ITT_AE_sez_m_tv   treat_num   m   m_spline_*,
cluster(trial)
gen hr_trt_ITT = exp(_b[treat_num])

* baseline only models:
foreach var in age_t fi_t t_t {
      * fit logistic model for remission outcome
      xi:   logistic   rcens_ITT_AE_sez_m_tv   treat_num   `var'   m
m_spline_*, cluster(trial)
      gen hr_trt_ITT_`var' = exp(_b[treat_num])
      * determine ratio of HR with and without variables
      gen r_hr_trt_ITT_`var' = hr_trt_ITT_`var'/hr_trt_ITT
      }
foreach var in sex ni feb rels hist type eeg stype {
      * fit logistic model for remission outcome
      * NOTE: no need for weights as no time-dependent variables
included in model
      xi:   logistic   rcens_ITT_AE_sez_m_tv   treat_num   i.`var'   m
m_spline_*, cluster(trial)
      gen hr_trt_ITT_`var' = exp(_b[treat_num])
      * determine ratio of HR with and without variables
      gen r_hr_trt_ITT_`var' = hr_trt_ITT_`var'/hr_trt_ITT
```

```
}
foreach var in age_t fi_t t_t sex ni feb rels hist type eeg stype {
      di "`var'" " " r_hr_trt_ITT_`var'
}

foreach var in age_t fi_t t_t sex ni feb rels hist type eeg stype {
      gen hr_diff_trt_ITT_`var' = abs(1-r_hr_trt_ITT_`var')
      di "`var'" " " hr_diff_trt_ITT_`var'
}
```

Much more complex programming was required when it was necessary to include weights, as in the excerpt below using FS to determine "significant" variables when adjusting for all treatment changes:

```
* baseline only models:
foreach var in age_t fi_t t_t {
      * NOTE: calculating weights for future use only
      * fit logistic model for wcens outcome, separate for each
treatment (as causes of withdrawal differ between treatments) to
determine weights for logistic model for remission outcome
      xi: logistic wcens_rcens_wr0123_AE_sez_m_tv `var' m
original_m_spline4_VPS* if treat_num==0
      predict p_wr0123_VPS_trt_`var' if e(sample)
      gen p_nwr0123_VPS_trt_`var' = 1 - p_wr0123_VPS_trt_`var'
      sort trial m
      by trial: replace p_nwr0123_VPS_trt_`var' =
p_nwr0123_VPS_trt_`var'*p_nwr0123_VPS_trt_`var'[_n-1] if _n>1
      * p_nwr0123_ are weights (probability of remaining uncensored
at each time point)
      xi: logistic wcens_rcens_wr0123_AE_sez_m_tv `var' m
original_m_spline4_LTG* if treat_num==1
      predict p_wr0123_LTG_trt_`var' if e(sample)
      gen p_nwr0123_LTG_trt_`var' = 1 - p_wr0123_LTG_trt_`var'
      sort trial m
      by trial: replace p_nwr0123_LTG_trt_`var' =
p_nwr0123_LTG_trt_`var'*p_nwr0123_LTG_trt_`var'[_n-1] if _n>1
      * create weight variable including both treatment arms
      gen p_nwr0123_trt_`var' = p_nwr0123_VPS_trt_`var' if treat_num
== 0
      replace p_nwr0123_trt_`var' = p_nwr0123_LTG_trt_`var' if
treat_num == 1
      gen ip_nwr0123_trt_`var' = 1/p_nwr0123_trt_`var'
      * fit logistic model for remission outcome
      * NOTE: no need for weights as no time-dependent variables
included in model
      xi: logistic rcens_wcens_wr0123_AE_sez_m_tv treat_num `var' m
m_spline4_*, cluster(trial)
      gen hr_trt_wr0123_`var' = exp(_b[treat_num])
      * determine ratio of HR with and without variables
      gen r_hr_trt_wr0123_`var' = hr_trt_wr0123_`var'/hr_trt_wr0123
            }

foreach var in sex ni feb rels hist hist2 type eeg stype {
      * NOTE: calculating weights for future use only
```

```
      * fit logistic model for wcens outcome, separate for each
treatment (as causes of withdrawal differ between treatments) to
determine weights for logistic model for remission outcome
      xi: logistic wcens_rcens_wr0123_AE_sez_m_tv i.`var' m
original_m_spline4_VPS* if treat_num==0
      predict p_wr0123_VPS_trt_`var' if e(sample)
      gen p_nwr0123_VPS_trt_`var' = 1 - p_wr0123_VPS_trt_`var'
      sort trial m
      by trial: replace p_nwr0123_VPS_trt_`var' =
p_nwr0123_VPS_trt_`var'*p_nwr0123_VPS_trt_`var'[_n-1] if _n>1
      * p_nwr0123_ are weights (probability of remaining uncensored
at each time point)
      xi: logistic wcens_rcens_wr0123_AE_sez_m_tv i.`var' m
original_m_spline4_LTG* if treat_num==1
      predict p_wr0123_LTG_trt_`var' if e(sample)
      gen p_nwr0123_LTG_trt_`var' = 1 - p_wr0123_LTG_trt_`var'
      sort trial m
      by trial: replace p_nwr0123_LTG_trt_`var' =
p_nwr0123_LTG_trt_`var'*p_nwr0123_LTG_trt_`var'[_n-1] if _n>1
      * create weight variable including both treatment arms
      gen p_nwr0123_trt_`var' = p_nwr0123_VPS_trt_`var' if treat_num
== 0
      replace p_nwr0123_trt_`var' = p_nwr0123_LTG_trt_`var' if
treat_num == 1
      gen ip_nwr0123_trt_`var' = 1/p_nwr0123_trt_`var'
      * fit logistic model for remission outcome
      * NOTE: no need for weights as no time-dependent variables
included in model
      xi: logistic rcens_wcens_wr0123_AE_sez_m_tv treat_num i.`var'
m m_spline4_*, cluster(trial)
      gen hr_trt_wr0123_`var' = exp(_b[treat_num])
      * determine ratio of HR with and without variables
      gen r_hr_trt_wr0123_`var' = hr_trt_wr0123_`var'/hr_trt_wr0123
}
foreach var in age_t fi_t t_t sex ni feb rels hist hist2 type eeg
stype {
      di "`var'" " " r_hr_trt_wr0123_`var'
}

* tvcs: AE/seizure variables
* need to stabilise weights: rename p_nwr0123_ p_nwr0123_denom;
refit without tvcs: rename p_nwr0123_ p_nwr0123_num; gen
stab_wgt_AE/sez = p_nwr0123_num/p_nwr0123_denom

* fit logistic model for wcens outcome, separate for each treatment
(as causes of withdrawal differ between treatments) to determine
weights for logistic model for remission outcome
xi: logistic wcens_rcens_wr0123_AE_sez_m_tv cum_sez_t m
original_m_spline4_VPS* if treat_num==0
predict p_wr0123_VPS_trt_sez if e(sample)
gen p_nwr0123_VPS_trt_sez = 1 - p_wr0123_VPS_trt_sez
sort trial m
by trial: replace p_nwr0123_VPS_trt_sez =
p_nwr0123_VPS_trt_sez*p_nwr0123_VPS_trt_sez[_n-1] if _n>1
* p_nwr0123_ are weights (probability of remaining uncensored at
each time point) but need to be stabilised
xi: logistic wcens_rcens_wr0123_AE_sez_m_tv cum_sez_t m
original_m_spline4_LTG* if treat_num==1
predict p_wr0123_LTG_trt_sez if e(sample)
```

```
gen p_nwr0123_LTG_trt_sez = 1 - p_wr0123_LTG_trt_sez
sort trial m
by trial: replace p_nwr0123_LTG_trt_sez =
p_nwr0123_LTG_trt_sez*p_nwr0123_LTG_trt_sez[_n-1] if _n>1
* create weight variable including both treatment arms
gen p_nwr0123_trt_sez = p_nwr0123_VPS_trt_sez if treat_num == 0
replace p_nwr0123_trt_sez = p_nwr0123_LTG_trt_sez if treat_num == 1
* create stabilised weight: use for numerator the weights from a
model without same baseline vars but no tvcs
gen ip_nwr0123_trt_sez_stab = p_nwr0123_trt/p_nwr0123_trt_sez
* fit weighted logistic model for remission outcome
logistic rcens_wcens_wr0123_AE_sez_m_tv treat_num cum_sez_t m
m_spline4_* [pw=ip_nwr0123_trt_sez_stab], cluster(trial)
gen hr_trt_wr0123_sez = exp(_b[treat_num])
* determine ratio of HR with and without variables
gen r_hr_trt_wr0123_sez = hr_trt_wr0123_sez/hr_trt_wr0123
di r_hr_trt_wr0123_sez

* tvcs: AE/seizure variables
* need to stabilise weights: rename p_nwr0123_ p_nwr0123_denom;
refit without tvcs: rename p_nwr0123_ p_nwr0123_num; gen
stab_wgt_AE/sez = p_nwr0123_num/p_nwr0123_denom

* fit logistic model for wcens outcome, separate for each treatment
(as causes of withdrawal differ between treatments) to determine
weights for logistic model for remission outcome
xi: logistic wcens_rcens_wr0123_AE_sez_m_tv cum_AE_t m
original_m_spline4_VPS* if treat_num==0
predict p_wr0123_VPS_trt_AE if e(sample)
gen p_nwr0123_VPS_trt_AE = 1 - p_wr0123_VPS_trt_AE
sort trial m
by trial: replace p_nwr0123_VPS_trt_AE =
p_nwr0123_VPS_trt_AE*p_nwr0123_VPS_trt_AE[_n-1] if _n>1
* p_nwr0123_ are weights (probability of remaining uncensored at
each time point) but need to be stabilised
xi: logistic wcens_rcens_wr0123_AE_sez_m_tv cum_AE_t m
original_m_spline4_LTG* if treat_num==1
predict p_wr0123_LTG_trt_AE if e(sample)
gen p_nwr0123_LTG_trt_AE = 1 - p_wr0123_LTG_trt_AE
sort trial m
by trial: replace p_nwr0123_LTG_trt_AE =
p_nwr0123_LTG_trt_AE*p_nwr0123_LTG_trt_AE[_n-1] if _n>1
* create weight variable including both treatment arms
gen p_nwr0123_trt_AE = p_nwr0123_VPS_trt_AE if treat_num == 0
replace p_nwr0123_trt_AE = p_nwr0123_LTG_trt_AE if treat_num == 1
* create stabilised weight: use for numerator the weights from a
model without same baseline vars but no tvcs
gen ip_nwr0123_trt_AE_stab = p_nwr0123_trt/p_nwr0123_trt_AE
* fit weighted logistic model for remission outcome
xi: logistic rcens_wcens_wr0123_AE_sez_m_tv treat_num cum_AE_t m
m_spline4_* [pw=ip_nwr0123_trt_AE_stab], cluster(trial)
gen hr_trt_wr0123_AE = exp(_b[treat_num])
* determine ratio of HR with and without variables
gen r_hr_trt_wr0123_AE = hr_trt_wr0123_AE/hr_trt_wr0123
di r_hr_trt_wr0123_AE

* fit logistic model for wcens outcome, separate for each treatment
(as causes of withdrawal differ between treatments) to determine
weights for logistic model for remission outcome
```

```
xi: logistic wcens_rcens_wr0123_AE_sez_m_tv dose_m m
original_m_spline4_VPS* if treat_num==0
predict p_wr0123_VPS_trt_dose if e(sample)
gen p_nwr0123_VPS_trt_dose = 1 - p_wr0123_VPS_trt_dose
sort trial m
by trial: replace p_nwr0123_VPS_trt_dose =
p_nwr0123_VPS_trt_dose*p_nwr0123_VPS_trt_dose[_n-1] if _n>1
* p_nwr0123_ are weights (probability of remaining uncensored at
each time point) but need to be stabilised
xi: logistic wcens_rcens_wr0123_AE_sez_m_tv dose_m m
original_m_spline4_LTG* if treat_num==1
predict p_wr0123_LTG_trt_dose if e(sample)
gen p_nwr0123_LTG_trt_dose = 1 - p_wr0123_LTG_trt_dose
sort trial m
by trial: replace p_nwr0123_LTG_trt_dose =
p_nwr0123_LTG_trt_dose*p_nwr0123_LTG_trt_dose[_n-1] if _n>1
* create weight variable including both treatment arms
gen p_nwr0123_trt_dose = p_nwr0123_VPS_trt_dose if treat_num == 0
replace p_nwr0123_trt_dose = p_nwr0123_LTG_trt_dose if treat_num ==
1
* create stabilised weight: use for numerator the weights from a
model without same baseline vars but no tvcs
gen ip_nwr0123_trt_dose_stab = p_nwr0123_trt/p_nwr0123_trt_dose
* fit weighted logistic model for remission outcome
xi: logistic rcens_wcens_wr0123_AE_sez_m_tv treat_num m m_spline4_*
[pw=ip_nwr0123_trt_dose_stab], cluster(trial)
gen hr_trt_wr0123_dose = exp(_b[treat_num])
* determine ratio of HR with and without variables
gen r_hr_trt_wr0123_dose = hr_trt_wr0123_dose/hr_trt_wr0123
di r_hr_trt_wr0123_dose

foreach var in age_t fi_t t_t sex ni feb rels hist hist2 type eeg
stype AE sez dose {
      gen hr_diff_trt_wr0123_`var' = abs(1-r_hr_trt_wr0123_`var')
      di "`var'" " " hr_diff_trt_wr0123_`var'
}
```

These commands demonstrate the need for careful labelling in order to facilitate easy iterative programming. This is especially true in the case of BE, as demonstrated in the excerpt below. This extract considers only the effect of baseline variable age, adjusting for all treatment changes:

```
* only baseline variables included (for stabilised weights)
* predict probability of treatment failure
(wcens_rcens_wr0123_AE_sez_m_tv)
xi: logistic wcens_rcens_wr0123_AE_sez_m_tv age_t fi_t t_t i.sex
i.ni i.feb i.rels i.hist i.eeg i.stype m original_m_spline_LTG* if
treat_num==0
predict p_wr0123_L_bl if e(sample)
gen p_nwr0123_L_bl = 1 - p_wr0123_L_bl
sort trial m
```

```
by trial: replace p_nwr0123_L_bl = p_nwr0123_L_bl*p_nwr0123_L_bl[_n-
1] if _n>1
* p_nwr0123_ are weights (probability of remaining uncensored at
each time point)
xi: logistic wcens_rcens_wr0123_AE_sez_m_tv age_t fi_t t_t i.sex
i.ni i.feb i.rels i.hist i.eeg i.stype m original_m_spline_VPS* if
treat_num==1
predict p_wr0123_V_bl if e(sample)
gen p_nwr0123_V_bl = 1 - p_wr0123_V_bl
sort trial m
by trial: replace p_nwr0123_V_bl = p_nwr0123_V_bl*p_nwr0123_V_bl[_n-
1] if _n>1
* create weight variable including both treatment arms
gen p_nwr0123_bl = p_nwr0123_L_bl if treat_num == 0
replace p_nwr0123_bl = p_nwr0123_V_bl if treat_num == 1

* all variables included
* predict probability of treatment failure
(wcens_rcens_wr0123_AE_sez_m_tv)
xi: logistic wcens_rcens_wr0123_AE_sez_m_tv age_t fi_t t_t i.sex
i.ni i.feb i.rels i.hist i.eeg i.stype cum_sez_t cum_AE_t m
original_m_spline_LTG* if treat_num==0
predict p_wr0123_L_all if e(sample)
gen p_nwr0123_L_all = 1 - p_wr0123_L_all
sort trial m
by trial: replace p_nwr0123_L_all =
p_nwr0123_L_all*p_nwr0123_L_all[_n-1] if _n>1
* p_nwr0123_ are weights (probability of remaining uncensored at
each time point)
xi: logistic wcens_rcens_wr0123_AE_sez_m_tv age_t fi_t t_t i.sex
i.ni i.feb i.rels i.hist i.eeg i.stype cum_sez_t cum_AE_t m
original_m_spline_VPS* if treat_num==1
predict p_wr0123_V_all if e(sample)
gen p_nwr0123_V_all = 1 - p_wr0123_V_all
sort trial m
by trial: replace p_nwr0123_V_all =
p_nwr0123_V_all*p_nwr0123_V_all[_n-1] if _n>1
* create weight variable including both treatment arms
gen p_nwr0123_all = p_nwr0123_L_all if treat_num == 0
replace p_nwr0123_all = p_nwr0123_V_all if treat_num == 1
* create stabilised weight: use for numerator the weights from a
model without same baseline vars but no tvcs
gen p_nwr0123_all_stab = p_nwr0123_bl/p_nwr0123_all
* fit weighted logistic model for remission outcome
xi: logistic rcens_wcens_wr0123_AE_sez_m_tv treat_num age_t fi_t t_t
i.sex i.ni i.feb i.rels i.hist i.eeg i.stype cum_sez_t cum_AE_t m
m_spline_* [pw=p_nwr0123_all_stab], cluster(trial)
gen hr_wr0123_all = exp(_b[treat_num])

* remove each variable in turn

* remove age_t
* only baseline variables included (for stabilised weights)
excluding age_t
* predict probability of treatment failure
(wcens_rcens_wr0123_AE_sez_m_tv)
xi: logistic wcens_rcens_wr0123_AE_sez_m_tv fi_t t_t i.sex i.ni
i.feb i.rels i.hist i.eeg i.stype m original_m_spline_LTG* if
treat_num==0
```

```
predict p_wr0123_L_bl_eag if e(sample)
gen p_nwr0123_L_bl_eag = 1 - p_wr0123_L_bl_eag
sort trial m
by trial: replace p_nwr0123_L_bl_eag =
p_nwr0123_L_bl_eag*p_nwr0123_L_bl_eag[_n-1] if _n>1
* p_nwr0123_ are weights (probability of remaining uncensored at
each time point)
xi: logistic wcens_rcens_wr0123_AE_sez_m_tv fi_t t_t i.sex i.ni
i.feb i.rels i.hist i.eeg i.stype m original_m_spline_VPS* if
treat_num==1
predict p_wr0123_V_bl_eag if e(sample)
gen p_nwr0123_V_bl_eag = 1 - p_wr0123_V_bl_eag
sort trial m
by trial: replace p_nwr0123_V_bl_eag =
p_nwr0123_V_bl_eag*p_nwr0123_V_bl_eag[_n-1] if _n>1
* create weight variable including both treatment arms
gen p_nwr0123_bl_eag = p_nwr0123_L_bl_eag if treat_num == 0
replace p_nwr0123_bl_eag = p_nwr0123_V_bl_eag if treat_num == 1

* predict probability of treatment failure
(wcens_rcens_wr0123_AE_sez_m_tv)
xi: logistic wcens_rcens_wr0123_AE_sez_m_tv fi_t t_t i.sex i.ni
i.feb i.rels i.hist i.eeg i.stype cum_sez_t cum_AE_t m
original_m_spline_LTG* if treat_num==0
predict p_wr0123_L_all_eag if e(sample)
gen p_nwr0123_L_all_eag = 1 - p_wr0123_L_all_eag
sort trial m
by trial: replace p_nwr0123_L_all_eag =
p_nwr0123_L_all_eag*p_nwr0123_L_all_eag[_n-1] if _n>1
* p_nwr0123_ are weights (probability of remaining uncensored at
each time point)
xi: logistic wcens_rcens_wr0123_AE_sez_m_tv fi_t t_t i.sex i.ni
i.feb i.rels i.hist i.eeg i.stype cum_sez_t cum_AE_t m
original_m_spline_VPS* if treat_num==1
predict p_wr0123_V_all_eag if e(sample)
gen p_nwr0123_V_all_eag = 1 - p_wr0123_V_all_eag
sort trial m
by trial: replace p_nwr0123_V_all_eag =
p_nwr0123_V_all_eag*p_nwr0123_V_all_eag[_n-1] if _n>1
* create weight variable including both treatment arms
gen p_nwr0123_all_eag = p_nwr0123_L_all_eag if treat_num == 0
replace p_nwr0123_all_eag = p_nwr0123_V_all_eag if treat_num == 1
* create stabilised weight: use for numerator the weights from a
model without same baseline vars but no tvcs
gen ip_nwr0123_all_eag_stab = p_nwr0123_bl_eag/p_nwr0123_all_eag
* fit weighted logistic model for remission outcome
xi: logistic rcens_wcens_wr0123_AE_sez_m_tv treat_num fi_t t_t i.sex
i.ni i.feb i.rels i.hist i.eeg i.stype cum_sez_t cum_AE_t m
m_spline_* [pw=ip_nwr0123_all_eag_stab], cluster(trial)
gen hr_wr0123_all_eag = exp(_b[treat_num])
* determine ratio of HR with and without variables
gen r_hr_wr0123_all_eag = hr_wr0123_all_eag/hr_wr0123_all
di r_hr_wr0123_all_eag
```

9.4.Key considerations as part of IPCW modelling process

There were a number of further modelling complications which arose during the somewhat evolutionary process of carrying out this sequential investigation. Thus this analysis demonstrates how an initial causal analysis plan may need to change when practical problems or limitations arise while undertaking such analyses due to features of the data or modelling techniques. The first issue related to missing TVC data.

9.4.1. Missing seizure dates

Forty-two patients (at 58 visits) had missing values for their total number of seizures experienced between current and previous visit, despite recoded valid dates of first and most recent seizures entered for the corresponding visit. This missing data field suggests that the clinician was not able to estimate the number of seizures experienced by the patient since the previous visit but that seizures did occur between the given dates. Given the potential importance of cumulative seizure count as a TVC, it was necessary to decide how to address this missing data problem.

There were a number of options available, the simplest being to simply exclude these patients from the analysis, or to assume that in fact no seizures had occurred. Alternatively, it may seem more plausible to assume that two seizures had in fact occurred (one each on the first and last dates of seizures given) or just one seizure (if the first and last dates of seizures were the same). More sophisticated imputation methods could have been employed, for example by estimating the missing total number of seizures based on the patient's seizure rates for previous (and potentially future) visits, perhaps using only those visits at which the patient was taking the randomised drug. This approach could be extended further using multiple imputation

regression techniques to more accurately predict missing seizures rates based on factors such as dose, treatment and patient characteristics, potentially considering the relative importance of within- compared to between-patient variability, the distribution of seizure counts over time and the potential for carryover treatment effects affecting seizure rates even after the patient stopped taking their randomised treatment.

Each of these imputation approaches were deemed either to be unrealistically simplistic or overly complicated in this setting, relying on assumptions that could not be easily or accurately verified. As such, it was decided that the approach used in the original analysis (194) would be employed here, namely that patients with missing seizure counts would simply be censored at the date of the first seizure recorded for this interval, as this was the last point at which the cumulative seizure count was known. The only exception to this was when the dates of first and last seizure were equal, and the patient's seizure history suggested that the patient typically experienced very few seizures; in this case, it was assumed that the patient experienced only one seizure on that date (so that a count of one was assumed for the missing total number of seizures).

This censoring (and simple imputation) resulted in eight patients' remission (or censoring) time being brought forward, changing the event status of three (and one) patients who had achieved remission (or a treatment change) but were instead censored. It is acknowledged that this censoring could introduce bias if the unknown seizure counts were correlated with the frequency of seizures. However, given that seizure intervals varied considerably among those patients with missing seizure counts, this bias was not overtly obvious.

9.4.2. Extreme weights

The other major problem that was encountered during this modelling exercise was that of extreme weights. In order to address this issue, it was necessary to understand the underlying reasons; on inspection of the data, it became obvious that this problem was related to two features of the data.

As mentioned in section 9.2.2, a potential cause for extreme weights is when an incorrect functional form for covariates is specified in the PLR model, thus implying an incorrect relationship between each covariate and the outcome. In particular, if the model wrongly assumes linearity (in other words, the untransformed covariates are included in the model), patients with extreme values of TVCs will be disproportionately influential. Thus, alternative models with log or inverse links (that asymptote rather than increase linearly as the TVC increases) were explored.

Martingale residuals are useful in determining the most appropriate functional form of covariates to be included in the model. As such, in order to decide on the best model fit between linear, log and inverse fit of continuous variables, a locally weighted smoothing (lowess) curve of the martingale residuals from the (constant only) Cox model was plotted against each transformation of each TVC in turn, namely untransformed values (X), logged values ($\log(1+X)$) and inverse values ($1/(1+X)$). An appropriate transformation would exhibit an approximately linear lowess smooth; thus, on visual inspection of these smoothing curves, it became apparent that the logarithmic transformation was appropriate for the TVC seizure count and three baseline continuous variables (interval between first ever seizure and randomisation, total number of tonic clonic seizures prior to randomisation and age).

Although this transformation helped to reduce the problem somewhat, extreme weights occurred, though to a lesser extent, particularly as more covariates were added to the model. This disproportionate influence of certain individuals had a large effect on estimated treatment effects and hence the selection process as a whole. For example, censoring one patient's follow up 15 months early made a large difference to the treatment effect (and range of weights), which changed from 0.05 (0.6, 2888.7) to 0.35 (0.6, 27.0).

Secondly, as expected, these extreme weights tended to occur in patients with large values of TVCs (particularly seizure counts and dose) late on in follow up, whose data were overly influential because there remained few other similar patients on which to base weight estimation. For example, one patient's prolonged follow up data was leading to extreme weights because of their high treatment dose received late on in follow up, when few patient-months of follow up remained in their treatment group. Thus, the available options to counter the extreme weights that continued to dominate the analysis were either to truncate continuous variables at the 95th (instead of just the 99th) centile; to censor patient's follow up when their weights exceeded a predefined cut-off level; or to categorise continuous variables in order to reduce the influence of outlying values. The last option is not ideal, as it would lead to a great loss of information. Thus, firstly the effect of truncation at the 95th centile was investigated.

Extreme weights still occurred in a small number of cases despite truncation at the 95th (rather than 99th) centile of continuous variables. Thus, although it is acknowledged that censoring patient's follow up when their weights become problematically high is likely to introduce a certain amount of bias, it is also recognised that simply ignoring the problem of extreme weights is likely to lead to a grossly incorrect model, given the

influence that extreme weights have on the resultant treatment effects. Thus it was decided that an individual's follow up would be censored when their weight exceeded 30; this cut-off level was approximately equal to one tenth of each treatment group size, a level at which it seemed reasonable to restrict a single individual's contribution. Although such censoring may introduce bias, it is worth noting that this censoring affected only approximately 0.01% of patient-time intervals in each analysis.

9.4.3. Revised plan

Thus, following these model considerations and investigations, the FS and BE selection procedures were carried out according to a slightly revised analysis plan. Cox regression was compared to PLR (using three and five knots); all continuous variables were analysed following logarithmic transformation and truncation at the 95% centile, and certain categorical variables were collapsed to remove the problem of small event numbers within combinations of factors. Patients' follow up was censored if they had missing seizure count data or if their weight exceeded 30 for a given time interval (month, fortnight or week). Variables were included in the initial pool of baseline and time-varying covariates only if their inclusion in the treatment-only WD model altered the WO treatment effect by at least 1%.

9.5.Summary of initial results

In order to determine which of the baseline and TVCs should be adjusted for in this IPCW analysis (thus addressing the first research question listed in 9.3.7), the Cox and PLR IPCW models were completed for the sequence of treatment change scenarios (none (ITT), ISC alone, ISC/personal choice and all changes) for each of time intervals (weekly, fortnightly and monthly) used for the creation of time-updated covariates

(seizures, AEs and dose). Both forward and backward selection (based on an initial pool of explanatory variables which alter the treatment hazard (or odds) ratio by at least 1%) were used to determine final models which included variables affecting the HR (or OR) by at least 2%, 5% or 10%. The corresponding bootstrapped CIs were also calculated, as the usual SEs for the Cox regression could not be calculated accurately due to time-varying weights within individual. These results have been summarised in forest plots of treatment effects and BS CIs (Figure 11 to Figure 14) as well as tables (Table 49 to Table 54) to allow easy comparison across treatment change scenarios and model types (see Appendix B).

Generally, the results show that baseline covariates were not usually selected into the WD model, but time-varying covariates are often selected, most commonly seizures but also dose and occasionally AE counts.

The forward and backward selection procedures were not always consistent in their selection of different variables; this is not surprising, however, given that the selection criterion was defined in terms of the change in treatment effect relative to that from the previous model, which differed when using FS (which, for example, starts with an initial treatment-only model) compared to BE (which begins with the full model containing all potential variables).

The PLR bootstrap confidence intervals (CIs) were generally wider than the usual CIs, due to the added uncertainty associated with bootstrapping. Indeed, initially, the bootstrap procedure for the PLR models resulted in very high sample failure rates due to lack of convergence (using both three and five knots), especially for the month spline variable. Investigation into the problems relating to knot positions and extreme weights (discussed above) helped to address this problem to a certain extent; however, this

model instability continued to affect the PLR five knot models in particular, leading to a continued problem of non-convergence of some bootstrap samples, extreme estimated covariate coefficients (typically $>10^7$) and artifactual peaks in the underlying baseline odds plots towards the end of follow up, suggesting instability of splines beyond the 95th centile. Further evidence of overfitting was provided by the coefficients for the spline variables (which were often very close to zero) and low EPV ratios (especially for the VPS treatment change model) as more variables were added to the models, suggesting that it may be more appropriate to use fewer knots for the PLR spline variables.

One potential consequence of this overfitting was that the number of variables selected into each PLR model was consistently smaller than for the corresponding Cox model, perhaps due to lack of power. These problems suggest that it may have been more appropriate to fit the model with fewer knots, potentially even with no knots (which equates to an underlying linear function of time).

In general, more covariates were selected into the model as more treatment changes were adjusted for (from none (ITT) to ISC alone to ISC/choice to all treatment changes), because of the increased power in the treatment change model as more events are available for model estimation. The number of covariates selected into the model also tended to increase as the frequency of follow up information increased (in other words, as the length of time intervals decreased, from month to fortnight to week), again because the amount of information available (from time-varying covariates) is increased.

As more variables were selected into the model, the treatment effect also tended to move away from one (towards zero, in other words further in favour of the control arm

VPS). However, given the fact that the number of variables selected into the model increased as more treatment changes were adjusted for, this variable selection process did not clearly reveal whether the reduction of the treatment effect (away from one) was due to the increased number of (and type of) treatment changes being considered or because of the increased number of variables being adjusted for.

9.6.Conclusions

As such, although this investigation into the selection of variables across the whole range of model features (model type (including the effect of bootstrapping on CIs), time interval, variable selection threshold and procedure, and the treatment changes accounted for) provides guidance on which (of the many) baseline and TVCs should be adjusted for in order to adequately satisfy the assumption of NUC, there are a number of questions of interest which remain to be answered, most notably whether the treatment effect is affected most by model type, reasons for treatment change or time interval.

Thus, in order to allow direct comparisons to be made between each scenario and thus an assessment of the independent effect of each of these sources of variation on the resultant treatment effect, it became apparent that it would be necessary to fit a standard set of models which adjust for the *same* covariates for each of the model type/treatment change/time interval combinations.

To this end, it was necessary to choose a standard set of covariates in order to investigate the influence of other factors (choice of model type and time interval, selection threshold and procedure, and the validity of normal or bootstrapped CIs for the PLR). Given that baseline covariates were not often selected into the ITT version of

each model, it seemed sensible to consider including a variety of combinations of only TVCs in the IPCW models. The results of this variable selection procedure across different modelling scenarios demonstrated that, on the whole, seizure count was selected most often into the IPCW models, followed by dose and then AEs. Thus it seemed appropriate that the following sequence of models should be considered: adjusting for time alone; adjusting for time and seizures; adjusting for time, seizures and dose; and finally adjusting for time, seizures, dose and AEs.

Furthermore, given the evidence of overfitting with three and five knots for the PLR spline variables (with spline coefficients close to zero and frequent non-convergence of bootstrapped CIs), it was decided that a series of PLR models with only a linear time variable (in other words, without any splines) should be compared to those with three and five knots, in order to assess any subsequent effect on treatment effect estimates and evidence of overfitting.

This proposed analysis using a standard set of covariates, the choice of which has been informed by the general trends of variable selection observed in this chapter's investigational analysis, will be described in the next chapter. Furthermore, the use of a standard set of IPCW models will allow direct comparison with the corresponding RPSFTM analysis to determine the effect on inferences drawn using these contrasting methods.

10. Comparison of IPCW and RPSFTM

In Chapter 9, IPCW models were fit using forward and backward selection procedures with a range of selection thresholds in order to determine which of the numerous available covariates would most usefully be adjusted for in the IPCW analysis of the VPS vs LTG comparison of SANAD arm B. The results of this investigation suggested that the three TVCs (dose and cumulative seizure and AE counts) were all (to varying degrees) important confounders of the relationship between treatment received and remission, while the baseline covariates were less important.

Thus, in this chapter, the IPCW and RPSFTM will be compared in the context of the SANAD arm B, using this standard set of TVCs determined by investigations described

in Chapter 9. The use of a standard set of covariates allows the impact of various statistical factors on IPCW modelling to be investigated, including the choice of regression model (PLR or Cox), time interval for determining time-varying covariates (month, fortnight or week) and competing reasons for treatment changes. PLR modelling required further consideration of the method of confidence interval (CI) estimation (normal versus bootstrapped CIs) and number of knots for the spline time variable.

10.1. Final standardised analysis

For this standardised analysis, the IPCW analyses were conducted using Cox and PLR models, adjusting for an increasing number of TVCs (none; seizures alone; seizures and dose; seizures, dose and AEs) for each set of treatment changes (none (mimicking ITT); ISC only; ISC or choice; all treatment changes) and each time interval (month; fortnight; week). Normal and bootstrapped CIs were compared for the PLR models, fitted using zero, three and five knots for the time spline variable in order to evaluate the likelihood of overfitting in these models. This comparison allows a clear assessment to be made of the relative importance and influence of each of these factors on the treatment effect.

Furthermore, the RPSFTM introduced in Chapter 8 was adapted in order to allow direct comparison with these IPCW models, thus only adjusting for those switches that relate to each particular treatment change scenario; therefore a series of RPSFTMs (adjusting only for "true" switches, as defined in Chapter 8) were fit according to the hierarchical structure of the causal research questions defined for the IPCW models. Note that it is possible to adjust the RPSFTM for any baseline factors that may be imbalanced between groups; however, given that no such covariates were identified as

particularly important in the IPCW selection procedure, this was not necessary. In contrast, it is inappropriate to adjust the RPSFTM for the TVCs that were selected for inclusion in the RPSFTM, as such adjustment would introduce selection bias.

10.1.1. Causal questions of interest

When faced with modelling decisions, it is important to keep focused on the clinical questions of interest, which in this case related to the treatment effect of LTG compared to VPS allowing for various reasons for treatment change. In addition to facilitating the comparison between the RPSFTM and IPCW in this trial, the use of a standard set of covariates in the IPCW models helps to answer the following (unanswered) questions from Chapter 9 regarding the impact of model type, reasons for treatment change or time interval on IPCW treatment effects:

1. Do IPCW treatment effect estimates differ between **Cox and PLR** (and according to **number of knots** in PLR)?
2. Do IPCW treatment effects change as more **TVCs** are adjusted for in the weight estimation model?
3. How does the choice of **time interval** impact on IPCW treatment effects? In particular, does the use of smaller intervals (weekly intervals) for IPCW analysis make Cox and PLR models more similar?
4. What are the general differences in IPCW treatment effects when different **treatment changes** are considered?
5. How do **IPCW and RPSFTM** treatment effects compare for each of these treatment change scenarios?

Treatment change scenarios

Table 41 demonstrates the practical differences between each of the modelling scenarios, in particular according to which treatment changes are adjusted for (which relates to the causal question of interest) and which TVCs are included in the WD model (which relates to the NUC assumption). When no treatment changes are accounted for, the results across all sets of TVC models are identical, given that no weights are generated. Artificial censoring is introduced at the time of a patient's first treatment change (F_i) if the reason for this change corresponds to the treatment change scenario under consideration; therefore, for each treatment change scenario, models vary only according to the number of TVCs that are used to generate time-varying weights in the WD models (in other words, according to the underlying NUC assumption).

Bootstrap CIs compared with normal CIs

Furthermore, these standardised analyses allow comparisons to be made between normal and bootstrapped CIs, as outlined in Table 42. As described in Chapter 5, normal SEs estimated from the Cox model are invalid, given that no allowance can be made for patient identity when applying their time-varying weights. Although such within-patient time-varying weights do not present a problem for PLR models, the actual process of estimating weights is known to artificially reduce the estimated SEs, suggesting the need for bootstrapped CI estimation for PLR models also.

However, in scenarios requiring no artificial censoring (and hence no weights), namely those ITT scenarios where no treatment changes are adjusted for or for the treatment only models which do not adjust for any TVCs, the standardised analyses described in

this chapter allow direct assessment of the effect of bootstrapping compared to normal CI estimation methods. Without the complication of within-patient time-varying weights, the normal CIs for these Cox models are valid, thus providing an opportunity to directly compare the usual and bootstrapped CIs, such that any general observed differences between CIs can be attributed to the impact of bootstrapping (in order to verify the (generally accepted) validity of bootstrapping in this context).

Furthermore, comparison between normal and bootstrapped CIs from PLR models which *do* require weight estimation will provide evidence of whether the process of estimating weights impacts on CI estimation. Although it is recognised that any such differences between these two CIs will reflect the impact of bootstrapping as well as weight estimation, the general effect of bootstrapping (alone) on PLR models is evident from the comparison between CIs for models that do *not* require weights (namely the ITT (or treatment only) models). Therefore, the difference in general trends seen between these two forms of CIs in models *with* and *without* weights give an indication of the effect of weight estimation on CI estimation over and above that of bootstrapping.

These models were also assessed for evidence of overfitting, by considering EPV ratios for each model, as well as any non-convergence of bootstrapped samples or extreme model coefficients (as displayed in Table 42).

EPV ratios

As discussed in Chapter 9, a restricted cubic spline variable with k knots introduces $(k-1)$ extra terms into the model; however given that a linear term is also required to accompany the spline variable, the total number of estimated coefficients required to model the underlying time effect for a restricted spline variable with k knots is equal to

k. A PLR model that includes no spline variables simply assumes a linear effect of treatment (and is indeed equivalent to a restricted cubic spline with one knot, given that such a spline restricts the tails (in other words, the whole spline) to be linear) and thus requires estimation of a single coefficient (for the linear term). The EPV ratios are interpreted in light of the recommendation that at least ten events (treatment change or remission, as appropriate) are required per variable included in the model, in order to ensure model stability (180).

Table 41 Artificial censoring and weighting, according to treatment changes (determined by causal question) and TVCs in WD model

	TVCs adjusted for in analysis (relating to underlying NUC)			
	Treatment only	Treatment + seizures	Treatment, seizures + dose	Treatment, seizures, dose + AEs
Treatment change scenario				
None (ITT)	**No artificial censoring, therefore no weights required**			
ISC	**Artificial censoring at F_i (if first treatment change is due to ISC)**			
	No weights (as no TVCs are accounted for)	Weights allowing for seizure count	Weights allowing for seizure count and dose	Weights allowing for seizure count, dose, AEs
ISC/choice	**Artificial censoring at F_i (if first treatment change is due to ISC or personal choice)**			
	No weights (as no TVCs are accounted for)	Weights allowing for seizure count	Weights allowing for seizure count and dose	Weights allowing for seizure count, dose, AEs
All	**Artificial censoring at F_i**			
	No weights (as no TVCs are accounted for)	Weights allowing for seizure count	Weights allowing for seizure count and dose	Weights allowing for seizure count, dose, AEs

Table 42 Assessment of BS results, according to treatment changes and TVCs in WD model

	Treatment alone	Treatment + seizures	Treatment, seizures + dose	Treatment, seizures, dose + AEs
ITT	No treatment changes (and hence no weights used) in ITT analysis: normal CI is valid; thus assess validity of BS by comparison with normal CI*			
ISC	No TVCs and hence no weights: therefore normal CI is valid; assess validity of BS by comparison with normal CI*	TVCs and treatment changes are both present; hence weights due to treatment changes are necessary: 1) Cox: a. Normal CI is invalid (with no account made for multiple (time interval) observations for each patient) 2) PLR (3 or 5 knots): a. Normal CI may be affected by weight estimation process: compare normal/BS CI to assess effect of weighting on SE (taking into account validity of BS from *) b. Overfitting of PLR models (especially with 5 knots) may lead to extreme coefficients and non-convergence of BS samples i. Typically very small coefficients (~0, i.e. OR~1) for spline variables, demonstrating little variation over time, but occasionally very large (>10^7) coefficients for 5th spline (leading to non-convergence of the bootstrapping model) indicating instability of model when fitting peak at the end of the range of time intervals. ii. Thus investigate PLR models with no spline variables (i.e. linear time effect only)		
ISC/choice				
All				

10.2. Results

In order to interpret this relatively complex analysis, it is helpful to focus, in turn, on each of the main research questions of interest listed in 10.1.1, while bearing in mind that the main purpose of the modelling exercise is ultimately to determine the treatment effect according to different reasons for treatment change.

1. Do IPCW treatment effect estimates differ between **Cox and PLR** (and according to **number of knots** in PLR)?

First, the results of the PLR and Cox models need to be assessed in terms of the effect of time interval, spline knots (for PLR models) and bootstrapping on model estimation and stability.

The results in Table 58 to Table 59 (in Appendix C) demonstrate that spline variables (with 3 or 5 knots) cause low EPV, especially for withdrawal models. These low EPV ratios lead to overfitting of the treatment change models, evident by extreme coefficients for the tails of the spline distribution and non-convergence of the bootstrapped samples (indicated by subscripts in Table 60 to Table 63 (in Appendix C) as the number of bootstrapped samples (out of a total of 200) for which the model failed to converge). This is a particular problem for models where the number of observations on which to base the PLR is low (in other words, when using the month interval). This problem worsened as the number of events (treatment changes) on which to base this estimation fell (in other words, with more restricted treatment change scenarios) and as more variables were added to the treatment change model (either in the form of TVCs or due to the increasing number of knots in the spline variables). Thus non-convergence is a particular problem for the PLR models with five spline

knots assessing restricted treatment change scenarios with a monthly time interval, where more than half of the bootstrapped samples fail to reach convergence.

However, despite this evidence of overfitting with increasing number of spline knots, the variation in terms of the treatment effect (OR) according to the number of spline knots is minimal. Interestingly, the difference between PLR treatment effects with different numbers of spline knots is greater for scenarios without adjustment for any treatment changes (ITT); otherwise the variation between OR estimates (between models with zero, three and five knots) tends to be at most 0.02.

In contrast, the difference between treatment effects is far greater between model types (in other words between PLR and Cox models), as the Cox model seems to be more susceptible to changes in experimental factors (such as time interval and choice of TVCs).

CI estimation

The independent effects on CI estimation of **weight estimation**, **bootstrapping** and **failing to account for patient identity** (when Cox estimation involves within-patient time-varying weights) need to be investigated by considering three separate comparisons, summarised in Table 42.

First, the comparison between usual and BS CIs for the models not involving any weight estimation (in other words, when no treatment changes or TVCs are adjusted for) will give an indication of the effect of **bootstrapping** on CI estimation. The PLR models without weights demonstrate the expected result that BS CIs tend to be wider than normal CIs (though not usually by more than 0.1) with the usual CI contained entirely within the BS CI (in other words, the upper (lower) limit from BS estimation

tends to be higher (lower) than the corresponding usual CI limit), but this is not true in all cases. In contrast the BS CIs from the corresponding Cox models (without weight estimation) tend to be *narrower* than the normal CIs.

The effect of **failing to account for patient identity in weighted Cox estimation** is evident from the comparison between the normal and BS CIs for Cox models which adjust for TVCs (and which hence involve weights). Standard CIs are much narrower than BS CIs in these cases, because they fail to account for correlation between within-patient time-varying weights. However, given the above result (that BS CIs tend in fact to be *narrower* than the corresponding (valid) normal CIs), this observation demonstrates just how misleading these invalid CIs are.

Finally, the effect of **weight estimation** can be observed by comparing normal and BS CIs obtained for those PLR models involving weight estimation. The difference between these CIs comprises the effect of both weight estimation and bootstrapping; however given the finding above (that PLR BS CIs tend to be wider than their corresponding standard CIs but not usually by more than 0.1), the independent effect of weight estimation can be inferred by comparing this finding with the observation that normal CIs from weighted PLR models were consistently narrower (and tended to be contained within) the corresponding BS CIs, with a width difference in the region of at least 0.3.

Thus, this informal analysis of the general trends across scenarios supports the theory that weight estimation does in fact lead to smaller SEs, and that standard Cox CIs estimated without regard for correlation for each patient's time-varying weights are grossly inaccurate. In settings where normal CIs for PLR treatment effects can be validly estimated (not involving weight estimation), BS CIs tend to be marginally wider

than the corresponding normal CIs; surprisingly, however, BS CIs from corresponding Cox models in these analyses tend to be *narrower* than the normal CIs.

The Cox models seem to be more affected than PLR by the need for a reliable pool of patients on which to base weight estimation; in the ISC treatment change scenarios, the Cox model CIs are particularly wide, reflecting the uncertainty with which the weights are estimated.

For both the PLR and Cox models, as more treatment changes and TVCs are accounted for, the treatment effects move away from the null towards zero (in favour of VPS); this is especially true for Cox models; for example in Table 67, the HR fell from 0.77 (adjusting for no treatment changes) to 0.16 (adjusting for all treatment changes) for the monthly interval version of the Cox model.

This meant that, without adjustment for any treatment changes (ITT), Cox treatment effect estimates were higher (closer to one) than the corresponding estimates from PLR regression. However, as Cox estimates were affected more markedly than PLR by changing model factors, once treatment changes were adjusted for, Cox treatment effect estimates fell and were lower (closer to zero) than their corresponding PLR estimates.

2. Do IPCW treatment effects change as more **TVCs** are adjusted for in the weight estimation model?

As more TVCs are added to the model, the treatment effects tend to move further towards zero; although this is true for both the PLR and Cox models, it is especially obvious in the Cox model. Furthermore, as the number of TVCs increases, the resulting treatment effect CIs widen, reflecting the uncertainty with which the weights are

estimated within each pool of participants with each covariate combination at each time.

3. How does the choice of **time interval** impact on IPCW treatment effects? In particular, does the use of smaller intervals (weekly intervals) for IPCW analysis make Cox and PLR models more similar?

Generally, it can be seen in Table 64 to Table 67 (in Appendix C) that varying time interval width does not greatly impact on treatment effects estimated by PLR, though BS CIs do vary between time intervals (though showing no systematic or obvious pattern, thus potentially simply due to Monte Carlo error, which would be overcome with a higher number of repeat samples). In contrast, treatment effects tend to fluctuate more markedly across time intervals when using Cox regression, with the HR tending to decrease (in favour of VPS) as the time interval lengthens (from week to fortnight to month). Furthermore, as would be expected, the difference between Cox and PLR treatment effect estimates generally (but not always) increases with time interval.

4. What are the general differences in IPCW treatment effects when different **treatment changes** are considered?

As more reasons for treatment change are adjusted for, treatment effects tend to drop further away from one, in favour of VPS. This effect is again observed much more markedly for Cox rather than PLR models.

When weight estimation is based on few treatment changes (for example, when only adjusting for changes due to ISC), treatment effect CIs tend to be wide, especially for the Cox model. As more treatment changes are adjusted for, the CIs become narrower

due to increased stability of model weights because of the larger pool of patients on which the weights have been modelled.

Generally, although susceptible to overfitting (due to spline variables and TVCs added to weight estimation models), PLR models appeared more stable than corresponding Cox models. Cox models seemed to be more susceptible to changes to modelling factors (BS CI estimation, time interval and TVC adjustment) and displayed greater fluctuations in treatment effect than PLR models. This apparent greater stability of PLR, even when subject to severe overfitting, suggests an advantage of PLR modelling in this context, which counter the inevitable complications in modelling relating to the fitting of spline variables.

5. How do **IPCW and RPSFTM** treatment effects compare for each of these treatment change scenarios?

In order to allow direct comparison between the IPCW and RPSFTM analyses, it is now of interest to approach the RPSFTM analysis using the same specific research questions as carried out for the IPCW; as such, the same hierarchical series of analyses allowing for different reasons for treatment changes were carried out using the RPSFTM, the results of which are shown in Table 68. Given that no baseline covariates were included in the IPCW models (and it would be inappropriate to include any TVCs in the RPSFTM), no extra covariates are added to the RPSFTM here. However, it should be noted that these results differ slightly from those in Chapter 8, as this analysis relates only to those patients who were included in the IPCW analysis, namely those with available data on dose, seizure and AE counts.

As with the IPCW, the main disadvantage of focusing on more specific reasons for treatment changes is the resulting reduction in power due to the reduced number of treatment change events on which to base the analysis. This issue is all the more problematic in the RPSFTM setting, as it compounds the existing limitation that all treatment changes other than direct switches between trial treatments are necessarily censored in the analysis. Therefore, even when adjusting for *all* treatment switches (due to ISC, choice or UAEs), these direct switches between trial treatments make up only one sixth (one third) of all treatment changes occurring prior to remission time in the VPS (LTG) group; the remaining five sixths (two thirds) of treatment changes are necessarily censored. Thus when analysis is restricted to adjust for treatment switches due to ISC alone, this proportion (of treatment changes that are appropriately accounted for in the analysis) drops to less than one tenth (one quarter) of the VPS (LTG) arm.

Bearing in mind these limitations, it is however interesting to note that adjustment for treatment switches due to ISC alone causes the HR to move closer towards one, suggesting a smaller advantage of VPS compared to the result reflecting no adjustment for any treatment switches. (Note that this scenario does not entirely reflect the original ITT analysis, given that censoring is applied at the time of all treatment changes (other than just direct switches between LTG and VPS) as for the other sequential analyses.) This implies that if, rather than switching from VPS to LTG and vice versa, patients were treated with more appropriate dosing to tackle poor seizure control, the longer term beneficial effect of LTG in terms of providing seizure control, and ultimately remission, would be enhanced, although VPS would continue to be the treatment of choice.

However, it is equally important to note the relatively dramatic fall in HR (from 0.877 to 0.772) when treatment switches occurring for reasons of personal choice (as well as ISC) are adjusted for. This change in treatment effect is especially surprising, given that the ISC/choice analysis adjustment differs from the ISC analysis adjustment only with respect to data from two VPS patients who switched out of choice, suggesting that the RPSFTM is relatively sensitive to small changes in the data. However, this trend continues as more treatment switches are adjusted for; the HR falls further (to 0.756) when treatment switches due to UAEs are also accounted for.

In comparison to the IPCW analyses, adjustment for treatment switches using RPSFTM has a less dramatic effect on treatment effect estimates. This is expected, given that these treatment switches between LTG and VPS make up only one sixth (one third) of the treatment changes accounted for in the IPCW methods.

Overall, this investigation into more specific, clinically relevant causal research questions demonstrated that (across all model variations), as more treatment changes were accounted for, the advantage of VPS over LTG increased; in other words, without deviation from randomised assignment, the observed (ITT) treatment effect would have indicated a further advantage of VPS over LTG.

10.3. Conclusions

There were a number of reasons why the IPCW was likely to perform better than the RPSFTM in this trial: first, in contrast to the RPSFTM, the IPCW is easily able to handle multiple randomisation arms (although for the sake of comparison, the same two-way comparison will be presented here as in the previous chapter). Furthermore, unlike for most other models, the initial 12 month period being devoid of events will

not violate the assumptions of the underlying Cox model used in IPCW, as this model makes use only of the ranking of events, rather than actual event times.

The explicit allowance for time-varying confounders (under the NUC assumption) is a particular advantage of the IPCW, given that this form of confounding is especially relevant in the SANAD trial, where variables such as prescribed dose, and cumulative seizure and AE counts are likely to influence the probability not only of remission but also of treatment changes. For example, a high dose of treatment increases the chance of seizure control while simultaneously increasing the probability of adverse drug reactions, which in turn may trigger treatment changes. Indeed, it may be particularly important to consider adjustment for treatment doses in this trial, given the pragmatic trial design (where prescribed dose was completely at the discretion of clinician, and even the initial dose of randomised drug was not standardised) and lack of blinding.

The greatest advantage of the IPCW analysis over the RPSFTM, however, is its ability to deal with any form of treatment change, without the need to bias analysis by exclusion or censoring of patients. This is because an IPCW analysis handles all types of (relevant) treatment change identically: namely by artificially introducing censoring of patients at the point of treatment change while appropriately accounting for this censoring using weighting, under the NUC assumption. Thus, in adjusting for all forms of treatment changes, the application of the IPCW method to the SANAD data addresses a slightly different research question from that to which the RPSTFM related, by providing an estimate of the relative treatment effect between treatments that would occur in the absence of *any* departure from randomised AED.

Indeed, a direct comparison between IPCW and RPSFTM methods was not possible, as RPSFTM can only reliably account for a specific type of treatment change (in this case, direct switches between trial treatments); in order to allow a direct comparison between

these methodologies, adjustment in the IPCW analysis would have need to have been restricted to treatment switches alone. However, a further reduction in the number of treatment changes (on which weight estimation is based) to treatment switches alone would compound the existing problem of model instability due to low EPV ratios in the IPCW analysis. Unsurprisingly, therefore, the effect of RPSFTM on treatment effect reduction was less pronounced than with IPCW methods, as the majority of treatment changes were necessarily ignored (and in fact introduced censoring) in the RPSFTM.

10.4. IPCW

This chapter has demonstrated how a complex trial scenario can be translated into relevant causal contrasts to enable estimation using different approaches to survival analysis. However, there were a number of complications in the IPCW leading to necessary compromises and simplifications in the modelling exercise. It is now necessary to consider the conclusions in light of the inevitable limitations of this simplification process.

10.4.1. Limitations

There were a number of limitations relating to the application of IPCW methods in this context. Firstly, although the PLR version of the WD model seemed in general to be the most reliable and appropriate method to apply in this context, this model was complicated by the issue of how many knots to include in the spline variables. In this analysis, the same number of knots was used for the spline variables in both the remission and treatment change models. However, given the less complex pattern of probability of treatment change (compared to remission, which is complicated by the zero probability of event occurring prior to 12 months coupled with the peak of events at 12 months) and fewer treatment change events (compared to remission), it may have

been more appropriate to use a smaller number of knots (perhaps two or three) in the treatment change model, while using more knots in the remission model (clustered around the 12 month point to more accurately reproduce the evident non-linearity).

Additionally, PLR modelling was developed originally for a large study (with over 5000 subjects) (183), which provided many more treatment changes on which to base the WD model than in SANAD; thus, given the potential bias which increases as the number of (treatment change) events decreases when modelling is implemented without conditioning (187), it would have been interesting to investigate use of *conditional* (rather than pooled) logistic regression in this context.

Furthermore, PLR is known to approximate well to the Cox model only when the probability of event within each discretised time interval is small (for example, <10%) (183). This was generally the case for all time intervals in the SANAD study apart from one major exception: given the large number of remission events occurring at 12 months (with approximately 30% of patients achieving I12mR), this approximation is unlikely to hold at this time point. The magnitude of the bias introduced by this peak in event rates at the very start of the follow up period is unknown; thus it may have been useful to consider a simulation study (with known true HR and varying degree of peaks in event rates) in order to investigate the impact of this peak of event rates on PLR estimates.

However, despite complications in PLR modelling due to necessary considerations regarding time interval, CI estimation and spline variables (and evidence of severe overfitting in some cases), PLR models were more stable than corresponding Cox models, which were highly susceptible to fluctuations in treatment effect with changes to modelling factors. For example, the effect of time interval (week, fortnight or month) was observed in the Cox model when all other factors were held constant, suggesting

that the differences in TVC values (held fixed at the start of each interval) between the different time interval analyses impacted on treatment effect estimation.

The differences between the Cox and PLR models may have arisen because of the different way in which these models estimate the baseline hazard (or risk of event): PLR modelling explicitly introduces extra parameters (by way of spline variables) in the model to describe the hazard, while Cox models condition the baseline hazard out of the model likelihood.

A further limitation of the application of the IPCW method in SANAD was the fact that the reason for treatment change was provided only for the *first* observed treatment change (as the first deviation from randomised treatment related to the primary outcome, time to treatment withdrawal). Although drug and dose data were available regarding each patient's first three treatment changes, the *reasons* for these subsequent treatment changes were not collected. As such, when considering analyses adjusting only for certain reasons for treatment change, treatment changes occurring for *other* reasons were ignored without introducing artificial censoring; however, ignoring these changes (due to other reasons) invoked the implicit assumption that no further treatment changes relating to the reason of interest occurred subsequently in that patient. This demonstrates that, when considering the design of studies, trialists should consider carefully how data collection relating to treatment changes beyond the initial changes may impact on analysis limitations.

Another drawback of the IPCW was the reliance on patient-reported information for the TVCs relating to seizure and AE counts. Chronic disease, because of their long term nature, cannot be as closely assessed as acute conditions; as such, follow up information in trials of diseases such as epilepsy often rely on patient self-reports for information on covariates such as adherence, adverse events or clinical events such as seizures. As

previously discussed, it is well known that patients may over-exaggerate their adherence in order to avoid disapproval from their health care provider; however there is also evidence to suggest that patients may under-report the occurrence of symptoms of their disease, particularly if their occurrence will lead to unwanted measures being taken.

In the case of epilepsy, patients may not want to disclose the occurrence of seizures, especially if they have been seizure free for some time and are aiming to achieve a period of remission in order to regain their driver's license. Previous (unpublished) analysis suggests a difference in the proportion of males and females achieving remission in SANAD, despite there being no clinical reason for the apparent disadvantage in females with respect to this outcome. It is hypothesized that males may have a greater incentive to regain their license (for example, when their job relies on their ability to drive) and thus may under-report their seizures. As such, given the importance of accurate data on all relevant TVCs in order to meet the NUC assumption, the potential for biased under-reporting of seizures may limit the value of the IPCW analysis in this setting.

10.4.2. Assumptions

Having discussed the practical application of the IPCW and RPSFTM methods in detail, it is useful to provide a reminder of the more general key factors for consideration when deciding whether to apply each of these methods. Given that these methods differ in terms of their assumptions and data requirements, their relative appropriateness depends on the circumstances to which they are applied; thus, it is necessary to consider the plausibility of their assumptions with each application. However, given that these assumptions are on the whole untestable, it is more realistic to consider whether these assumptions are generally reasonable rather than necessarily wholly satisfied in the given context (142).

IPCW is based on the fundamental assumption that groups are appropriately rebalanced after reweighting each (uncensored) individual's contribution to analysis (195). However this in factor relies on two separate assumptions, not only that all factors used to match those who did and did not crossover (but who otherwise had identical prognosis) are known and measured, and therefore able to be accounted for when calculating weights (NUC assumption), but also that the trial setting provides a pool of exchangeable patients for each combination of prognostic factor, such that for each patient who change treatment, there are sufficient patients with equal prognosis and similar likelihood of changing treatment who did not switch treatments (positivity). If too few patients of each prognostic profile level either did or did not change treatment at a particular time point, the weight generation model will be unstable and extreme weights will result for the corresponding patients. These patients would then receive undue influence in the model, which will increase the chance of spurious model results (often manifested as extreme coefficients or large SEs).

Rather than sample size per se, therefore, it is the number of events on which to base weight estimation that is important (along with NUC) in order to ensure that the exchangeability assumption is met. Indeed the IPCW method has been shown to be prone to high levels of bias in relatively large sample sizes (500) when a high proportion (>90%) of control patients switch to experimental treatment, even when the NUC assumption holds (142). Therefore, as expected, in smaller sample sizes where very few events have been observed, the IPCW method is likely to perform poorly (196).

However, the NUC assumption represents a key limitation of the IPCW method (which, unlike the positivity assumption, cannot be overcome by increasing sample size) and may be particularly problematic in an RCT context. Typically RCT datasets are much smaller than observational datasets and when fewer data are available, the IPCW

method may become less stable and CIs may become wide. In addition, key predictors of treatment switching are not always collected in RCTs (such as patient switching preferences) and often data collection on key indicators is stopped at some point (for example, upon treatment discontinuation or disease progression) which hampers the applicability of the IPCW method (142).

With time-varying confounders, the exchangeability assumption (dependent on the NUC and positivity assumptions) is particularly unlikely to be true throughout follow up, especially as number of remaining patients receiving treatment as randomised falls. Furthermore, as observed in this analysis, large values of time-dependent confounders strongly associated with treatment change may lead to extreme weights (even after stabilising) and subsequent invalid causal estimands.

The NUC assumption as part of the IPW method introduces further conundrums. Although the probability of treatment change is assumed to depend on known confounders alone, in order to allow valid IPW estimation, it is necessary that *not all* patients with these confounding features actually do change treatment. Thus it must be considered why these patients did not experience a treatment change; it may simply be a random occurrence, or it may be that these patients differ systematically from those who did change treatment, indicating that an additional selection effect has been missed (and the NUC assumption is therefore invalid) (111).

Thus Robins (111) stipulates that the IPTW method should be accompanied by sensitivity analyses to assess the degree to which violations of the NUC assumption may change analysis conclusions. Robins demonstrates how such sensitivity analyses may be carried out, first by defining a measure that quantifies the degree of confounding due to unmeasured factors and then assessing whether conclusions alter materially when this measure is allowed to vary over a range of non-zero values.

10.4.3. Further investigations

As demonstrated in this chapter, even in a large trial like SANAD with common treatment changes and excellent follow up (without the usual problems when deviation from treatment protocol leads to deviation from follow up protocol), there remained uncertainty regarding the most appropriate method to use, with obvious major limitations and assumption violations with each method. In particular, the limited number of events relating to the most important clinical question (of adjusting for treatment changes due to ISC alone) meant there was very little scope for further extensions to the model in this context. This is unfortunate, given that there were a number of issues that warranted further investigation in this trial.

Firstly, given that loss to follow up is often related to prognosis, it would have been interesting to assess the impact of (potentially informative) loss to follow up (when patients' outcome data are censored prior to the administrative censoring date of the trial or to achieving remission) on conclusions, for example using IPCW methods to calculate weights relating to the probability of loss to follow up (rather than treatment change). However, there were insufficient numbers of patients who experienced censoring prior to remission or treatment change to allow valid estimation of such weights.

Furthermore, it would have been useful to investigate the impact of assumptions regarding those 42 patients with missing seizure data for at least one of their clinic visits, comparing simple and multiple imputation techniques (with estimation of missing values based on predictive patient characteristics) as described in Chapter 9.

The IPCW model could have been extended to compare all three randomised arms in arm B of the SANAD trial. However, this three-way comparison was not attempted,

not only because of the additional complexity that would be introduced by such an analysis, but also because the clinical question of interest related particularly to the two-way comparison between VPS and LTG alone.

A further data quality issue affecting the IPCW model was the assumption that long periods without follow up data indicated *no* change in outcome or treatment prescription. Given that the occurrence of any missed clinic visits may well be related to a patient's condition, it would have been interesting to carry out a sensitivity analysis where patients' data were censored if the interval since their previous follow up information exceeded a certain (clinically informed) time period (for example, six or twelve months), with adjustment for potential bias introduced by such censoring accounted for using IPW (adjusting for all confounding factors predicting loss to follow up and outcome). If the overall conclusions of the IPCW were affected greatly by varying the time before follow up was censored, this would demonstrate the sensitivity of the analysis to the assumption that all follow up data were accurately accounted for.

A further consideration regarding data accuracy is that of measurement error. For example, it is known that patients with extreme numbers of seizures (some with multiple seizures per day) were only able to provide approximations of the number of seizures experienced between clinic visits. The impact of this inevitable measurement error is likely to be inflated due to the increased influence of extreme values of TVCs in determining time-varying weights. Although this issue would be addressed to a certain degree through the use of truncation of seizure counts (at the 95^{th} centile), it would have been useful to assess the sensitivity of conclusions drawn when extreme seizure counts were varied systematically (for example, increased or decreased by a certain amount).

It would also be interesting to investigate further the impact of extreme weights on analysis, as were observed in this application, potentially using simulation studies.

Indeed, the practical and statistical problems associated with weight estimation that were encountered during this analysis are not unusual, and research into improving methods for weight estimation is a continuing area of debate (197).

Adherence to treatment prescription

However, perhaps of greatest interest and impact would be the consideration of adherence to treatment prescription. Given that there were no detailed compliance information collected as part of the SANAD trial, it was not possible to assess the effect of treatment adherence, rather than simply treatment *prescription* on remission. Given that AEDs can lead to unpleasant side effects, it is likely that a substantial proportion of patients do not always adhere to their daily dose. It is quite plausible that the effect of nonadherence to treatment prescription on outcome could be of a similar magnitude to that of changes in treatment prescription; however, without any adherence data, it was not possible to investigate this in the causal analyses presented in these chapters.

In order to explore the likely impact of adherence on causal estimation, it would be useful to carry out a simulation study, assessing the effect on outcome of various levels of compliance and various degrees of association between a patient's compliance and their prognosis. The parameters used to simulate these compliance data would be informed by the small amount of compliance information collected in the annual quality of life questionnaire in SANAD. In estimating the *impact* of (ignoring) adherence information on *clinical* interpretation of results, this simulation study would differ from that previously suggested, which instead would assess the *statistical validity* of applying this methodology in this setting.

An additional simulation study assessing impact would also be useful to investigate adjustment for particular reasons for treatment change. In particular, although clinicians

were primarily interested in the question of treatment effect adjusted for treatment changes caused by ISC, there was inadequate power in the SANAD data to estimate this treatment effect with any confidence. The reason for this is that patients are likely to experience treatment changes due to other events before making changes due to ISC. If seizures persist, rather than giving up and changing treatment immediately, clinicians often initially increase the dose of randomised drug to see if seizure control can be achieved through more appropriate dosing (198). However given that increasing dose increases the likelihood of adverse reactions, patients may then be censored for reasons related to UAEs prior to switching treatment (because of ISC) or achieving remission. Alternatively, treatment changes for other indications may well occur (prior to remission or treatment changes due to ISC), such as the patient requesting to change treatment (for example, because they believe themselves to be in a state of remission (even though they have not yet achieved 12 months of remission from seizures) or because of a desire to become pregnant, which is contraindicated with some AEDs). In providing more power than the original SANAD data, a simulation study based on observed treatment change profiles in SANAD would allow valid and independent estimation of specific weights for each cause of treatment change (ISC, personal choice or UAEs); thus it would be possible to use IPCW methods to obtain individual's weights (and hence adjust) for each (relevant) cause of treatment change.

Furthermore, it would then be possible to use results from such analyses to adjust for *multiple* reasons for treatment change, with patients' overall weights obtained simply by multiplication of their individual (relevant) reason-specific weights. The resulting treatment effects, adjusted for each combination of reason for treatment change, could then be compared to those obtained using the sequential approach demonstrated in this chapter; if broadly similar, this would validate and support the use of the hierarchical

approach in model scenarios with low power (given that weight estimation for individual reasons for treatment change is compromised when there are insufficient treatment changes for each individual reason within each treatment group).

10.4.4. Alternative methods

In censoring follow up at the point of each patient's first treatment change, IPCW methods may present an advantage, if follow up data post-change are considered misleading or unhelpful, and hence may be more powerful than ITT. If, instead, the follow up data after first treatment change are believed to contain potentially useful information, it may be advantageous to fit a MSM, which uses information from all periods of treatment, in other words, *after as well as before* the change from randomised treatment.

For example, IPTW weighting could have been applied to the SANAD data, where these weights are estimated from the inverse probability of *observed* (rather than simply randomised) treatment. Rather than artificially censoring at the time of treatment change, an IPTW model regresses the outcome on a suitable summary of treatment history (such as a time-varying binary treatment indicator) throughout follow up, adjusting for each individual's estimated probability of receiving their observed treatment. The difference between IPTW (MSM) and IPCW methods may best be understood by considering their interpretation in the absence of any relevant TVCs: is such a case, an IPTW model would reduce to a simple (unweighted, baseline covariate-adjusted) AT analysis (with a time-varying treatment indicator); in contrast, without TVCs, the IPCW method reduces to an (unweighted, baseline covariate-adjusted) PP analysis (censored at the point of first relevant treatment change).

Such a method would also have been feasible for the honey trial, in which factors likely to confound the relationship between treatment and outcome included a patient's age and their (time-varying) wound size. Clinicians were inclined to treat younger, fitter patients with honey as they believed it to be a promising treatment for such patients; this is evident when comparing the median (IQR) age of the five patients who were switched from conventional to honey treatment by their clinician [44 (38, 53) years] to that of the remaining patients [72 (56.5, 79) years]. Conversely, treatment of large wounds with a poor prognosis was likely to be switched to more extreme modalities, such as surgery. Thus, in theory, it would have been feasible to fit an IPTW model to this trial data, by first using a Cox or PLR model to predict each patient's probability of receiving their actual observed treatment for each time interval, dependent on their age at randomisation and (time-varying) wound size. The inverse of these predicted probabilities could then be used to weight analysis of time to healing (or time to 50% reduction in wound size) regressed on treatment received, thus providing a treatment effect estimate free from selection bias. Furthermore, it would have been interesting to account for LTFU (for example, when patients moved wards or hospitals) using separate IPCW (where censoring is due to LTFU rather than artificial censoring at the time of treatment changes). However, the validity of this model would rest on the questionable assumption that these two prognostic factors wholly account for the association between treatment changes and prognosis. Unfortunately there were insufficient numbers of treatment changes to allow the IPCW method to be performed with any confidence.

10.5. RPSFTM

10.5.1. Limitations

As discussed in Chapter 8, there were a number of severe limitations when the RPSFTM was applied to analysis of time to 12 month remission in the SANAD B trial; in particular, the model is unable to accurately represent the peak in remission events at 12 months or the lack of events prior to 12 months, resulting in model predictions of remission event times prior to 12 months.

Furthermore, in handling only one form of treatment change (direct switches between randomised treatments), patients' follow up was necessarily censored at the time of all other treatment changes (thus introducing bias in a similar manner to that of PP analyses). The application to the RPSFTM of the hierarchical structure (adjusting for certain reasons for treatment change) in this chapter presents a further challenge, due to the reduction in power as fewer numbers of relevant treatment switches are adjusted for in the restricted treatment change scenarios.

10.5.2. Further investigations

As such, given the severity of these fundamental limitations of the RPSFTM in this context relating to violations of basic necessary assumptions and limited power, a discussion of potential refinements or extensions to the model may seem somewhat redundant. However, for the sake of completeness, these will be presented here, particularly as a number of these model variations would, in theory at least, provide partial solutions to some of these major flaws.

10.5.3. Assumptions

The first of these potential solutions relate to the likely violation of the common treatment effect in SANAD, given that the treatment effect may vary according to whether treatment is received on diagnosis (which was the time of randomisation for the majority (87%) of recruited patients in these treatment groups) or following treatment failure of the alternative (randomised) treatment. Although in theory it would be possible to extend the RPSFTM to include more than one treatment effect (perhaps with a different treatment covariate for treatment received prior to and after treatment failure (of randomised drug), thus differentiating between treatment received from randomisation or following treatment switch), the resulting complexity of the model would mean that, in practice, the additional parameter would be not be estimable with any degree of certainty.

The RPSFTM assumes the treatment effect is multiplicative, immediate and constant (which may not be true, as a patient's response to treatment may change as patient progresses through disease); therefore the acceleration factor is interpreted as an average effect over time, across patients and disease stages (195).

Thus it is important to consider whether the assumptions underlying the modelled treatment effect are likely to reflect the true treatment effect. For example, if the common treatment effect assumption does not hold (such that the treatment effect differs according to when treatment is initiated), the RPSFTM will produced biased results (as demonstrated by Morden (137)).

Given that treatment switches are likely to be associated with prognosis, it is not possible to confirm the validity of the common treatment effect assumption using observed data alone; instead it is important to seek expert clinical guidance in order to

back up claims regarding the plausibility of this assumption. In particular, it is useful to ascertain the likely variation in effectiveness of treatment as disease progresses (142). For example, the plausibility of the common treatment effect may be questionable if treatment switches only occur when clinically indicated, such as on disease progression, at which time the capacity for a patient to benefit may be different compared to pre-progression because of their advanced disease.

Although it would be of interest in many clinical scenarios, therefore, to estimate multiple treatment effects, the RPSFTM is limited in that it becomes increasingly unstable when trying to estimate more than one such parameter. As such, alternative methods are required to address this issue.

10.5.4. Alternative methods

One such method, an alternative two-stage version of the RPSFTM, was proposed and demonstrated by Robins (95) and Yamaguchi (199) respectively, in an attempt to overcome the problem of violations of the common treatment effect assumption. These methods are appropriate when there is a definitive secondary baseline (for example, date of disease progression) at which point *alone* switching may occur, such that randomisation is assumed to be adhered to until this point. These methods effectively recognise that the clinical trial is randomised up until the point of disease progression, for example, but beyond that point, essentially becomes an observational study.

This two-stage approach first uses a counterfactual SNM (which, as described in Chapter 5, is similar to the RPSFTM but relies on the assumption of NUC rather than randomisation for g-estimation) to obtain estimates of the treatment effect specific to control patients in order to allow estimation of their counterfactual event times (in other words, as if they had not switched to receive experimental treatment); these

"shrunk" survival times then form a counterfactual dataset unaffected by switching, thus providing an inferred "uncontaminated" control arm against which the treatment effect specific to experimental group patients can be estimated. However, given that the SNM is essentially an observational version of the RPSFTM and attempts to account for time-dependent confounding using the NUC assumption, this method has similar limitations to the IPCW (135). In particular, when very high switching proportions are combined with small sample sizes, these two-stage methods are likely to be prone to error and bias.

Latimer (142) proposes that these methods can be simplified to remove dependence on g-estimation, if it is plausible to assume no time-dependent confounding between time of disease progression and time of treatment switch, such that only those factors collected *at the point* of treatment switch which determine treatment switch are adjusted for, rather than requiring data on all time-dependent confounders (as required by the more general two-stage approaches).

Such two-stage methods remove the need to rely on the potentially implausible common treatment effect assumption, but they do depend on identifiability of a suitable secondary baseline (at which point (alone) switches occur and at which all patients are assumed to be at a similar stage of disease) and on the NUC assumption to hold at this secondary baseline time point. When switching occurs soon after the secondary baseline, the scope for such time-dependent confounding is limited, but this is not the case if switching happens substantially after the secondary baseline, in which case the potential for such bias becomes substantial (142).

In summary, this simplified two-stage model removes the dependence on the common treatment effect assumption, and does not require data to be collected on confounders at time points other than at the time of the treatment switch. However, such a method

may be prone to time-dependent confounding if switches occur beyond this time point. Furthermore, it may not be generalisable to all disease areas, given its reliance on the ability to identify a secondary baseline at which point all switches occur.

These methods may be particularly relevant for cancer clinical trials, where the decision when and whether to initiate second-line treatment on disease progression is often left largely at the discretion of the treating clinician, though the protocol may provide brief or broad guidance on this. Factors which influence this decision include the patient's clinical condition, abnormal laboratory markers or adverse drug reactions; as such, the effect of second line therapies may be viewed as arising from observational studies (136). The argument against the common treatment effect in this clinical setting stems from the fact that a constant AF essentially consider first-line and second-line patients equally, whereas in clinical practice they are considered to be different patient populations.

However, in disease areas such as epilepsy with less clearly defined points of treatment change (as a clinician's decision to alter treatment prescription varies on an individual patient basis, depending on their own tolerability and priorities), such a secondary baseline is less obvious, therefore rendering such methods less useful.

A further potential refinement to the RPSFTM in this setting relates to the typical lengthy withdrawal periods when patients switch AEDs; in theory it would therefore be relevant to consider allowing for lagged effects of treatment, whereby the treatment effect is extended by a certain time period beyond the recorded date when withdrawal began, either up to the recorded date when the drug was taken for a final time, or by assuming a general withdrawal period (informed by clinical opinion) for all patients coming off each treatment.

As discussed in Chapter 8, mixture modelling may be applied in order to get around the problem of necessary truncation of AF prior to 12 months (as 12 month remission cannot, by definition, occur before this time) and the peak of events at 12 months. Mixture modelling involves fitting separate models for immediate and delayed remission, where the immediate remission model needs to reflect treatment over 12 months but is complicated by the fact that a basic time-varying treatment covariate is appropriate only until the time at which first seizure occurs.

An alternative approach to adjustment for treatment changes would be to consider adjustment for changes to prescribed treatment using joint modelling of survival outcome and longitudinal dose data (under an NUC identifying assumption), as demonstrated by Williamson et al (198) for the CBZ versus LTG comparison among arm A SANAD patients. Given that dose titration differed between these treatment arms (as clinicians were more familiar with the standard CBZ drug and hence tended to prescribe higher doses more quickly, whereas they were more cautious with newer LTG), there was a concern that UAEs would be more likely in patients randomised to the more well-known standard drugs. Hence analysis was carried out, adjusting for standardised dose levels between treatment groups (calculated by dividing patients' doses by their treatment-group-specific median initial doses) in joint modelling of dose and survival. This analysis demonstrated that the observed benefit of LTG in terms of preventing AEs was maintained even allowing for the difference in titration rates between arms, and that the two arms remained similar in terms of seizure control once doses had been accounted for.

Although useful in adjusting for differential rate titration between groups, this joint modelling analysis did not, however, account for treatment changes or switches between arms: supplementation with additional drugs was ignored, and once

randomised treatment was withdrawn (in other words, the dose of randomised treatment dropped to zero), analysis was censored. Given that such censoring is likely to introduce bias, it would be desirable to extend the joint model to allow for between-group switches using RPSFTM methods. However, given the complexity of both the joint modelling and RPSFTM models, extensions to allow for between-arm switches as well as dose titration would be overly complicated and beyond the scope of this PhD.

Alternatively, it would be possible to combine the methodology of the RPSTFM (to model the effect of direct switches between randomised treatments of interest) with that of the IPCW (to remove the bias caused by the necessary censoring of events that occur prior to or after the "true" switches of interest and also to potentially allow for competing events). However, any benefit gained by reduction in bias when undertaking such an analysis would be unlikely to outweigh the problems caused by the additional model complexity or underlying flaws associated with the use of the RPSFTM in this context.

Thus, in applying the hierarchical structure regarding reasons for treatment change to the analysis of T12mR in this chapter, it is apparent that the severe limitations of the RPSFTM, discussed in Chapter 8, are compounded further. In adjusting only for those switches occurring for particular reasons, the usefulness of this model is reduced, while the bias is increased (because of additional necessary censoring when patients experience treatment switches for other reasons).

10.6. Conclusions

This comparatory analysis has demonstrated a number of key conclusions regarding the use of IPCW and RPSFTM methods in a context such as SANAD.

In particular, as regards IPCW modelling, analysts must consider whether the underlying assumption of NUC is convincing, and whether accurate data are available on all relevant confounders between treatment changes and outcome (for example, information used by clinicians when deciding to change a patient's treatment prescription or when recommending a treatment switch, as this clinical information will often also relate to outcome). Furthermore, it is necessary to determine whether there is a sufficiently large pool of those who did (and did not) change treatment at each particular time point within each category (defined by cross-classification of all confounders in the model) to allow reliable weight estimation; in practical terms, this requires assessment of the proportion changing treatment in relation to the sample size at each time point for each combination of patient factors; analysts must also be vigilant regarding evidence of extreme weights or coefficients.

Technically speaking, when applying IPCW methods, it is not necessary to continue outcome follow up beyond the time of the first relevant treatment change. However, given that it may be of interest to apply sensitivity analyses (for example, adjusting for other forms of treatment change) or to compare IPCW methods with those obtained using the RPSFTM, it is wise to collect survival information on patients beyond the time of treatment changes until the outcome (or censoring) event occurs.

In contrast to the IPCW, which accommodates all types of treatment changes (indeed, treatment changes are not differentiated in the analysis; they are all treated identically, by artificial censoring at the point of the first treatment change), the RPSTFM cannot easily handle more than one type of treatment change. Although the RPSFTM could in theory be extended to adjust for multiple forms of treatment change, in practice such complexity leads to model instability (94). Generally, the RPSFTM will be most appropriate when the treatment deviations mostly take the form of switches between

trial treatments and when it is plausible to assume that the effect of treatment is constant (regardless of when treatment is received).

Thus, the main considerations when applying RPSFTM relate to the assumption of a common treatment effect and limitations regarding adjustment for only a single form of treatment deviation. Thus, analysts may need to consider whether the treatment effect may continue after discontinuing treatment (in which case it may be appropriate to introduce a lagged treatment effect; for example, this may have been particularly important in SANAD given the long withdrawal phases *following* the date when treatment was officially stopped) or whether the impact of treatment may differ according to the clinical state of the patient (in which case the common treatment effect assumption is violated). Although these assumptions are untestable with empirical data, it is possible to evaluate the performance of the RPSFTM by examining the model results; for example, analysts should consider the extent of recensoring (and its potential impact on treatment effects, if a treatment-time interaction is clinically plausible) and judge the success of G-estimation (by assessing the similarity of the distribution of the adjusted counterfactual ("untreated" or "control" state, as appropriate) event times between randomised groups).

10.7. Summary

As demonstrated in this chapter, even in a large trial like SANAD with common treatment changes and excellent follow up (without the usual problems when deviation from treatment protocol leads to deviation from follow up protocol), there remained uncertainty regarding the most appropriate method to use, with obvious major limitations with each available method. In particular, the limited number of events relating to the most important clinical question (of adjusting for treatment changes due

to ISC alone) meant there was very little scope for further extensions to the model in this context.

In the next chapter, general guidelines will be presented for trialists facing similar decisions on how best to adjust for treatment changes in order to estimate causal effects in a randomised trial with survival outcomes, taking into account the modelling requirements and practical limitations of the alternative methods.

11. Discussion

The aim of this thesis has been to explore, explain, demonstrate and compare the use of causal methodologies in the analysis of RCTs, in order to make the available (but somewhat technical) statistical methods more readily accessible and comprehensible by non-specialist analysts.

The thesis began with an overview of the impact of treatment nonadherence in general practice as well as in trials, highlighting that nonadherence to medical advice has been recognised as a widespread problem since the 1970s but its impact on clinical trial analyses is less well appreciated. Increased understanding of the reasons for deviation from prescribed treatment and recognition of the expertise patients often have regarding their own disease state have led to a move away from the traditional paternalistic, authoritative relationship of a doctor with their patients, towards a more shared decision-making model of concordance. However this notion of concordance does not easily translate into the world of RCTs, because of the need for inferences to

be made based on a known treatment structure, required for even the most pragmatic of trial protocols.

Any deviation from assigned treatment renders ITT analysis (the only analysis that properly mirrors the randomisation process) defective in estimation of any effect except effectiveness of treatment as implemented in the trial. Examples of real life trials featuring different patterns of nonadherence and various research questions of interest demonstrated the need for methods beyond ITT to answer such questions.

Thus an overview of statistical methods was presented, intended as an aid to researchers who are new to the field of causal inference, thus presented with an emphasis on practical considerations necessary to ensure appropriate implementation of techniques and a particular focus on interpretation of methods, rather than technical detail. This background research and consideration of the wider issues (for example, when it is appropriate to use different methods, and how to approach such analyses with necessary consideration of both the clinical and statistical issues) has been usefully summarised in a number of guidance tools, complemented by demonstration of analyses, intended to provide guidance for trialists and statisticians on how to approach causal analysis.

In particular, this statistical literature review demonstrated the bias inherent in naïve methods which fail to respect randomisation, and presented various statistical methods to more appropriately estimate causal effects, with a particular focus on survival methods, as these are often relevant in settings typically associated with nonadherence (featuring long term treatment and follow up).

A review of published trials then demonstrated that, despite the fact that deviation from allocated treatment occurs in the vast majority of trials, issues relating to bias caused by

deviation from treatment are rarely considered or addressed appropriately by trial analysts, suggesting the need for wider appreciation of the availability of other causal methods.

Following a description of the necessary practical and statistical considerations required for such causal analyses, two methods to account for treatment deviation were presented and applied in the analysis of survival outcomes, in trials of contrasting complexity. In particular, the SANAD trial provided numerous analysis issues and challenges, due to the complexities of the trial design and nature of treatment changes that occurred. These challenges are likely to be common in other chronic disease areas, where treatment changes may occur for a number of competing reasons and interest lies in achieving remission of symptoms. As such, this novel application of complex methods in this complicated trial scenario provided a useful example for discussion of the analysis issues and methods, striking a balance between simplifications of compliance scenarios and associated assumptions in order to make the methods sufficiently easy to apply, and yet without compromising on the necessary detail on nature of and reasons for non-compliance which importantly inform analysis.

This application of methods has provided a transparent discussion of issues, potential biases and limitations, as well as how to interpret results obtained from causal methods for analysis of survival outcomes. In particular, these analyses demonstrated a number of issues which are likely to be common among trials with long term treatments and deviation from protocol, for example, when difficulties arise when trying to account for numerous forms of treatment deviation (in RPSFTM) or problems relating to extreme weights (in IPCW modelling). These issues were addressed using various approaches, for example, truncation and transformation of skewed covariates in order to reduce the impact of the resulting extreme weights in IPCW modelling. Other complications in

IPCW modelling included the need to apply variable selection processes independently of (invalid) p-values, the use of bootstrapping to address the problem of deflated SEs resulting from weight estimation, data preparation and modelling complexities associated with time-varying covariates, issues of low power when investigating competing reasons for treatment change, and the need to consider knot positions for underlying time spline variables necessary for PLR modelling.

The analyses demonstrated in these final chapters have demonstrated how the two pictograms presented in Chapter 7 can be used, first to clarify the compliance scenario such that a relatively complicated compliance situation, such as that in the SANAD trial, can be simplified into a feasible scenario for analysis, and secondly to guide the analyst when choosing which appropriate method of analysis to employ. Sketching a cDAG may also help to clarify the causal scenario, for example in helping researchers to determine whether factors are likely to be time-dependent confounders affected by prior treatment (and thus appropriate for adjustment using IPCW methods) or simply time-dependent covariates.

A final reference tool for researchers is now provided, presenting general guidelines for trialists when faced with the decision of whether to use RPSFTM or IPCW methodology to adjust for treatment deviations when analysing survival outcomes.

11.1. Choice of survival analysis methods

When planning to carry out adjustment for treatment changes in trials with survival outcomes, analysts need to consider which of the available methods would be most appropriate, while recognising that no one method will be appropriate for all circumstances; instead the performance of each method will depend on the particular trial setting to which they are applied. Table 43 and Table 44 provide a reminder for

trialists of the modelling assumptions and practical limitations of the alternative methods.

Table 43 RPSFTM considerations

Model feature	Consideration required
Common treatment effect	Consider whether treatment effect is likely to be **constant** regardless of when treatment commenced.
Secondary baseline	If treatment switches tend to occur at a common "secondary baseline" (for example, on progression), at which point treatment effect is likely to differ from treatment started at randomisation (for example, on diagnosis) thus **violating common treatment effect**, consider instead using **two-stage RPSFTM**.
Various forms of treatment change	Consider **how to address treatment changes** (other than those directly accounted for by definition of "on" and "off" treatment in model) in relation to causal research question of interest: substantial numbers of treatment changes (which are not relevant to the causal scenario in question) will undermine validity of analysis, due to necessary censoring at the time of such changes.
Impact of recensoring	Assess effect of **recensoring** by checking the number of event times (and events) that were recensored, considering the potential impact on treatment effect estimation (if a treatment-time interaction is possible).
Test model performance	Assess **success of G-estimation** by comparing the counterfactual distributions (the control state event times estimated by applying the optimal AF to the RPSFTM) for the treatment and control arms; these should be similar under the randomisation assumption.

Table 44 IPCW modelling considerations

Model feature	Consideration required
Selection of TVCs	Consider how best to determine which TVCs are important in predicting treatment change and outcome: consult **clinical opinion**; may be necessary to apply **selection procedure** (if numerous TVCs)
Functional form of covariates	Check optimal functional form using `lowess` curve of **martingale** residuals (from Cox model)
Extreme covariate values	**Truncate** at the 99th centile to **avoid extreme weights** (which in turn distort treatment effect estimate) due to influential outlying values of important predictors of treatment change/outcome
Time intervals (for discretised TVCs)	Strike the **balance** between greater **accuracy** (increases as interval length decreases) and **computational intensity** (increases with interval length)
Model type: **Cox or PLR**	**PLR** is useful if using lagged variables or if TVCs change frequently (and therefore are too complicated to be analysed without discretising)
	Cox modelling avoids the need to consider splines to mirror underlying risk function in PLR model
Splines **(for PLR only)**	Create and use **treatment-specific spline** variables for WD model, but use **overall splines** for the WO model
	Consider **shape of underlying risk**, in order to identify times where risk changes, in order to inform **positioning of knots**
CI estimation	Estimate CIs using **bootstrapping** to overcome correlation due to within-patient time-varying weights (in Cox model) and the reduction in SEs due to the weight estimation procedure
EPV ratio	Consider ratio of number of variables in model to number of treatment change events, in particular when considering the **number of knots** to use for spline variables (spline with ***k*** **knots** requires ***k*** **variables**)
Model assumptions	Consider plausibility of **NUC** assumption (whether all confounding variables have been accounted for) by seeking clinical expert opinion
	Examine model weights for evidence of violation of **positivity** assumption: extreme weights may indicate unreliable pool of patients who do (or do not) change treatment at a particular time (within a given subgroup of patients defined by cross-classification of model covariates)

Given their differing assumptions and data requirements, neither IPCW nor RPSFTM methods will be directly applicable in every trial setting. It may be useful to apply both (or alternative variations of these) methods, as a means of assessing the sensitivity of results to the associated assumptions (200, 201). Furthermore, although it is good practice to decide on an analysis plan before undertaking analysis, the application presented in the previous chapters demonstrate that the analysis plan may need to be altered as analysis proceeds. However, in order to avoid accusation of selective reporting, it is important to be transparent and report fully on all modifications of analyses that take place. Thus, if several methods have been applied, it is important to provide a discussion of the limitations and potential biases associated with each method when presenting results, as an aid for interpretation.

11.2. Economic evaluation

These two causal methods for survival outcomes are important, not only from a patient or clinician perspective, but also for economic evaluations of drugs by regulatory authorities. This is a particular problem in cancer trials, given that patients are often permitted (or indeed encouraged) to switch treatments on disease progression or relapse (142), leading to distortion of OS results. Furthermore, licensing bodies such as the United States Food and Drug Administration and the European Medicines Agency may accept PFS as a primary endpoint for drug approval, reducing the incentives to maintain trial randomisation beyond disease progression. However, although showing an OS advantage may not be essential for obtaining marketing authorisation, a lifetime horizon is generally advocated in economic evaluations, especially for interventions that impact upon survival (as recommended, for example, in the NICE Guide to the Methods of Technology Appraisal). Thus interest is growing in how to address the issue of treatment switches in order to provide unbiased and clear treatment OS comparisons

which are not available from ITT, PP or AT analyses. Causal effect modelling techniques are receiving attention as a potential solution to the problem, and there have been a number of submissions made to NICE and the Scottish Medicines Consortium which have applied IPCW or RPSFTM when considering cost-effectiveness of a number of cancer drugs (202-204).

However these applications tend to be post hoc, secondary analyses devised to cope with problems arising with the interpretation and use of ITT estimates. In order to increase their credence, more thought needs to be given to the use of such methods at the protocol development stage, considering the degree to which treatment changes are likely to occur, what information is likely to be used in making the decision to change treatment and which methods of statistical analysis are likely to be the most appropriate.

Lack of available compliance data provide further barriers to reliable assessment of non-compliance in CE assessments. As such, despite the significant impact of non-compliance on clinical and economic factors, the impact of non-compliance on drug therapeutics is rarely considered. Reporting of the extent and nature of clinically relevant compliance measures by trialists would facilitate assessment of non-compliance in CE assessments, an area of potential further research.

11.3. Future work

In addition to the proposed extensions to RPSFTM and IPCW modelling discussed in Chapter 10, it would be interesting to use qualitative research methods to explore factors relating to the reasons for non-compliance, either on the part of the patient or treatment provider. Greater understanding of the reasons why patients do not adhere could usefully inform compliance analyses, in a manner analogous to using data on reasons for missing outcome data to inform sensitivity analyses. Furthermore,

knowledge regarding typical barriers to adherence could be used to improve likelihood of participant adherence by introducing relevant adherence-enhancing measures in treatment protocols.

Following on from the review of adherence information reported in published randomised, it would also be interesting to examine the nature and quality of such data typically obtained from routine clinical records, as these records may provide a useful source of adherence information to inform analysis of observational studies.

11.4. Statistical analysis plan

In order to ensure availability of necessary data and protect against selective reporting, trialists need to consider at the design stage the statistical analysis methods that will be employed to adjust for non-compliance. First, it is necessary to consider the trial aims and likely directions of bias introduced by any anticipated adherence problems, in order to determine whether analysis by ITT is likely to be appropriate or sufficient (135).

Regardless of trial aims, however, reporting information on the uptake and acceptance of treatment is important for the interpretation of the success of the trial treatments, even when analysis does not aim to adjust for nonadherence to treatment protocol. Even when clinicians are not interested in an explanatory analysis *per se* (but instead are interested in the effectiveness of the policy of *starting* with a certain treatment, for example), it is nevertheless important for clinicians to be aware of what changes did occur; otherwise, without an understanding of what the trial treatment policy entailed or how it panned out, it is not possible to fully appreciate or interpret the effectiveness of the trial treatment policy. Even when a trial does not involve many treatment changes, this fact should be communicated so that those interpreting the results are aware that the ITT result is likely to closely mirror the explanatory effect of treatment. Thus,

regardless of whether a trial is designed and analysed to demonstrate effectiveness or efficacy of treatment, or any measure in between, it is important to provide a clear description of the degree and nature of treatment changes.

To this end, it may be helpful to create a **monitoring plan** which specifies how all relevant compliance data will be collected, recorded and reported during the course of the trial. Data should be collected in order to ensure clinically relevant summaries of compliance can be created (3). Reporting missing data is as important as disclosure of treatment deviations, given that the two are often related and interlinked (121). In particular, trial reports should distinguish withdrawal due to LTFU from active decisions to exclude patients from analysis (for example, due to withdrawal or deviation from treatment protocol) (41).

Furthermore, without the collection of required data, any necessary statistical methods will never be realized; thus it is important to plan which variables should be collected, relevant to the chosen statistical methods (for example, which baseline and post-randomisation time-varying covariates will predict treatment change, to facilitate IPCW estimation), with details given on how these compliance and covariate measures will be measured in practice. The choice of statistical methods should be discussed and justified in a detailed **statistical analysis plan**, given that each method has its own different advantages and disadvantages and relies on different assumptions, considering the use of sensitivity analyses to assess departures from identifying assumptions.

Likewise, investigators must interpret the results of analyses presented in trial reports in light of (lack of) information regarding (or adjustment for) treatment adherence changes, critically appraising the trial design, conduct and analysis (50).

11.5. Power and sample size

Analysts should also consider the effect of nonadherence or treatment changes on the power of trial analyses. Non-compliance in a trial typically reduces the power of ITT analyses because the treatment experiences of randomised groups are in fact more similar than intended. Although it may seem natural to aim to recover this lost power, it is often impossible to do so using the methods discussed above without making additional unverifiable assumptions regarding the comparability of those who do and do not comply, such as those underlying PP and AT analyses (69).

For this reason, potential loss of power caused by compliance should also be considered when planning the sample size of a trial which aims to demonstrate treatment efficacy. Given that it will rarely be possible to regain the associated lost power, the initial sample size of a trial should incorporate an inflation factor based on realistic projections of relevant forms of treatment non-compliance (53).

Snapinn (179) demonstrates how informative non-compliance impacts on sample size and power, discussing different methods developed to allow for likely non-compliance rates when planning trial sample sizes. Snapinn describes how the majority of the sample size methods available assume only treatment switches to the alternative treatment and all assume that such discontinuation is independent of outcome (i.e. uninformative). Snapinn argues that this (latter) assumption can lead to greatly underestimated sample sizes, because it is not the rate of non-compliance *per se*, but rather the proportion of *endpoints* occurring in non-compliant patients, that impacts on power; Snapinn goes on to demonstrate an alternative method of determining sample size which allows for informative dropout (205).

White (69) considers whether it is possible to regain the lost power in an ITT analysis caused by departures from randomised treatment without making assumptions about selection effects (which are implied, for example, when using PP or AT analyses in an attempt to regain power), and discusses two possible ways to achieve this: firstly by using a modelling approach to impute missing compliance-types using the observed outcomes (206, 207) rather than simple ITT analyses; White notes, however, that the gain in power stems from distributional assumptions, and as such these methods may be highly sensitive to departures from the distributional assumptions. Secondly, covariates that predict compliance may be used to regain power (for example, in CACE analyses or structural mean models) (144).

In survival analysis, a related idea of weighted analysis, where earlier event times are assigned more weight, may provide extra power; however, in practice, this gain in power is unlikely to be substantial. Lagakos (208) investigated this issue when a constant HR among compliers yields a non-constant ITT HR (because of early termination of treatment among non-compliers); although the inefficiency of the ordinary log rank test can be improved in these situations (with weights dependent on the amount and pattern of early termination), the efficiency gains were small in many of the scenarios investigated.

11.6. Trial designs to address deviation from randomised treatment

Dunn (58) points out that nonadherence is not always due to a lack of cooperation on the part of the patient or a flaw in the design or methodology of a trial; indeed, nonadherence issues may be inherent to the disease and treatment process, and if foreseen, may be incorporated into the design of a trial during the planning stage, for

example in the case of encouragement or Zelen's RCD trials, both of which incorporate the possibility of all-or-nothing compliance by patients in the experimental arm.

In Zelen's randomised consent design (209), introduced to improve trial recruitment, the usual order of consent followed by randomisation is reversed, so that eligible patients are first randomised into one of the treatment arms, and then their consent is sought regarding their willingness to receive this randomised treatment.

Similarly, encouragement designs may be employed if the consent process is likely to lead to adherence problems, either because patients may not accept idea of randomisation in the given setting or if informing patients about treatment may then affect adherence of patients who end up in control arm. For example, if it is considered unethical to withhold treatment from one treatment arm, it may be possible to instead randomise *incentives* for treatment, such that one group receives no incentive different to usual (introducing no ethical problem) while the other group receives a treatment incentive (for example, in (127)).

Other simple design features may be incorporated to aid adherence analyses; for example, compliance measurements may be taken during a baseline placebo run in phase to obtain information on baseline predictors of compliance (noting that such a placebo run in phase will be useful in determining *which* of these variables should be recorded because, although the magnitude of relationship between such variables and placebo/treatment compliance may differ, the actual variables which correlate with placebo and treatment compliance are likely to be the same (144)), or it may be possible to seek out patient preferences before they are randomised, in order to obtain information on preference effects for better prediction of underlying compliance (210).

Likewise, if a trial is likely to involve numerous forms of treatment changes (for example, if a variety of treatment options are available on disease progression) or if extreme degrees of nonadherence are expected (for example, when randomisation is unlikely to be acceptable to most patients), it may be helpful to pre-empt the problem by trying to design the trial such that specific sequences of treatment are assigned from randomisation, or repeat randomisations take place as and when patients require different treatment lines. However, such designs may themselves be subject to nonadherence (when clinicians prescribe alternatives to the randomised treatment sequences) and may be overly complex or lengthy.

SMART designs

Sequential multiple assignment randomised trials (SMART) for the estimation of dynamic treatment regimens, in order to better estimate the optimal treatment package (sequence of treatments) for *individuals* rather than for *diagnosis*, are of particular importance in chronic conditions where individual patient's response may change over time, thus seeking out *personalised medicine* tailored to suit patients (for example, according to their genetic profile or at a more macro level of characteristics, such as side effect profiles) (211). Using functions with covariates, treatment and response history as arguments, with outputs of which actions should be taken (in other words, the decision rule on how to choose treatments rather than an "optimal" actual treatment) these designs assume, in a similar manner to IPCW and SMMs, that one has access to all known covariates which predict outcome and treatment changes (for example, patient and clinician preferences).

SMART designs are particularly appropriate for diseases where sequential phases of treatment are common, for example chronic conditions like asthma, epilepsy and cancer (requiring different first- and second-line treatments following diagnosis and

progression, respectively) and behavioural or psychological interventions. For example, the efficacy of warfarin (in treating thrombosis) is highly dependent on individual patient factors (diet, genetic variants and interactions with other medication), with serious side effects if inappropriate dosing; as such dose needs to be monitored regularly. Use of SMART designs in such settings is complicated, however, by the need to take account of repeated continuous dose data.

Depending on the particular trial characteristics, SMART designs may therefore result in better retention of patients (given the availability of multiple treatment options for patients while in the trial) and greater generalisability to clinical care (as the adaptive design and treatment flexibility allows for wider recruitment than in general) but typically involve long term follow up and large sample sizes, with associated high costs. Furthermore this relatively new and unfamiliar trial design is often met with reticence from funders and statisticians; thus it will be important for this growing area of research into study design to continue to be explained and explored more fully, in order to maximise its potential use in addressing adherence issues.

11.7. Software

Finally, it is important for analysts to consider whether software is available for their chosen analysis methods or whether specialist help and programming will be required. Certain methods (such as RPSFTM "strbee" code in Stata for survival outcomes adjusting for switching between trial treatments) are available as statistical programs that are user-friendly and easy to implement without specialist statistical knowledge or programming. IPCW methods are also becoming increasingly incorporated into statistical programs (for example, IPCW methods have been included in the latest update of Stata); however, there may be extra coding requirements required for

implementation of these methods depending on the complexity presented by the trial data (as in the case of these applications in SANAD).

11.8. Conclusion

This thesis has demonstrated how an ITT effect estimate in the presence of treatment deviations will be biased as an estimate of treatment efficacy or effectiveness in a setting different from that in the trial. There is therefore often a need for other causal analysis methods to supplement or replace ITT analysis, depending on the research question, pattern of treatment deviations, trial design and particular outcome of interest.

Despite the importance and increasing pertinence of causal analysis estimation in RCTs, however, whenever causal methodology is employed, analysts must remember that such analyses are not intended to replace or overrule analysis according to ITT; instead causal methodology seeks to complement analysis by randomisation, by providing greater insight and additional evidence regarding treatment efficacy. As encapsulated by Efron (39), "None of this spells the end of intent-to-treat analysis. It is and will remain a steadfast beacon in the foggy vistas of biomedical experimentation, but old friends need not feel abandoned when exciting new possibilities are explored".

References

1. Wright EC. Non-compliance--or how many aunts has Matilda? Lancet. 1993;342(8876):909.
2. Sackett D, Haynes B, editors. Compliance with therapeutic regimens. Baltimore: Johns Hopkins University Press; 1976.
3. Vrijens B, De Geest S, Hughes DA, Przemyslaw K, Demonceau J, Ruppar T, et al. A new taxonomy for describing and defining adherence to medications. British Journal of Clinical Pharmacology. 2012;73(5):691-705.
4. Donovan J. Patient Decision Making: The Missing Ingredient in Compliance Research. International Journal of Technology Assessment in Health Care. 1995;11(3):443-55.
5. World Health Organisation. Adherence to Long-term Therapies: Evidence for Action. E S, editor. Geneva: World Health Organisation; 2003 07 April 2003. 212 p.
6. Haynes RB, McDonald HP, Garg AX. Helping Patients Follow Prescribed Treatment. JAMA: The Journal of the American Medical Association. 2002;288(22):2880-3.
7. L L, PB H. Health Care, Research and Regulatory Impact of Noncompliance. In: JA C, B S, editors. Patient Compliance in Medical Practice and Clinical Trials. New York: Raven Press; 1991. p. 393-403.
8. Everitt BCP, editor. Encyclopaedic companion to medical statistics: Hodder Arnold; 2005.
9. Vermeire E, Hearnshaw H, PV R, J D. Patient adherence to treatment:three decades of research. A comprehensive review. Journal of Clinical Pharmacy and Therapeutics. 2001;26:331-42.
10. Carter S, Taylor D, Levenson R. A question of choice — compliance in medicine taking. 2005.
11. Haynes RB, Dantes R. Patient compliance and the conduct and interpretation of therapeutic trials. Controlled Clinical Trials. 1987;8(1):12-9.
12. DiMatteo R, Chow MSS. Patient adherence to pharmacotherapy: The importance of effective communication. Formulary. 1995;30(10):596.
13. Bond W, Hussar D. Detection methods and strategies for improving medication compliance. American Journal of Hospital Pharmacy. 1991;48:1978-88.
14. Osterberg L, Blaschke T. Adherence to Medication. New England Journal of Medicine. 2005;353:487-97.
15. Urquhart J, De Klerk E. Contending paradigms for the interpretation of data on patient compliance with therapeutic drug regimens. Statistics in Medicine. 1998;17(3):251-67; discussion 387-9.
16. Escalada P, Griffiths P. Do people with cancer comply with oral chemotherapy treatments? British Journal of Community Nursing. 2006;11(12):532 - 6
17. Düsing R, Lottermoser K, Mengden T. Compliance with drug therapy—new answers to an old question. Nephrology Dialysis Transplantation. 2001;16(7):1317-21.
18. Jin J, Sklar GE, Oh VMS, Li SC. Factors affecting therapeutic compliance: A review from the patient's perspective. Therapeutics and Clinical Risk Management. 2008;4(1):269-86.
19. Hughes DA, Bagust A, Haycox A, Walley T. The impact of non-compliance on the cost-effectiveness of pharmaceuticals: a review of the literature. Health Economics. 2001;10(7):601-15.

20. Vrijens B, Urquhart J. Patient adherence to prescribed antimicrobial drug dosing regimens. Journal of Antimicrobial Chemotherapy. 2005;55(5):616-27.
21. Nunes V NJ, O'Flynn N, Calvert N, Kuntze S, Smithson H, Benson J, Blair J, Bowser A, Clyne W, Crome P, Haddad P, Hemingway S, Horne R, Johnson S, Kelly S, Packham B, Patel M, Steel J. Clinical Guidelines and Evidence Review for Medicines Adherence: involving patients in decisions about prescribed medicines and supporting adherence. London: National Collaborating Centre for Primary Care and Royal College of General Practitioners; 2009.
22. Urquhart J. Patient non-compliance with drug regimens:measurement, clinical correlates, economic impact. European Heart Journal. 1996;17(Supplement A):8-15.
23. Meier P. Compliance as an Explanatory Variable in Clinical Trials: Comment. Journal of the American Statistical Association. 1991;86(413):19-22.
24. Dusing R, Weisser B, Mengden T, H V. Changes in Antihypertensive Therapy - The Role of Adverse Effects and Compliance. Blood Pressure. 1998;7(5-6):313-5.
25. Taylor SA, Galbraith SM, Mills RP. Causes Of Non-Compliance With Drug Regimens In Glaucoma Patients: A Qualitative Study. Journal of Ocular Pharmacology and Therapeutics. 2002;18(5):401-9.
26. Townsend A, Hunt K, Wyke S. Managing multiple morbidity in mid-life: a qualitative study of attitudes to drug use. BMJ. 2003;327(7419):837.
27. Blackwell B. Counselling and compliance. Patient Counselling and Health Education. 1978;1(2):45-9.
28. Zolnierek K, DiMatteo M. Physician Communication and Patient Adherence to Treatment: A Meta-Analysis. Medical Care August. 2009;47(8):826-34.
29. Tarn DM, Heritage J, Paterniti DA, Hays RD, Kravitz RL, Wenger NS. Physician Communication When Prescribing New Medications. Arch Intern Med. 2006;166(17):1855-62.
30. Becker M, Maiman L. Sociobehavioural determinants of compliance with health and medical care recommendations. Medical Care. 1975;13:10-24.
31. Russell S, Daly J, Hughes E, Hoog Cot. Nurses and 'difficult' patients: negotiating non-compliance. Journal of Advanced Nursing. 2003;43(3):281-7.
32. Mullen P. Compliance becomes concordance. BMJ. 1997;314(7082):691.
33. Heath I. A wolf in sheep's clothing: a critical look at the ethics of drug taking. BMJ. 2003;327(7419):856-8.
34. Aronson JK. Time to abandon the term 'patient concordance'. Br J Clin Pharmacol 2007;64(5):711–3.
35. Serebruany V, Oshrine B, Alex I, Atar D, Michelson A, JJ F. Noncompliance in cardiovascular clinical trials. American Heart Journal. 2005;150:882-6.
36. Vrijens B. Presentation of draft ABC proposal on taxonomy and terminology 2009. Available from: http://www.espacomp.eu/file/presentations/Bernard%20Vrijens_ppt%20.pdf.
37. Hughes DA, Bagust A, Haycox A and Walley T. Accounting for Noncompliance in Pharmacoeconomic Evaluations. Pharmacoeconomics. 2001;19(12):1185-97.
38. Haynes RB, Ackloo E, Sahota N, McDonald HP, Yao X. Interventions for enhancing medication adherence. 2008.
39. Efron B. Foreword: Limburg Compliance Symposium. Statistics in Medicine. 1998;17(3):249-50.
40. Rabeneck L VC, Horwitz RI. Problems in the conduct and analysis of randomized clinical trials: are we getting the right answers to the wrong questions? Arch Intern Med. 1992;152(3):507-12.

41. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, et al. CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340.
42. Godwin M, Ruhland L, Casson I, MacDonald S, Delva D, Birtwhistle R, et al. Pragmatic controlled clinical trials in primary care: the struggle between external and internal validity. BMC Medical Research Methodology. 2003;3(1):28.
43. May G, DeMets D, Friedman L, Furberg C, Passamani E. The randomized clinical trial: bias in analysis. Circulation. 1981;64(4):669-73.
44. Revicki DA, Frank L. Pharmacoeconomic Evaluation in the Real World: Effectiveness Versus Efficacy Studies. Pharmacoeconomics. 1999;15(5):423-34.
45. DeMets D, Friedman L, Furberg C. Counting Events in Clinical Trials (Correspondence). New England Journal of Medicine. 1980;302(16):924-5.
46. Devereaux PJ, Manns BJ, Ghali WA, Quan H, Lacchetti C, Montori VM, et al. Physician Interpretations and Textbook Definitions of Blinding Terminology in Randomized Controlled Trials. JAMA. 2001;285(15):2000-3.
47. Sommer A, Zeger SL. On estimating efficacy from clinical trials. Statistics in Medicine. 1991;10(1):45-52.
48. Gartlehner G, Hansen RA, Nissman D, Lohr KN, Carey TS. A simple and valid tool distinguished efficacy from effectiveness studies. Journal of Clinical Epidemiology. 2006;59(10):1040-8.
49. Ioannidis JPA, Evans SJW, Gøtzsche PC, O'Neill RT, Altman DG, Schulz K, et al. Better Reporting of Harms in Randomized Trials: An Extension of the CONSORT Statement. Annals of Internal Medicine. 2004;141(10):781-8.
50. Roland M, Torgerson DJ. Understanding controlled trials: What are pragmatic trials? BMJ. 1998;316(7127):285-.
51. Porta N, Bonet C, Cobo E. Discordance between reported intention-to-treat and per protocol analyses. Journal of Clinical Epidemiology. 2007;60(7):663-9.
52. Sheiner LB, Rubin DB. Intention-to-treat analysis and the goals of clinical trials. Clin Pharmacol Ther. 1995;57(1):6-15.
53. Newcombe RG. Explanatory and pragmatic estimates of the treatment effect when deviations from allocated treatment occur. Statistics in Medicine. 1988;7(11):1179-86.
54. Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutical trials. Journal of Chronic Diseases. 1967;20(8):637-48.
55. Bratton DJ, Nunn AJ. Alternative approaches to tuberculosis treatment evaluation: the role of pragmatic trials [State of the art series. Operational research. Number 4 in the series]. The International Journal of Tuberculosis and Lung Disease. 2011;15(4):440-6.
56. Lagakos SW, Lim LL, Robins JM. Adjusting for early treatment termination in comparative clinical trials. Stat Med. 1990;9(12):1417-24; discussion 33-7.
57. Sheiner LB, Rubin DB. Intention-to-treat analysis and the goals of clinical trials[ast]. Clin Pharmacol Ther. 1995;57(1):6-15.
58. Dunn G, Maracy M, Tomenson B. Estimating treatment effects from randomized clinical trials with noncompliance and loss to follow-up: the role of instrumental variable methods. Statistical Methods in Medical Research. 2005;14(4):369-95.
59. Marson AG, Williamson PR. Interpreting regulatory trials in epilepsy. Current Opinion in Neurology. 2009;22(2):167.
60. Simon G, Wagner E, Vonkorff M. Cost-effectiveness comparisons using "real world" randomized trials: The case of new antidepressant drugs. Journal of Clinical Epidemiology. 1995;48(3):363-73.

61. International Conference on Harmonisation E9 Expert Working Group. ICH Harmonised Tripartite Guideline. Statistical principles for clinical trials. Statistics in Medicine. 1999;18(15):1905-42.
62. Dodd S, White IR, Williamson P. Nonadherence to treatment protocol in published randomised controlled trials: a review. Trials. 2012;13(1):84-99.
63. Emsley R, Dunn G, White IR. Mediation and moderation of treatment effects in randomised controlled trials of complex interventions. Statistical Methods in Medical Research. 2010;19(3):237-70.
64. Epstein LH. The direct effects of compliance on health outcome. Health Psychology. 1984;3(4):385-93.
65. Coronary Drug Project Research Group. Influence of adherence to treatment and response of cholesterol on mortality in the Coronary Drug Project New England Journal of Medicine. 1980;303:1038-41.
66. Horwitz RI, Viscoli CM, Donaldson RM, Murray CJ, Ransohoff DF, Horwitz RI, et al. Treatment adherence and risk of death after a myocardial infarction. The Lancet. 1990;336(8714):542-5.
67. Chêne G, Morlat P, Leport C, Hafner R, Dequae L, Charreau I, et al. Intention-to-Treat vs. On-Treatment Analyses of Clinical Trial Data: Experience from a Study of Pyrimethamine in the Primary Prophylaxis of Toxoplasmosis in HIV-Infected Patients. Controlled Clinical Trials. 1998;19(3):233-48.
68. White IR PS. Statistical reporting of clinical trials with individual changes from allocated treatment. Statistics in Medicine. 1996;15(3):249-62.
69. White IR. Uses and limitations of randomization-based efficacy estimators. Statistical Methods in Medical Research. 2005;14:327-47.
70. Cox D. Discussion. Statistics in Medicine. 1998;17(3):387-9.
71. Pocock SJ, Abdalla M. The hope and the hazards of using compliance data in randomized controlled trials. Statistics in Medicine. 1998;17(3):303-17.
72. Kim LG, White IR. Compliance-adjusted intervention effects in survival data. The Stata Journal. 2004;4(3):257-64.
73. White IR DG. Adjustment for non-compliance in randomised controlled trials. In: Everitt BS PC, editor. London: Hodder Arnold; 2005.
74. Walker S, White IR, Babiker AG. Parametric randomization-based methods for correcting for treatment changes in the assessment of the causal effect of treatment. Statistics in Medicine. 2004;23:571-90.
75. Lewis JA, Machin D. Intention to treat – who should use ITT? Br J Cancer. 1993;68(4):647-50.
76. International Conference on Harmonisation E3 Expert Working Group. ICH Harmonised Tripartite Guideline. Structure and content of clinical study reports. 1995.
77. Baker SG, Kramer BS. Simple maximum likelihood estimates of efficacy in randomized trials and before-and-after studies, with implications for meta-analysis. Statistical Methods in Medical Research. 2005;14:349-67.
78. Hughes DA, Cowell W, Koncz T, Cramer J. Methods for Integrating Medication Compliance and Persistence in Pharmacoeconomic Evaluations. Value in Health. 2007;10(6):498–509.
79. Consortium TIWP. Estimation of the Warfarin Dose with Clinical and Pharmacogenetic Data. New England Journal of Medicine. 2009;360(8):753.
80. Mega JL, Simon T, Collet J-P, Anderson JL, Antman EM, Bliden K, et al. Reduced-Function CYP2C19 Genotype and Risk of Adverse Clinical Outcomes Among Patients Treated With Clopidogrel Predominantly for PCI : A Meta-analysis. JAMA: Journal of the American Medical Association. 2010;304(16):1821.

81. Mallal S, Phillips E, Carosi G, Molina J-M, Workman C, Toma, et al. HLA-B 5701 Screening for Hypersensitivity to Abacavir. New England Journal of Medicine. 2008;358(6):568.
82. McCormack M, Alfirevic A, Bourgeois S, Farrell JJ, Kasperavi, Dalia, et al. HLA-A 3101 and Carbamazepine-Induced Hypersensitivity Reactions in Europeans. New England Journal of Medicine. 2011;364(12):1134.
83. Sharma SP. Avastin Saga Reveals Debate Over Clinical Trial Endpoints. Journal of the National Cancer Institute. 2012;104(11):800-1.
84. Holland PW. Statistics and Causal Inference. Journal of the American Statistical Association. 1986;81(396):945-60.
85. Rubin DB. [On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.] Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies. Statistical Science. 1990;5(4):472-80.
86. Hernan MA, Hernandez-Diaz S, Robins JM. A Structural Approach to Selection Bias. Epidemiology. 2004;15(5):615-25.
87. Hernán MA, Hernández-Díaz S. Beyond the intention-to-treat in comparative effectiveness research. Clinical Trials. 2012;9(1):48-55.
88. Ten Have TRP, Normand S-LTP, Marcus SMP, Brown CHP, Lavori PP, Duan NP. Intent-to-treat vs. Non-intent-to-treat Analyses under Treatment Non-adherence in Mental Health Randomized Trials. Psychiatric Annals. 2008;38(12):772-83.
89. White IR GE. Clinical trials comparing two treatment policies: which aspects of the treatment policies make a difference? Statistics in Medicine. 1998;17(3):319-39.
90. Yusuf S, Collins R, Peto R. Why do we need some large, simple randomized trials? Statistics in Medicine. 1984;3(4):409-20.
91. Sommer A, Djunaedi E, Loeden AA, Tarwotjo I, West Jr K, Tilden R, et al. Impact of Vitamin A supplementation on childhood mortality: a Randomised Controlled Community Trial. The Lancet. 1986;327(8491):1169-73.
92. Goetghebeur EJT, Pocock SJ. Statistical Issues in Allowing for Noncompliance and Withdrawal. Drug Information Journal. 1993;27(3):837-45.
93. MRC Working Party. Medical Research Council trial of treatment of hypertension in older adults: principal results. . BMJ. 1992;304(6824):405-12.
94. White IR, Babiker AG, Walker S, Darbyshire JH. Randomization-based methods for correcting for treatment changes: examples from the Concorde trial. Statistics in Medicine. 1999;18:2617-34.
95. Robins JM, Greenland S. Adjusting for Differential Rates of Prophylaxis Therapy for PCP in High-Versus Low-Dose AZT Treatment Arms in an AIDS Randomized Trial. Journal of the American Statistical Association. 1994;89(427):737-49.
96. Booth CM, Eisenhauer EA. Progression-Free Survival: Meaningful or Simply Measurable? Journal of Clinical Oncology. 2012;30(10):1030-3.
97. Dancey JE, Dodd LE, Ford R, Kaplan R, Mooney M, Rubinstein L, et al. Recommendations for the assessment of progression in randomised cancer treatment trials. European Journal of Cancer. 2009;45(2):281-9.
98. Robson V, Dodd S, Thomas S. Standardized antibacterial honey (Medihoney™) with standard therapy in wound care: randomized clinical trial. Journal of Advanced Nursing. 2009;65(3):565-75.
99. Marson AJ, et al. The SANAD study of eff ectiveness of carbamazepine, gabapentin, lamotrigine, oxcarbazepine, or topiramate for treatment of partial epilepsy: an unblinded randomised controlled trial. Lancet. 2007;369:1000-15.

100. Marson AJ, et al. The SANAD study of eff ectiveness of valproate, lamotrigine or topiramate for generalised and unclassifi able epilepsy: an unblinded randomised controlled trial. Lancet. 2007;369:1016-26.
101. Powell C, Kolamunnage-Dona R, Lowe J, Boland A, Petrou S, Doull I, et al. Magnesium sulphate in acute severe asthma in children (MAGNETIC): a randomised, placebo-controlled trial. The Lancet Respiratory Medicine. 2013;1(4):301-8.
102. Powell C, Kolamunnage-Dona R, Lowe J, Boland A, Petrou S, Doull I, et al. MAGNEsium Trial In Children (MAGNETIC): a randomised, placebo controlled trial and economic evaluation of nebulised magnesium sulphate in acute severe asthma in children. Health Technol Assess. 2013;17(45).
103. Maldonado G, Greenland S. Estimating causal effects. International Journal of Epidemiology. 2002;31(2):422-9.
104. Pearl J. Causal Inference in the Health Sciences: A Conceptual Introduction. Health Services and Outcomes Research Methodology. 2001;2(3):189-220.
105. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology. 1974;66(5):688-701.
106. Sobel ME. Causal Inference in the Social Sciences. Journal of the American Statistical Association. 2000;95(450):647-51.
107. Toh S HM. Causal Inference from Longitudinal Studies with Baseline Randomization. International Journal of Biostatistics. 2008;4(1):Article 22.
108. Murphy SA. Optimal dynamic treatment regimes. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2003;65(2):331-55.
109. Little RJ, Rubin DB. Causal Effects in Clinical and Epidemiological Studies Via Potential Outcomes: Concepts and Analytical Approaches. Annual Review of Public Health. 2000;21(1):121-45.
110. Angrist JD, Imbens GW. Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity. Journal of the American Statistical Association. 1995;90(430):431-42.
111. Robins JM. Marginal Structural Models versus Structural Nested Models as Tools for Causal Inference. In: Halloran M.E. BD, editor. Statistical Models in Epidemiology: The Environment and Clinical Trials. New York: Springer; 1999. p. 95-134.
112. Bellamy SL, Lin JY, Have TRT. An introduction to causal modeling in clinical trials. Clinical Trials. 2007;4(1):58-73.
113. Shrier I, Steele RJ, Verhagen E, Herbert R, Riddell CA, Kaufman JS. Beyond intention to treat: What is the right question? Clinical Trials. 2014;11(1):28-37.
114. Heitjan DF. Ignorability and bias in clinical trials. Statistics in Medicine. 1999;18(17-18):2421-34.
115. Goetghebeur E, Molenberghs G. Causal Inference in a Placebo-Controlled Clinical Trial With Binary Outcome and Ordered Compliance. Journal of the American Statistical Association. 1996;91(435):928-34.
116. Vansteelandt S, Goetghebeur E. Sense and sensitivity when correcting for observed exposures in randomized clinical trials. Statistics in Medicine. 2005;24(2):191-210.
117. Angrist JD, Imbens GW, Rubin DB. Identification of Causal Effects Using Instrumental Variables. Journal of the American Statistical Association. 1996;91(434):444-55.
118. Hernán MA, Lanoy E, Costagliola D, Robins JM. Comparison of Dynamic Treatment Regimes via Inverse Probability Weighting. Basic & Clinical Pharmacology & Toxicology. 2006;98(3):237-42.

119. Little RJ, Long Q, Lin X. A Comparison of Methods for Estimating the Causal Effect of a Treatment in Randomized Clinical Trials Subject to Noncompliance. Biometrics. 2009;65(2):640-9.
120. Little RJ, Yau LHY. Statistical techniques for analyzing data from prevention trials: Treatment of no-shows using Rubin's causal model. Psychological Methods. 1998;3(2):147-59.
121. Frangakis CE, Rubin DB. Addressing Complications of Intention-to-Treat Analysis in the Combined Presence of All-or-None Treatment-Noncompliance and Subsequent Missing Outcomes. Biometrika. 1999;86(2):365-79.
122. Peng Y, Little RJA, Raghunathan TE. An Extended General Location Model for Causal Inferences from Data Subject to Noncompliance and Missing Values. Biometrics. 2004;60(3):598-607.
123. Nagelkerke N, Fidler V, Bernsen R, Borgdorff M. Estimating treatment effects in randomized clinical trials in the presence of non-compliance.[erratum appears in Stat Med 2001 Mar 30;20(6):982]. Statistics in Medicine. 2000;19(14):1849-64.
124. Greenland S. An introduction to instrumental variables for epidemiologists. International Journal of Epidemiology. 2000;29(4):722-9.
125. Frangakis CE, Brookmeyer RS, Varadhan R, Safaeian M, Vlahov D, Strathdee SA. Methodology for Evaluating a Partially Controlled Longitudinal Treatment Using Principal Stratification, with Application to a Needle Exchange Program. Journal of the American Statistical Association. 2004;99(465):239-49.
126. Imbens GW, Rubin DB. Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance. The Annals of Statistics. 1997;25(1):305-27.
127. Hirano K, Imbens GW, Rubin DB, Zhou X-H. Assessing the effect of an influenza vaccine in an encouragement design. Biostatistics. 2000;1(1):69-88.
128. Robins J. The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In: L S, H F, A M, editors. Health Service Research Methodology: A Focus on AIDS: NCHSR, US Public Health Service; 1989. p. 113-59.
129. Robins JM, Tsiatis AA. Correcting for non-compliance in randomized trials using rank preserving structural failure time models. Communications in Statistics: Theory & Methods. 1991;20(8):2609-31.
130. Hernán MA, Cole SR, Margolick J, Cohen M, Robins JM. Structural accelerated failure time models for survival analysis in studies with time-varying treatments. Pharmacoepidemiology and Drug Safety. 2005;14(7):477-91.
131. Robins J, Tsiatis AA. Semiparametric Estimation of an Accelerated Failure Time Model with Time- Dependent Covariates. Biometrika. 1992;79(2):311-9.
132. Fewell Z, Hernán MA, Wolfe F, Tilling K, Choi H, Sterne JAC. Controlling for time-dependent confounding using marginal structural models. Stata Journal. 2004;4(4):402-20.
133. Collett D. Modelling Survival Data in Medical Research. New York: Chapman and Hall/CRC; 2003.
134. White IR, Walker S, Babiker A. strbee: Randomization-based efficacy estimator. The Stata Journal. 2002;2 (2):140-50.
135. Latimer NR, Abrams KR, Lambert PC, Crowther MJ, Wailoo AJ, Morden JP, et al. Adjusting Survival Time Estimates to Account for Treatment Switching in Randomized Controlled Trials—an Economic Evaluation Context: Methods, Limitations, and Recommendations. Medical Decision Making. 2014;34(3):387-402.
136. Yamaguchi T, Ohashi Y. Adjusting for differential proportions of second-line treatment in cancer clinical trials. Part I: Structural nested models and marginal

structural models to test and estimate treatment arm effects. Statistics in Medicine. 2004;23(13):1991-2003.
137. Morden J, Lambert P, Latimer N, Abrams K, Wailoo A. Assessing methods for dealing with treatment switching in randomised controlled trials: a simulation study. BMC Medical Research Methodology. 2011;11(1):4.
138. Mark SD, Robins JM. A method for the analysis of randomized trials with compliance information: An application to the multiple risk factor intervention trial. Controlled Clinical Trials. 1993;14(2):79-97.
139. Goetghebeur E, Lapp K. The Effect of Treatment Compliance in a Placebo-controlled Trial: Regression with Unpaired Data. Journal of the Royal Statistical Society: Series C (Applied Statistics). 1997;46(3):351-64.
140. Tilling K, Sterne JAC, Didelez V. G-estimation for Accelerated Failure Time Models. In: Tu Y-K, Greenwood DC, editors. Modern Methods for Epidemiology. Dordrecht: Springer Science+Business Media; 2012.
141. White IR, Walker S, Babiker AG and Darbyshire JH. Impact of treatment changes on the interpretation of the Concorde trial. AIDS. 1997;11:999-1006.
142. Latimer N, KR A. NICE DSU Technical Support Document 16: Adjusting survival time estimates in the presence of treatment switching. 2014.
143. Robins J, Rotnitzky A. Estimation of Treatment Effects in Randomised Trials with Non-Compliance and a Dichotomous Outcome Using Structural Mean Models. Biometrika. 2004;91(4):763-83.
144. Fischer-Lapp K, Goetghebeur E. Practical Properties of Some Structural Mean Analyses of the Effect of Compliance in Randomized Trials. Controlled Clinical Trials. 1999;20(6):531-46.
145. Robins JA. Correcting for non-compliance in randomized trials using structural nested mean models. Communications in Statistics: Theory & Methods. 1994;23:2379-412.
146. Vansteelandt S, Goetghebeur E. Causal Inference with Generalized Structural Mean Models. Journal of the Royal Statistical Society Series B (Statistical Methodology). 2003;65(4):817-35.
147. Ten Have TR, Elliott MR, Joffe M, Zanutto E, Datto C. Causal Models for Randomized Physician Encouragement Trials in Treating Primary Care Depression. Journal of the American Statistical Association. 2004;99(465):16-25.
148. Goetghebeur E, Vansteelandt S. Structural mean models for compliance analysis in randomized clinical trials and the impact of errors on measures of exposure. Statistical Methods in Medical Research. 2005;14(4):397-415.
149. Robins J. Marginal structural models versus structural nested models as tools for causal inference. In: ME H, DA B, editors. Statistical Models in Epidemiology, the Environment and Clinical Trials. New York: Springer; 1999. p. 95-133.
150. Robins JM, Finkelstein DM. Correcting for Noncompliance and Dependent Censoring in an AIDS Clinical Trial with Inverse Probability of Censoring Weighted (IPCW) Log-Rank Tests. Biometrics. 2000;56(3):779-88.
151. Matsuyama Y, Yamaguchi T. Estimation of the marginal survival time in the presence of dependent competing risks using inverse probability of censoring weighted (IPCW) methods. Pharmaceutical Statistics. 2008;7(3):202-14.
152. Hernan MA, Brumback B, Robins JM. Marginal Structural Models to Estimate the Joint Causal Effect of Nonrandomized Treatments. Journal of the American Statistical Association. 2001;96(454):440-8.
153. DerSimonian R, Charette LJ, McPeek B, Mosteller F. Reporting on Methods in Clinical Trials. New England Journal of Medicine. 1982;306(22):1332-7.

154. Freiman JA, Chalmers TC, Smith H, Kuebler RR. The Importance of Beta, the Type II Error and Sample Size in the Design and Interpretation of the Randomized Control Trial. New England Journal of Medicine. 1978;299(13):690-4.
155. Gøtzsche PC. Multiple publication of reports of drug trials. European Journal of Clinical Pharmacology. 1989;36(5):429-32.
156. Gøtzsche PC. Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal antiinflammatory drugs in rheumatoid arthritis. Controlled Clinical Trials. 1989;10(1):31-56.
157. Meinert CL, Tonascia S, Higgins K. Content of reports on clinical trials: A critical review. Controlled Clinical Trials. 1984;5(4, Supplement 1):328-47.
158. Mosteller F. Problems of omission in communications. Clin Pharmacol Ther. 1979;25:761-4.
159. Reiffenstein RJ, Schiltroth AJ, Todd DM. Current standards in reported drug trials. CMAJ. 1968;99(23):1134-5.
160. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Annals of Internal Medicine. 2001;134(8):663-94.
161. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, et al. Improving the Quality of Reporting of Randomized Controlled Trials. JAMA: The Journal of the American Medical Association. 1996;276(8):637-9.
162. Toerien M ea. A review of reporting of participant recruitment and retention in RCTs in six major journals. Trials. 2009;10(52).
163. Wood AM, White IR, Thompson SG. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. Clinical Trials. 2004;1:368-76.
164. Heiss G, Wallace R, Anderson GL, Aragaki A, Beresford SAA, Brzyski R, et al. Health risks and benefits 3 years after stopping randomized treatment with estrogen and progestin. JAMA. 2008;299(9):1036-45.
165. Tyrer P, Oliver-Africano PC, Ahmed Z, Bouras N, Cooray S, Deb S, et al. Risperidone, haloperidol, and placebo in the treatment of aggressive challenging behaviour in patients with intellectual disability: a randomised controlled trial.[see comment]. Lancet. 2008;371(9606):57-63.
166. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical Evidence of Bias. JAMA: The Journal of the American Medical Association. 1995;273(5):408-12.
167. Schulz KF, Grimes DA, Altman DG, Hayes RJ. Blinding and exclusions after allocation in randomised controlled trials: survey of published parallel group trials in obstetrics and gynaecology. BMJ. 1996;312(7033):742-4.
168. Schulz KF, Grimes DA. Sample size slippages in randomised trials: exclusions and the lost and wayward. The Lancet. 2002;359(9308):781-5.
169. Schulz KF, Altman DG, Moher D. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340.
170. Molina J-M, Andrade-Villanueva J, Echevarria J, Chetchotisakd P, Corral J, David N, et al. Once-daily atazanavir/ritonavir versus twice-daily lopinavir/ritonavir, each in combination with tenofovir and emtricitabine, for management of antiretroviral-naive HIV-1-infected patients: 48 week efficacy and safety results of the CASTLE study.[see comment]. Lancet. 2008;372(9639):646-55.
171. Supplementation with Multiple Micronutrients Intervention Trial Study G, Shankar AH, Jahari AB, Sebayang SK, Aditiawarman, Apriatni M, et al. Effect of maternal multiple micronutrient supplementation on fetal loss and infant death in Indonesia: a double-blind cluster-randomised trial.[see comment]. Lancet. 2008;371(9608):215-27.

172. Russell JA, Walley KR, Singer J, Gordon AC, Hebert PC, Cooper DJ, et al. Vasopressin versus norepinephrine infusion in patients with septic shock. New England Journal of Medicine. 2008;358(9):877-87.
173. Fergusson D, Aaron SD, Guyatt G, Hebert P. Post-randomisation exclusions: the intention to treat principle and excluding patients from analysis. BMJ. 2002;325(7365):652-4.
174. Gravel J, Opatrny L, Shapiro S. The intention-to-treat approach in randomized controlled trials: Are authors saying what they do and doing what they say? Clinical Trials. 2007;4(4):350-6.
175. Herman A, Botser IB, Tenenbaum S, Chechick A. Intention-to-Treat Analysis and Accounting for Missing Data in Orthopaedic Randomized Clinical Trials. J Bone Joint Surg Am. 2009;91(9):2137-43.
176. Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. BMJ. 1999;319(7211):670-4.
177. Kruse RL, Alper BS, Reust C, Stevermer JJ, Shannon S, Williams RH. Intention-to-treat analysis: who is in? Who is out? J Fam Pract. 2002;51(11):969-71.
178. Gross D, Fogg L. A Critical Analysis of the Intent-to-Treat Principle in Prevention Research. The Journal of Primary Prevention. 2004;25(4):475-89.
179. Snapinn S, Jiang Q, Iglewicz B. Informative noncompliance in endpoint trials. Current Controlled Trials in Cardiovascular Medicine. 2004;5(1):5.
180. Peduzzi P, Wittes J, Detre K, Holford T. Analysis as-randomized and the problem of non-adherence: An example from the veterans affairs randomized trial of coronary artery bypass surgery. Statistics in Medicine. 1993;12(13):1185-95.
181. Dunn G, Maracy M, Tomenson B. Estimating treatment effects from randomized clinical trials with noncompliance and loss to follow-up: the role of instrumental variable methods. Statistical Methods in Medical Research. 2005;14:369-95.
182. Greenland S, Pearl J, Robins JM. Causal Diagrams for Epidemiologic Research. Epidemiology. 1999;10(1):37-48.
183. D'Agostino RB, Lee M-L, Belanger AJ, Cupples LA, Anderson K, Kannel WB. Relation of pooled logistic regression to time dependent cox regression analysis: The framingham heart study. Statistics in Medicine. 1990;9(12):1501-15.
184. Cole SR, Hernán MA, Robins JM, Anastos K, Chmiel J, Detels R, et al. Effect of Highly Active Antiretroviral Therapy on Time to Acquired Immunodeficiency Syndrome or Death using Marginal Structural Models. American Journal of Epidemiology. 2003;158(7):687-94.
185. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Statistics in Medicine. 2004;23(19):2937-60.
186. Williamson E, Morley R, Lucas A, Carpenter J. Propensity scores: From naïve enthusiasm to intuitive understanding. Statistical Methods in Medical Research. 2012;21(3):273-93.
187. Harrell Jr F. Regression Modelling Strategies. New York: Springer; 2001.
188. Durrleman S, Simon R. Flexible regression models with cubic splines. Statistics in Medicine. 1989;8(5):551-61.
189. Stone C. Additive Splines in Statistics. Proceedings of the Statistical Computing Section, American Statistical Association. 1986:45.
190. Hosmer DW, Lemeshow S. Applied logistic regression: New York: Wiley, 1989.; 1989.
191. Maldonado G, Greenland S. Simulation Study of Confounder-Selection Strategies. American Journal of Epidemiology. 1993;138(11):923-36.

192. Williamson PR TSC, Sander JW and Marson AG. Importance of competing risks in the analysis of anti-epileptic drug failure. Trials. 2007;8(12).
193. Royston P, Sauerbrei W. Improving the robustness of fractional polynomial models by preliminary covariate transformation: A pragmatic approach. Computational Statistics & Data Analysis. 2007;51(9):4240-53.
194. Marson AG AR, Baker GA, Chadwick DW, Doughty J, Eaton B, Gamble C, Jacoby A, Shackley P, Smith DF, Tudur-Smith C, Vanoli A, Williamson PR. A randomised controlled trial examining the longer-term outcomes of standard versus new antiepileptic drugs. The SANAD trial. Health Technology Assessment. 2007;11(37).
195. Ishak KJ, Proskorovsky I, Korytowsky B, Sandin R, Faivre S, Valle J. Methods for Adjusting for Bias Due to Crossover in Oncology Trials. PharmacoEconomics. 2014;32(6):533-46.
196. Howe CJ, Cole SR, Chmiel JS, Muñoz A. Limitation of Inverse Probability-of-Censoring Weights in Estimating Survival in the Presence of Strong Selection Bias. American Journal of Epidemiology. 2011;173(5):569-77.
197. Gruber S, Logan RW, Jarrín I, Monge S, Hernán MA. Ensemble learning of inverse probability weights for marginal structural modeling in large observational datasets. Statistics in Medicine. 2015;34(1):106-17.
198. Williamson PR K-DR, Philipson P, Marson AG. Joint modelling of longitudinal and competing risks data. Statistics in Medicine. 2008;27(30):6426-38.
199. Yamaguchi T, Ohashi Y. Adjusting for differential proportions of second-line treatment in cancer clinical trials. Part II: An application in a clinical trial of unresectable non-small-cell lung cancer. Statistics in Medicine. 2004;23(13):2005-22.
200. Robins JM. Correction for non-compliance in equivalence trials. Statistics in Medicine. 1998;17(3):269-302.
201. Greenland S, Lanes S, Jara M. Estimating effects from randomized trials with discontinuations: the need for intent-to-treat design and G-estimation. Clinical Trials. 2008;5:5-13.
202. National Institute for Health and Clinical Excellence. Sunitinib for the treatment of gastrointestinal stromal tumours. 2009.
203. National Institute for Health and Clinical Excellence. Everolimus for the second-line treatment of advanced renal cell carcinoma. 2010.
204. National Institute for Health and Clinical Excellence. Pazopanib for the first-line treatment of advanced renal cell carcinoma. 2011.
205. Jiang Q, Snapinn S, Iglewicz B. Calculation of Sample Size in Survival Trials: The Impact of Informative Noncompliance. Biometrics. 2004;60(3):800-6.
206. Imbens GW, Rubin DB. Estimating Outcome Distributions for Compliers in Instrumental Variables Models. The Review of Economic Studies. 1997;64(4):555-74.
207. Rubin DB. More powerful randomization-based p-values in double-blind trials with non-compliance. Statistics in Medicine. 1998;17(3):371-85.
208. Lagakos SW, Lim LLY, Robins JM. Adjusting for early treatment termination in comparative clinical trials. Statistics in Medicine. 1990;9(12):1417-24.
209. Zelen M. A New Design for Randomized Clinical Trials. New England Journal of Medicine. 1979;300(22):1242-5.
210. Dunn G, et al. Estimating psychological treatment effects from a randomised controlled trial with both non-compliance and loss to follow-up. British Journal of Psychiatry. 2003;183:323-31.
211. Shortreed SM, Moodie EEM. Estimating the optimal dynamic antipsychotic treatment regime: evidence from the sequential multiple-assignment randomized

Clinical Antipsychotic Trials of Intervention and Effectiveness schizophrenia study. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2012;61(4):577-99.

Appendix A: Effect of knots on treatment effect from PLR and Cox models

Table 45 Effect of number of knots on treatment effect from PLR and Cox models

Time interval	Treatment changes	Number of observations (patients)	Number of withdrawals (remissions)			Cox PH regression treatment effect (log likelihood) for treatment only model	PLR treatment effect (log pseudolikelihood) for treatment + time only model		
			LTG	VPS	Total		3 knots	4 knots	5 knots
Monthly	ITT	8315 (395)	- (143)	- (154)	- (297)	0.774 (-1587.9)	0.246 (-1135.2)	0.824 (-1152.4)	0.155 (-1110.7)
	ISC	6544 (392)	50 (110)	26 (141)	76 (251)	0.773 (-1318.8)	1.018 (-882.4)	0.602 (-915.4)	1.947 (-863.0)
	ISC/choice	6494 (392)	53 (108)	29 (139)	82 (247)	0.774 (-1295.1)	1.019 (-869.3)	0.592 (-901.6)	1.933 (-850.3)
	All	5405 (373)	64 (90)	56 (114)	120 (204)	0.736 (-1039.9)	0.889 (-699.5)	3.897 (-681.8)	0.005 (-637.5)
Biweekly	ITT	17985 (395)	- (143)	- (154)	- (297)	0.770 (-1582.7)	0.839 (-1349.9)	1.031 (-1351.5)	0.316 (-1327.6)
	ISC	14196 (395)	52 (110)	26 (141)	78 (251)	0.769 (-1313.4)	1.017 (-1071.4)	1.274 (-1080.4)	0.923 (-1068.5)
	ISC/choice	14087 (395)	55 (108)	29 (139)	84 (247)	0.770 (-1289.7)	1.018 (-1055.1)	1.264 (-1063.8)	0.917 (-1052.3)
	All	11766 (387)	74 (90)	59 (114)	133 (204)	0.731 (-1035.4)	0.298 (-851.8)	0.772 (-865.1)	0.058 (-845.1)
Weekly	ITT	36256 (396)	- (144)	- (155)	- (299)	0.762 (-1582.2)	0.701 (-1628.8)	1.034 (-1583.2)	0.156 (-1543.8)
	ISC	28606 (396)	52 (111)	26 (141)	78 (252)	0.767 (-1311.0)	1.021 (-1259.1)	1.176 (-1270.6)	0.086 (-1223.1)
	ISC/choice	28389 (396)	55 (109)	29 (139)	84 (248)	0.768 (-1287.4)	0.431 (-1237.0)	1.135 (-1251.5)	0.203 (-1218.4)
	All	23721 (392)	77 (91)	60 (114)	137 (205)	0.728 (-1033.0)	0.153 (-1003.2)	0.003 (-1004.3)	0.587 (-993.2)

Table 46 Effect of knots on unweighted remission model (monthly intervals)

Treatment changes	Number of observations (patients) [range]	Number of treatment changes (remission)	Cox PH treatment effect (ll)	Pooled logistic regression	1. 3 knots		2. 4 knots		3. 5 knots	
					Knot values	Treatment effect (pll)	Knot values	Treatment effect (pll)	Knot values	Treatment effect (pll)
1. ITT	8315 (395) [1,63]	- (297)	0.774 (-1587.9)	All follow up (knots at usual centiles)	3, 11, 36	NC	2,8,16,45	NC	2,6,11,20,45	NC
				All follow up (knots clustered at ~1yr)	3, 12, 36	NC	2,11,13,45	NC	2,6,11,13,45	NC
				All follow up (knots from time > 1 year; splines 0 < 1 year)	13, 22, 45	0.702 (-1115.8)	12,18,28,51	0.713 (-892.0)	12,16,22,32,51	0.715 (-879.6)
				Only >=1yr included		0.717 (-873.6)		0.713 (-892.0)		0.715 (-879.6)
2. ISC	4011 (379) [1,62]	76 (251)	0.772 (-1318.8)	All follow up (knots at usual centiles)	2, 9, 28	NC	1,7,12,37	NC	1,5,9,15,37	0.699 (-742.0)
				All follow up (knots clustered at ~1yr)	2, 12, 28	NC	1,7,12,37	NC	1,5,11,13,37	0.694 (-682.4)
				All follow up (knots from time > 1 year; splines 0 < 1 year)	12, 20, 40	0.694 (-968.0)	12,16,24,48	0.697 (-652.0)	12,15,20,27,48	0.693 (-644.7)
				Only >=1yr included		0.693 (-669.8)		0.697 (-652.0)		0.693 (-644.7)
3.ISC or choice	6494 (392) [1,62]	82 (247)	0.774 (-1295.1)	All follow up (knots at usual centiles)	2, 9, 28	NC	1,6,12,37	NC	1,5,9,15,37	0.705 (-731.8)
				All follow up (knots clustered at ~1yr)	2, 12, 28	NC	1,11,13,37	0.699 (-672.7)	1,5,11,13,37	0.699 (-672.7)
				All follow up (knots from time > 1 year; splines 0 < 1 year)	12, 20, 40	0.694 (-953.3)	12,16,24,48	0.700 (-642.4)	12,15,20,27,48	0.695 (-634.6)
				Only >=1yr included	12, 20, 40	0.699 (-660.1)	12,16,24,48	0.700 (-642.4)		0.695 (-634.6)
4. All	5405 (373) [1,62]	120 (204)	0.736 (-1039.9)	All follow up (knots at usual centiles)	2, 9, 24	NC	1,6,11,31	NC	1,5,9,13,31	NC
				All follow up (knots clustered at ~1yr)	2, 12, 24	NC	1,11,13,31	0.651 (-525.7)	1,5,11,13,31	NC
				All follow up (knots from time > 1 year; splines 0 < 1 year)	12, 18, 34	0.632 (-757.8)	12,15,22,42	NC	12,14,18,25,42	NC
				Only >=1yr included		0.650 (-516.1)		NC		NC

Table 47 Effect of knots on unweighted remission model (fortnightly intervals)

Treatment changes	Number of observations (patients) [range]	Number of treatment changes (remission)	Cox PH treatment effect (ll)	Pooled logistic regression	3 knots		4 knots		5 knots	
					Knot values	Treatment effect (pll)	Knot values	Treatment effect (pll)	Knot values	Treatment effect (pll)
1. ITT	17985 (395) [1,130]	- (297)	0.770 (-1582.7)	All follow up (knots at usual centiles)	5, 23, 76	NC	3,16,34,95	0.709 (-1239.7)	3,13,23,42,95	NC
				All follow up (knots clustered at ~1yr)	5, 26, 76	NC	3,25,27,95	0.699 (-1134.4)	3,13,25,27,95	NC
				All follow up (knots from time > 1 year; splines 0 < 1 year)	28, 49, 97	0.705 (-1346.9)	27,40,61,109	0.713 (-1026.4)	27,36,49,69,109	0.714 (-1012.7)
				Only >=1yr included		0.712 (-1045.0)		0.713 (-1026.4)		0.714 (-1012.7)
2. ISC	14196 (395) [1,130]	78 (251)	0.769 (-1313.4)	All follow up (knots at usual centiles)	4, 20, 60	NC	2,14,26,78	NC	2,11,20,32,78	0.706 (-934.9)
				All follow up (knots clustered at ~1yr)	4, 26, 60	NC	2,25,27,78	NC	2,11,25,27,78	NC
				All follow up (knots from time > 1 year; splines 0 < 1 year)	28, 44, 86	0.706 (-1053.1)	26,36,54,103	0.699 (-796.4)	26,33,44,60,103	0.694 (-781.4)
				Only >=1yr included		0.697 (-772.9)		0.699 (-796.4)		0.694 (-781.4)
3.ISC or choice	14087 (395) [1,130]	84 (247)	0.770 (-1289.7)	All follow up (knots at usual centiles)	4, 19, 60	NC	2,13,26,78	NC	2,11,19,32,78	0.712 (-921.1)
				All follow up (knots clustered at ~1yr)	4, 26, 60	NC	2,25,27,78	NC	2,11,25,27,78	NC
				All follow up (knots from time > 1 year; splines 0 < 1 year)	28, 44, 87	0.711 (-1037.9)	26,36,54,103	0.702 (-784.4)	26,33,44,60,103	0.695 (-768.6)
				Only >=1yr included		0.698 (-760.4)		0.702 (-784.4)		0.695 (-768.6)
4. All	11766 (387) [1,130]	133 (204)	0.731 (-1035.4)	All follow up (knots at usual centiles)	4, 18, 51	NC	2,12,24,65	NC	2,10,18,28,65	NC
				All follow up (knots clustered at ~1yr)	4, 26, 51	NC	2,25,27,65	NC	2,10,25,27,65	NC
				All follow up (knots from time > 1 year; splines 0 < 1 year)	27, 41, 75	0.662 (-816.7)	26,35,49,91	NC	26,32,41,54,91	NC
				Only >=1yr included		0.646 (-597.1)		NC		NC

Table 48 Effect of knots on unweighted remission model (weekly intervals)

Treatment changes	Number of observations (patients) [range]	Number of treatment changes (remission)	Cox PH treatment effect (ll)	Pooled logistic regression	3 knots		4 knots		5 knots	
					Knot values	Treatment effect (pll)	Knot values	Treatment effect (pll)	Knot values	Treatment effect (pll)
1. ITT	36122 (396) [1,263]	- (297)	0.768 (-1580.6)	All follow up (knots at usual centiles)	10, 47, 152	NC	5,33,68,191	NC	5,26,47,85,191	NC
				All follow up (knots clustered at ~1yr)	10, 52, 152	NC	5,51,53,191	NC	5,26,51,53,191	NC
				All follow up (knots from time > 1 year; splines 0 < 1 year)	58, 98, 196	0.707 (-1559.7)	54,81,123,219.65	0.708 (-1192.7)	54,73,98,139, 219.65	0.711 (-1170.2)
				Only >=1yr included		0.708 (-1223.6)		0.708 (-1192.7)		0.711 (-1170.2)
2. ISC	28535 (396) [1,263]	78 (251)	0.767 (-1311.0)	All follow up (knots at usual centiles)	8, 39, 119	NC	4,27,51,155	NC	4,21,39,63,155	NC
				All follow up (knots clustered at ~1yr)	8, 52, 119	NC	4,51,53,155	NC	4,21,51,53,155	NC
				All follow up (knots from time > 1 year; splines 0 < 1 year)	56, 88, 174	0.710 (-1229.9)	53,74,109,207	0.696 (-879.6)	53,68,88,122,207	0.702 (-854.5)
				Only >=1yr included		0.698 (-910.7)		0.696 (-879.6)		0.702 (-854.5)
3.ISC or choice	28318 (396) [1,263]	84 (247)	0.768 (-1287.4)	All follow up (knots at usual centiles)	8, 39, 120	NC	4,26,51,156	NC	4,21,39,63,156	NC
				All follow up (knots clustered at ~1yr)	8, 52, 120	NC	4,51,53,156	NC	4,21,51,53,156	NC
				All follow up (knots from time > 1 year; splines 0 < 1 year)	56, 88, 175	0.715 (-1211.7)	53,74,109,208	0.696 (-864.2)	53,68,88,122,208	0.703 (-839.9)
				Only >=1yr included		0.699 (-895.9)		0.696 (-864.2)		0.703 (-839.9)
4. All	23671 (392) [1,263]	137 (204)	0.728 (-1033.0)	All follow up (knots at usual centiles)	7, 35, 102	NC	4,24,48,131	NC	4,19,35,55,131	NC
				All follow up (knots clustered at ~1yr)	7, 52, 102	NC	4,51,53,131	0.657 (-763.0)	4,19,51,53,131	NC
				All follow up (knots from time > 1 year; splines 0 < 1 year)	56, 83, 151.3	0.666 (-969.5)	53,70,99,183	0.645 (-675.4)	53,65,83,110,183	0.652 (-661.3)
				Only >=1yr included		0.647 (-706.4)		0.645 (-675.4)		0.652 (-661.3)

Appendix B: IPCW modelling results

The treatment effect (CI) in the cell nearest to the left is valid for empty cells.

[] denotes the usual normal 95% CI for comparison with the bootstrapped CI for PLR.

Subscripts following CIs denote the number of bootstrap samples for which the model did not converge.

Key for tables demonstrating variables selected into model:

× Addition (or removal) of variable alters treatment effect by less than 1%, and variable is therefore not included in pool of potential variables for selection into model

. Variable is included in initial pool of potential variables for selection into model (i.e. addition of variable alters treatment effect by more than 1%) but fails to be included in the 2%, 5% or 10% models (as its addition (or removal) of variable alters treatment effect by less than 2%)

✓ Addition (or removal) of variable alters treatment effect between 2 and 5%

✓ Addition (or removal) of variable alters treatment effect between 5 and 10%

☑ Addition (or removal) of variable alters treatment effect by more than 10%

Table 49 Treatment effect (95% BS CI) [standard CI] with monthly intervals

	Treatment alone	Variable selection threshold 10%	5%	2%
No treatment changes				
PLR – 3 knots	0.71 [0.55, 0.92]			
FS		0.71 (0.57, 0.95) [0.55, 0.92]		
BE		0.71 (0.57, 0.95) [0.55, 0.92]		
PLR – 5 knots	0.72 (0.55, 0.93)			
FS		0.72 (0.57, 0.95) [0.55, 0.93]		
BE		0.72 (0.57, 0.95) [0.55, 0.93]		
Survival analysis	0.77 (0.62, 0.97)			
FS		0.77 (0.64, 0.94)		
BE		0.77 (0.64, 0.94)		
ISC changes only				
PLR – 3 knots	0.70 (0.52, 0.94)			
FS		0.70 (0.52, 0.996) [0.52, 0.94]		0.68 (0.41, 1.09)$_{1}$ [0.51, 0.91]
BE		0.70 (0.52, 0.996) [0.52, 0.94]		0.68 (0.50, 0.97) [0.51, 0.92]
PLR – 5 knots	0.70 (0.52, 0.93)			
FS		0.70 (0.51, 0.99) [0.52, 0.93]		0.68 (0.39, 1.11)$_{127}$ [0.50, 0.91]
BE		0.70 (0.51, 0.99) [0.52, 0.93]		
Survival analysis	0.77 (0.60, 0.99)			
FS		0.65 (0.29, 0.98)	0.71 (0.47, 1.62)	0.68 (0.43, 1.26)
BE		0.77 (0.62, 0.99)	0.74 (0.53, 1.00)	0.68 (0.43, 1.26)
ISC/choice changes only				
PLR – 3 knots	0.70 (0.52, 0.94)			
FS		0.70 (0.52, 0.99) [0.52, 0.94]		0.68 (0.45, 1.42) [0.51, 0.92]
BE		0.70 (0.52, 0.99) [0.52, 0.94]		
PLR – 5 knots	0.70 (0.52, 0.94)			
FS		0.70 (0.51, 0.996) [0.52, 0.94]		0.68 (0.37, 1.11)$_{104}$ [0.50, 0.93]
BE		0.70 (0.51, 0.996) [0.52, 0.94]		
Survival analysis	0.77 (0.60, 0.97)			
FS		0.63 (0.22, 0.92)	0.69 (0.42, 1.40)	0.63 (0.39, 1.34)
BE		0.63 (0.24, 0.92)		0.63 (0.37, 1.17)
All treatment changes				
PLR – 3 knots	0.65 (0.47, 0.90)			
FS		0.65 (0.48, 0.90) [0.47, 0.90]	0.60 (0.41, 1.34)$_{4}$ [0.43, 0.85]	0.63 (0.45, 0.75)$_{4}$ [0.44, 0.89]
BE		0.65 (0.48, 0.90) [0.47, 0.90]	0.58 (0.42, 1.02) [0.42, 1.02]	0.62 (0.45, 1.10)$_{4}$ [0.44, 0.88]
PLR – 5 knots	0.64 (0.46, 0.90)			
FS		0.64 (0.48, 0.90) [0.46, 0.90]	0.60 (0.39, 1.36)$_{23}$ [0.43, 0.84]	0.60 (0.38, 1.33)$_{25}$ [0.43, 0.86]
BE		0.64 (0.48, 0.90) [0.46, 0.90]	0.58 (0.41, 1.03)$_{23}$ [0.40, 0.82]	0.62 (0.45, 1.13)$_{22}$ [0.44, 0.88]
Survival analysis	0.74 (0.56, 0.97)			
FS		0.51 (0.31, 0.84)	0.55 (0.33, 0.92)	
BE		0.58 (0.34, 0.92)	0.58 (0.34, 0.99)	0.60 (0.34, 1.05)

Table 50 Variable selection (using 2%, 5% and 10% thresholds) with monthly intervals

Variables considered for selection (**bold** indicates time-varying covariates)

	Age	TCS	Interval	*Sex*	*NI*	*Feb*	*Rels*	*Hist*	*EEG*	*Type*	**Seizures**	**AEs**	**Dose**
No treatment changes													
PLR – 3 knots													
FS	×	×	·	×	·	×	·	·	·	×	-	-	-
BE	×	×	·	×	·	×	·	·	·	×	-	-	-
PLR – 5 knots													
FS	×	×	·	×	·	×	·	·	·	×	-	-	-
BE	×	×	·	×	·	×	·	·	·	×	-	-	-
Survival analysis													
FS	×	×	×	×	×	×	×	·	×	×	-	-	-
BE	×	×	×	×	×	×	×	·	×	×	-	-	-
ISC changes only													
PLR – 3 knots													
FS	·	×	×	·	×	×	·	·	×	·	✓	×	·
BE	✓	×	×	·	×	×	·	·	×	·	·	×	·
PLR – 5 knots													
FS	·	×	×	×	×	×	·	·	×	·	✓	×	·
BE	·	×	×	×	×	×	·	·	×	·	·	×	·
Survival analysis													
FS	·	×	×	×	×	×	✓	×	×	☑	☑	×	✓
BE	·	×	×	×	×	×	✓	×	×	✓	✓	×	✓
ISC/choice changes only													
PLR – 3 knots													
FS	·	×	×	·	×	×	·	·	×	✓	✓	×	✓
BE	·	×	×	·	×	×	·	·	×	·	·	×	·
PLR – 5 knots													
FS	·	×	×	·	×	×	·	·	×	✓	✓	×	✓
BE	·	×	×	·	×	×	·	·	×	·	·	×	·
Survival analysis													
FS	✓	×	×	×	×	×	✓	×	×	✓	☑	×	✓
BE	✓	×	×	×	×	×	·	×	×	·	☑	×	✓
All treatment changes													
PLR – 3 knots													
FS	·	×	×	✓	×	✓	✓	×	✓	×	✓	✓	✓
BE	✓	×	×	·	×	✓	✓	×	·	×	·	·	✓
PLR – 5 knots													
FS	·	×	×	·	×	✓	✓	×	·	×	✓	✓	✓
BE	✓	×	×	·	×	✓	✓	×	·	×	·	·	✓
Survival analysis													
FS	✓	×	×	✓	×	·	✓	×	✓	×	☑	·	☑
BE	✓	×	×	✓	×	✓	✓	×	✓	×	☑	✓	✓

Table 51 Treatment effect (95% BS CI) [standard CI] with fortnightly intervals

	Treatment alone	Variable selection threshold		
		10%	5%	2%
No treatment changes				
PLR – 3 knots	0.71 (0.55, 0.91)			
FS		0.71 (0.56, 0.94)		
		[0.55, 0.91]		
BE		0.71 (0.56, 0.94)		0.73 (0.59, 0.97)
		[0.55, 0.91]		[0.56, 0.95]
PLR – 5 knots	0.72 (0.56, 0.93)			
FS		0.72 (0.57, 0.94)		
		[0.56, 0.93]		
BE		0.72 (0.57, 0.94)		
		[0.56, 0.93]		
Survival analysis	0.77 (0.61, 0.97)			
FS		0.77 (0.63, 0.96)		
BE		0.77 (0.63, 0.96)		
ISC changes only				
PLR – 3 knots	0.70 (0.53, 0.93)			
FS		0.70 (0.55, 1.04)		0.70 (0.41, 1.22)
		[0.53, 0.93]		[0.52, 0.94]
BE		0.70 (0.55, 1.04)		
		[0.53, 0.93]		
PLR – 5 knots	0.70 (0.52, 0.94)			
FS		0.70 (0.54, 0.998)		
		[0.52, 0.94]		
BE		0.70 (0.54, 0.998)		0.71 (0.41, 1.26)
		[0.52, 0.94]		[0.53, 0.96]
Survival analysis	0.77 (0.60, 0.99)			
FS		0.59 (0.28, 0.87)		0.58 (0.28, 0.84)
BE		0.59 (0.28, 0.87)		
ISC/choice changes only				
PLR – 3 knots	0.70 (0.53, 0.94)			
FS		0.70 (0.57, 1.07)		0.70 (0.41, 1.18)
		[0.53, 0.94]		[0.52, 0.94]
BE		0.70 (0.57, 1.07)		
		[0.53, 0.94]		
PLR – 5 knots	0.70 (0.52, 0.94)			
FS		0.70 (0.55, 0.99)		0.71 (0.39, 1.24)
		[0.52, 0.94]		[0.52, 0.96]
BE		0.70 (0.55, 0.99)		
		[0.52, 0.94]		
Survival analysis	0.77 (0.60, 0.99)			
FS		0.52 (0.32, 0.75)	0.57 (0.30, 0.81)	0.54 (0.30, 0.76)
BE		0.52 (0.32, 0.75)	0.57 (0.30, 0.81)	0.54 (0.30, 0.76)
All treatment changes				
PLR – 3 knots	0.65 (0.47, 0.89)			
FS		0.65 (0.48, 0.91)	0.61 (0.44, 1.07)[1]	0.66 (0.41, 1.11)
		[0.47, 0.89]	[0.44, 0.84]	[0.48, 0.92]
BE		0.65 (0.48, 0.91)		0.67 (0.45, 1.20)[1]
		[0.47, 0.89]		[0.49, 0.93]
PLR – 5 knots	0.65 (0.47, 0.90)			
FS		0.65 (0.47, 0.88)	0.57 (0.32, 1.24)	0.62 (0.37, 1.05)
		[0.47, 0.90]	[0.40, 0.82]	[0.44, 0.87]
BE		0.65 (0.47, 0.88)		0.64 (0.46, 1.10)
		[0.47, 0.90]		[0.46, 0.90]
Survival analysis	0.73 (0.55, 0.97)			
FS		0.56 (0.33, 0.97)	0.57 (0.31, 1.01)	0.59 (0.35, 1.04)
BE		0.58 (0.30, 0.83)	0.55 (0.26, 0.81)	0.58 (0.32, 0.87)

Table 52 Variable selection (using 2%, 5% and 10% thresholds) with fortnightly intervals

Variables considered for selection (**bold** indicates time-varying covariates)

	Age	TCS	Interval	*Sex*	*NI*	*Feb*	*Rels*	*Hist*	*EEG*	*Type*	**Seizures**	**AEs**	**Dose**
No treatment changes													
PLR – 3 knots													
FS	×	×	·	×	·	×	·	·	·	×	-	-	-
BE	×	×	·	×	✓	×	✓	✓	✓	×	-	-	-
PLR – 5 knots													
FS	×	×	·	×	·	×	·	·	·	×	-	-	-
BE	×	×	·	×	·	×	·	·	·	×	-	-	-
Survival analysis													
FS	×	×	×	×	×	×	×	·	×	×	-	-	-
BE	×	×	×	×	×	×	×	·	×	×	-	-	-
ISC changes only													
PLR – 3 knots													
FS	·	×	×	✓	×	×	·	·	×	·	✓	×	·
BE	·	×	×	·	×	×	·	·	×	·	·	×	·
PLR – 5 knots													
FS	·	×	×	×	×	×	·	·	×	·	·	×	·
BE	✓	×	×	×	×	×	·	✓	×	·	✓	×	·
Survival analysis													
FS	✓	×	×	×	×	×	✓	×	×	☑	☑	×	☑
BE	·	×	×	×	×	×	·	×	×	☑	☑	×	☑
ISC/choice changes only													
PLR – 3 knots													
FS	✓	×	×	·	×	×	·	✓	×	·	✓	×	✓
BE	·	×	×	·	×	×	·	·	×	·	·	×	·
PLR – 5 knots													
FS	✓	×	×	·	×	×	·	✓	×	·	✓	×	✓
BE	·	×	×	·	×	×	·	·	×	·	·	×	·
Survival analysis													
FS	·	×	×	×	×	×	✓	×	×	✓	☑	×	☑
BE	·	×	×	×	×	×	✓	×	×	✓	☑	×	☑
All treatment changes													
PLR – 3 knots													
FS	·	×	×	✓	×	✓	✓	×	✓	×	✓	✓	·
BE	·	×	×	·	×	·	·	×	·	×	·	·	·
PLR – 5 knots													
FS	✓	×	×	·	×	·	✓	×	✓	×	✓	✓	·
BE	✓	×	×	·	×	✓	✓	×	·	×	·	·	✓
Survival analysis													
FS	✓	×	×	·	×	✓	✓	×	☑	×	☑	☑	✓
BE	✓	×	×	✓	×	✓	·	×	✓	×	·	☑	☑

Table 53 Treatment effect (95% BS CI) [standard CI] with weekly intervals

	Treatment alone	Variable selection threshold		
		10%	5%	2%
No treatment changes				
PLR – 3 knots	0.70 (0.55, 0.91)			
FS		0.70 (0.55, 0.91)		
		[0.55, 0.91]		
BE		0.70 (0.55, 0.91)		
		[0.55, 0.91]		
PLR – 5 knots	0.72 (0.56, 0.93)			
FS		0.72 (0.54, 0.93)		
		[0.56, 0.93]		
BE		0.72 (0.54, 0.93)		
		[0.56, 0.93]		
Survival analysis	0.76 (0.61, 0.96)			
FS		0.76 (0.60, 0.93)		
BE		0.76 (0.60, 0.93)		
ISC changes only				
PLR – 3 knots	0.70 (0.53, 0.93)			
FS		0.70 (0.55, 0.97)		0.71 (0.45, 1.15)
		[0.53, 0.93]		[0.53, 0.96]
BE		0.70 (0.55, 0.97)		0.67 (0.49, 0.98)
		[0.53, 0.93]		[0.50, 0.90]
PLR – 5 knots	0.70 (0.53, 0.94)			
FS		0.70 (0.54, 0.96)		0.67 (0.42, 1.30)
		[0.53, 0.94]		[0.50, 0.91]
BE		0.70 (0.54, 0.96)		0.67 (0.47, 1.02)
		[0.53, 0.94]		[0.50, 0.90]
Survival analysis	0.77 (0.60, 0.99)			
FS		0.60 (0.31, 1.12)		0.56 (0.32, 0.90)
BE		0.53 (0.30, 0.85)	0.58 (0.33, 1.07)	0.56 (0.32, 0.90)
ISC/choice changes only				
PLR – 3 knots	0.70 (0.53, 0.94)			
FS		0.70 (0.54, 0.96)		0.70 (0.45, 1.14)
		[0.53, 0.94]		[0.52, 0.94]
BE		0.70 (0.54, 0.96)		0.70 (0.44, 1.15)
		[0.53, 0.94]		[0.52, 0.94]
PLR – 5 knots	0.70 (0.52, 0.94)			
FS		0.70 (0.53, 0.94)		0.68 (0.44, 1.29)
		[0.52, 0.94]		[0.51, 0.92]
BE		0.70 (0.53, 0.94)		0.68 (0.44, 0.97)
		[0.52, 0.94]		[0.51, 0.92]
Survival analysis	0.77 (0.60, 0.99)			
FS		0.59 (0.33, 1.02)	0.59 (0.32, 1.03)	0.56 (0.36, 0.90)
BE		0.53 (0.29, 0.89)		0.56 (0.27, 0.91)
All treatment changes				
PLR – 3 knots	0.65 (0.47, 0.89)			
FS		0.65 (0.47, 0.87)	0.56 (0.36, 1.23)	0.61 (0.40, 1.01)
		[0.47, 0.89]	[0.39, 0.80]	[0.43, 0.86]
BE		0.65 (0.47, 0.87)	0.59 (0.38, 0.99)	0.61 (0.40, 1.01)
		[0.47, 0.89]	[0.43, 0.82]	[0.43, 0.86]
PLR – 5 knots	0.65 (0.47, 0.90)			
FS		0.65 (0.47, 0.88)	0.55 (0.37, 1.14)	0.57 (0.40, 1.15)
		[0.47, 0.90]	[0.38, 0.80]	[0.41, 0.81]
BE		0.65 (0.47, 0.88)	0.58 (0.34, 1.57)	0.62 (0.36, 1.31)
		[0.47, 0.90]	[0.41, 0.81]	[0.44, 0.87]
Survival analysis	0.73 (0.55, 0.96)			
FS		0.54 (0.27, 0.93)	0.50 (0.25, 0.85)	0.60 (0.35, 1.05)
BE		0.55 (0.29, 0.82)		0.60 (0.35, 1.05)

Table 54 Variable selection (using 2%, 5% and 10% thresholds) with weekly intervals

Variables considered for selection (**bold** indicates time-varying covariates)

	Age	TCS	Interval	*Sex*	*NI*	*Feb*	*Rels*	*Hist*	*EEG*	*Type*	**Seizures**	**AEs**	**Dose**
No treatment changes													
PLR – 3 knots													
FS	×	×	·	·	·	×	·	·	·	×	-	-	-
BE	×	×	·	·	·	×	·	·	·	×	-	-	-
PLR – 5 knots													
FS	×	×	·	×	·	×	·	·	·	×	-	-	-
BE	×	×	·	×	·	×	·	·	·	×	-	-	-
Survival analysis													
FS	×	×	×	×	×	×	×	·	×	×	-	-	-
BE	×	×	×	×	×	×	×	·	×	×	-	-	-
ISC changes only													
PLR – 3 knots													
FS	·	×	×	✓	×	×	✓	✓	✓	·	✓	✓	✓
BE	·	×	×	·	×	×	✓	·	·	·	·	✓	✓
PLR – 5 knots													
FS	✓	×	×	×	×	×	·	✓	×	·	✓	✓	✓
BE	·	×	×	×	×	×	·	✓	×	·	·	✓	✓
Survival analysis													
FS	☑	×	×	×	×	×	✓	×	×	✓	☑	☑	☑
BE	✓	×	×	×	×	×	✓	×	×	✓	☑	✓	☑
ISC/choice changes only													
PLR – 3 knots													
FS	·	×	×	✓	×	×	·	✓	✓	·	✓	✓	✓
BE	·	×	×	✓	×	×	·	✓	✓	·	✓	✓	✓
PLR – 5 knots													
FS	✓	×	×	·	×	×	✓	✓	×	·	✓	✓	✓
BE	·	×	×	·	×	×	·	·	×	·	·	·	✓
Survival analysis													
FS	✓	×	×	×	×	×	✓	×	×	☑	☑	✓	☑
BE	✓	×	×	×	×	×	✓	×	×	·	☑	✓	☑
All treatment changes													
PLR – 3 knots													
FS	·	×	×	✓	×	✓	·	×	✓	×	✓	✓	·
BE	·	×	×	✓	×	✓	·	×	✓	×	✓	✓	·
PLR – 5 knots													
FS	·	×	×	·	×	·	·	×	✓	×	✓	✓	·
BE	·	×	×	✓	×	✓	·	×	·	×	·	✓	✓
Survival analysis													
FS	✓	×	×	✓	×	·	✓	×	☑	×	☑	☑	✓
BE	✓	×	×	✓	×	·	✓	×	✓	×	☑	✓	✓

Table 55 Withdrawal (VPS): Number of variables in model (EPV, including spline variables (3 with 3 knots (2 spline variables + linear time variable), 5 with 5 knots (4 spline variables + linear time variable))

Treatment changes	Time interval / Variable selection threshold											
	Month				Fortnight				Week			
	10%	5%	2%	Full model	10%	5%	2%	Full model	10%	5%	2%	Full model
ITT (none)												
PLR_3												
FS												
BE												
PLR_5												
FS												
BE												
Survival												
FS												
BE												
ISC only												
PLR_3												
FS	0 (8.7)	0 (8.7)	1 (6.5)		0 (8.7)	0 (8.7)	2 (5.2)		0 (8.7)	0 (8.7)	8 (2.4)	
BE	0 (8.7)	0 (8.7)	1 (6.5)	7 (2.6)	0 (8.7)	0 (8.7)	0 (8.7)	7 (2.6)	0 (8.7)	0 (8.7)	3 (4.3)	10 (2.0)
PLR_5												
FS	0 (5.2)	0 (5.2)	1 (4.3)		0 (5.2)	0 (5.2)	0 (5.2)		0 (5.2)	0 (5.2)	5 (2.6)	
BE	0 (5.2)	0 (5.2)	0 (5.2)	6 (2.4)	0 (5.2)	0 (5.2)	3 (3.3)	6 (2.4)	0 (5.2)	0 (5.2)	3 (3.3)	7 (2.2)
Survival												
FS	2 (13)	3 (8.7)	4 (6.5)		3 (8.7)	3 (8.7)	5 (5.2)		4 (6.5)	4 (6.5)	6 (4.3)	
BE	0 (NA)	1 (26)	4 (6.5)	5 (5.2)	3 (8.7)	3 (8.7)	3 (8.7)	5 (5.2)	2 (13)	3 (8.7)	6 (4.3)	6 (4.3)
ISC/choice												
PLR_3												
FS	0 (9.7)	0 (9.7)	3 (4.8)		0 (9.7)	0 (9.7)	4 (4.1)		0 (9.7)	0 (9.7)	7 (2.9)	
BE	0 (9.7)	0 (9.7)	0 (9.7)	7 (2.9)	0 (9.7)	0 (9.7)	0 (9.7)	7 (2.9)	0 (9.7)	0 (9.7)	7 (2.9)	10 (2.2)
PLR_5												
FS	0 (5.8)	0 (5.8)	3 (3.6)		0 (5.8)	0 (5.8)	4 (3.2)		0 (5.8)	0 (5.8)	6 (2.6)	
BE	0 (5.8)	0 (5.8)	0 (5.8)	7 (2.4)	0 (5.8)	0 (5.8)	0 (5.8)	7 (2.4)	0 (5.8)	0 (5.8)	1 (4.8)	8 (2.2)
Survival												
FS	1 (29)	3 (9.7)	5 (5.8)		2 (14.5)	3 (9.7)	4 (7.3)		3 (9.7)	5 (5.8)	6 (4.8)	
BE	1 (29)	1 (29)	3 (9.7)	5 (5.8)	2 (14.5)	3 (9.7)	4 (7.3)	5 (5.8)	2 (14.5)	2 (14.5)	5 (5.8)	6 (4.8)
All												
PLR_3												
FS	0 (18.7)	1 (14)	8 (5.1)		0 (18.7)	1 (14)	7 (5.6)		0 (18.7)	2 (11.2)	6 (6.2)	
BE	0 (18.7)	2 (11.2)	4 (8)	9 (4.7)	0 (18.7)	0 (18.7)	6 (6.2)	9 (4.7)	0 (18.7)	1 (14)	6 (6.2)	9 (4.7)
PLR_5												
FS	0 (11.2)	1 (9.3)	5 (5.6)		0 (11.2)	2 (8)	6 (5.1)		0 (11.2)	2 (8)	4 (6.2)	
BE	0 (11.2)	2 (8)	4 (6.2)	9 (4)	0 (11.2)	0 (11.2)	4 (6.2)	9 (4)	0 (11.2)	2 (8)	4 (6.2)	9 (4)
Survival												
FS	2 (28)	2 (28)	7 (8)		4 (14)	6 (9.3)	8 (7)		4 (14)	5 (11.2)	8 (7)	
BE	1 (56)	5 (11.2)	9 (6.2)	9 (6.2)	2 (28)	3 (19.7)	7 (8)	9 (6.2)	1 (56)	1 (56)	8 (7)	9 (6.2)

Table 56 Withdrawal (LTG): Number of variables in model (EPV, including spline variables (3 with 3 knots (2 spline variables + linear time variable), 5 with 5 knots (4 spline variables + linear time variable))

Treatment changes	Time interval / Variable selection threshold											
	Month				Fortnight				Week			
	10%	5%	2%	Full model	10%	5%	2%	Full model	10%	5%	2%	Full model
ITT (none)												
PLR_3												
FS												
BE												
PLR_5												
FS												
BE												
Survival												
FS												
BE												
ISC only												
PLR_3												
FS	0 (16.7)	0 (16.7)	1 (12.5)		0 (16.7)	0 (16.7)	2 (10)		0 (16.7)	0 (16.7)	8 (4.5)	
BE	0 (16.7)	0 (16.7)	1 (12.5)	7 (5)	0 (16.7)	0 (16.7)	0 (16.7)	7 (5)	0 (16.7)	0 (16.7)	3 (8.3)	10 (3.8)
PLR_5												
FS	0 (10)	0 (10)	1 (8.3)		0 (10)	0 (10)	0 (10)		0 (10)	0 (10)	5 (5)	
BE	0 (10)	0 (10)	0 (10)	6 (4.5)	0 (10)	0 (10)	3 (6.3)	6 (4.5)	0 (10)	0 (10)	3 (6.3)	7 (4.2)
Survival												
FS	2 (25)	3 (16.7)	4 (12.5)		3 (16.7)	3 (16.7)	5 (10)		4 (12.5)	4 (12.5)	6 (8.3)	
BE	0 (NA)	1 (50)	4 (12.5)	5 (10)	3 (16.7)	3 (16.7)	3 (16.7)	5 (10)	2 (25)	3 (16.7)	6 (8.3)	6 (8.3)
ISC/choice												
PLR_3												
FS	0 (17.7)	0 (17.7)	3 (8.8)		0 (17.7)	0 (17.7)	4 (7.6)		0 (17.7)	0 (17.7)	7 (5.3)	
BE	0 (17.7)	0 (17.7)	0 (17.7)	7 (5.3)	0 (17.7)	0 (17.7)	0 (17.7)	7 (5.3)	0 (17.7)	0 (17.7)	7 (5.3)	10 (4.1)
PLR_5												
FS	0 (10.6)	0 (10.6)	3 (6.6)		0 (10.6)	0 (10.6)	4 (5.9)		0 (10.6)	0 (10.6)	6 (4.8)	
BE	0 (10.6)	0 (10.6)	0 (10.6)	7 (4.4)	0 (10.6)	0 (10.6)	0 (10.6)	7 (4.4)	0 (10.6)	0 (10.6)	1 (8.8)	8 (4.1)
Survival												
FS	1 (53)	3 (17.7)	5 (10.6)		2 (26.5)	3 (17.7)	4 (13.3)		3 (17.7)	5 (10.6)	6 (8.8)	
BE	1 (53)	1 (53)	3 (17.7)	5 (10.6)	2 (26.5)	3 (17.7)	4 (13.3)	5 (10.6)	2 (26.5)	2 (26.5)	5 (10.6)	6 (8.8)
All												
PLR_3												
FS	0 (21.3)	1 (16)	8 (5.8)		0 (21.3)	1 (16)	7 (6.4)		0 (21.3)	2 (12.8)	6 (7.1)	
BE	0 (21.3)	2 (12.8)	4 (9.1)	9 (5.3)	0 (21.3)	0 (21.3)	6 (7.1)	9 (5.3)	0 (21.3)	1 (16)	6 (7.1)	9 (5.3)
PLR_5												
FS	0 (12.8)	1 (10.7)	5 (6.4)		0 (12.8)	2 (9.1)	6 (5.8)		0 (12.8)	2 (9.1)	4 (7.1)	
BE	0 (12.8)	2 (9.1)	4 (7.1)	9 (4.6)	0 (12.8)	0 (12.8)	4 (7.1)	9 (4.6)	0 (12.8)	2 (9.1)	4 (7.1)	9 (4.6)
Survival												
FS	2 (32)	2 (32)	7 (9.1)		4 (16)	6 (10.7)	8 (8)		4 (16)	5 (12.8)	8 (8)	
BE	1 (64)	5 (12.8)	9 (7.1)	9 (7.1)	2 (32)	3 (21.3)	7 (9.1)	9 (7.1)	1 (64)	1 (64)	8 (8)	9 (7.1)

Table 57 Remission: Number of variables in model (EPV, including spline variables (3 with 3 knots (2 spline variables + linear time variable), 5 with 5 knots (4 spline variables + linear time variable))

Treatment changes	Time interval / Variable selection threshold											
	Month				Fortnight				Week			
	10%	5%	2%	Full model	10%	5%	2%	Full model	10%	5%	2%	Full model
ITT (none)												
PLR_3												
FS	1 (74.3)	1 (74.3)	1 (74.3)		1 (74.3)	1 (74.3)	1 (74.3)		1 (74.3)	1 (74.3)	1 (74.3)	
BE	1 (74.3)	1 (74.3)	2 (59.4)	6 (33)	1 (74.3)	1 (74.3)	6 (33)	7 (29.7)	1 (74.3)	1 (74.3)	1 (74.3)	8 (27)
PLR_5												
FS	1 (49.5)	1 (49.5)	1 (49.5)		1 (49.5)	1 (49.5)	1 (49.5)		1 (49.5)	1 (49.5)	1 (49.5)	
BE	1 (49.5)	1 (49.5)	1 (49.5)	7 (24.8)	1 (49.5)	1 (49.5)	1 (49.5)	7 (24.8)	1 (49.5)	1 (49.5)	1 (49.5)	7 (24.8)
Survival												
FS	1 (297)	1 (297)	1 (297)		1 (297)	1 (297)	1 (297)		1 (297)	1 (297)	1 (297)	
BE	1 (297)	1 (297)	1 (297)	2 (148.5)	1 (297)	1 (297)	1 (297)	2 (148.5)	1 (297)	1 (297)	1 (297)	2 (148.5)
ISC only												
PLR_3												
FS	1 (62.8)	1 (62.8)	1 (62.8)		1 (62.8)	1 (62.8)	2 (50.2)		1 (62.8)	1 (62.8)	6 (27.9)	
BE	1 (62.8)	1 (62.8)	2 (50.2)	6 (27.9)	1 (62.8)	1 (62.8)	1 (62.8)	6 (27.9)	1 (62.8)	1 (62.8)	2 (50.2)	8 (22.8)
PLR_5												
FS	1 (41.8)	1 (41.8)	1 (41.8)		1 (41.8)	1 (41.8)	1 (41.8)		1 (41.8)	1 (41.8)	3 (31.4)	
BE	1 (41.8)	1 (41.8)	1 (41.8)	5 (25.1)	1 (41.8)	1 (41.8)	3 (31.4)	5 (25.1)	1 (41.8)	1 (41.8)	2 (35.9)	5 (25.1)
Survival												
FS	2 (125.5)	2 (125.5)	3 (83.7)		2 (125.5)	2 (125.5)	4 (62.8)		2 (125.5)	2 (125.5)	4 (62.8)	
BE	1 (251)	1 (251)	3 (83.7)	4 (62.8)	2 (125.5)	2 (125.5)	2 (125.5)	4 (62.8)	1 (251)	1 (251)	4 (62.8)	4 (62.8)
ISC/choice												
PLR_3												
FS	1 (61.8)	1 (61.8)	2 (49.4)		1 (61.8)	1 (61.8)	3 (41.2)		1 (61.8)	1 (61.8)	5 (30.9)	
BE	1 (61.8)	1 (61.8)	2 (49.4)	6 (27.4)	1 (61.8)	1 (61.8)	1 (61.8)	6 (27.4)	1 (61.8)	1 (61.8)	5 (30.9)	8 (22.5)
PLR_5												
FS	1 (41.2)	1 (41.2)	2 (35.3)		1 (41.8)	1 (41.8)	3 (30.9)		1 (41.2)	1 (41.2)	4 (27.4)	
BE	1 (41.2)	1 (41.2)	1 (41.2)	6 (22.5)	1 (41.8)	1 (41.8)	1 (41.2)	6 (22.5)	1 (41.2)	1 (41.2)	1 (41.2)	6 (22.5)
Survival												
FS	1 (247)	2 (123.5)	4 (61.8)		1 (247)	2 (123.5)	3 (82.3)		2 (123.5)	3 (82.3)	4 (61.8)	
BE	1 (247)	1 (247)	2 (123.5)	4 (61.8)	1 (247)	2 (123.5)	3 (82.3)	4 (61.8)	1 (247)	1 (247)	3 (82.3)	4 (61.8)
All												
PLR_3												
FS	1 (51)	1 (51)	6 (22.7)		1 (51)	1 (51)	6 (22.7)		1 (51)	1 (51)	5 (25.5)	
BE	1 (51)	1 (51)	6 (22.7)	7 (20.4)	1 (51)	1 (51)	6 (22.7)	7 (20.4)	1 (51)	1 (51)	5 (25.5)	7 (20.4)
PLR_5												
FS	1 (34)	1 (34)	3 (25.5)		1 (34)	1 (34)	5 (20.4)		1 (34)	1 (34)	3 (25.5)	
BE	1 (34)	2 (29.1)	4 (22.7)	7 (17)	1 (34)	1 (34)	4 (22.7)	7 (17)	1 (34)	1 (34)	3 (25.5)	7 (17)
Survival												
FS	1 (204)	6 (34)	6 (34)		3 (68)	4 (51)	6 (34)		3 (68)	3 (68)	6 (34)	
BE	1 (204)	4 (51)	7 (29.1)	7 (29.1)	1 (204)	2 (102)	6 (34)	7 (29.1)	1 (204)	1 (204)	6 (34)	7 (29.1)

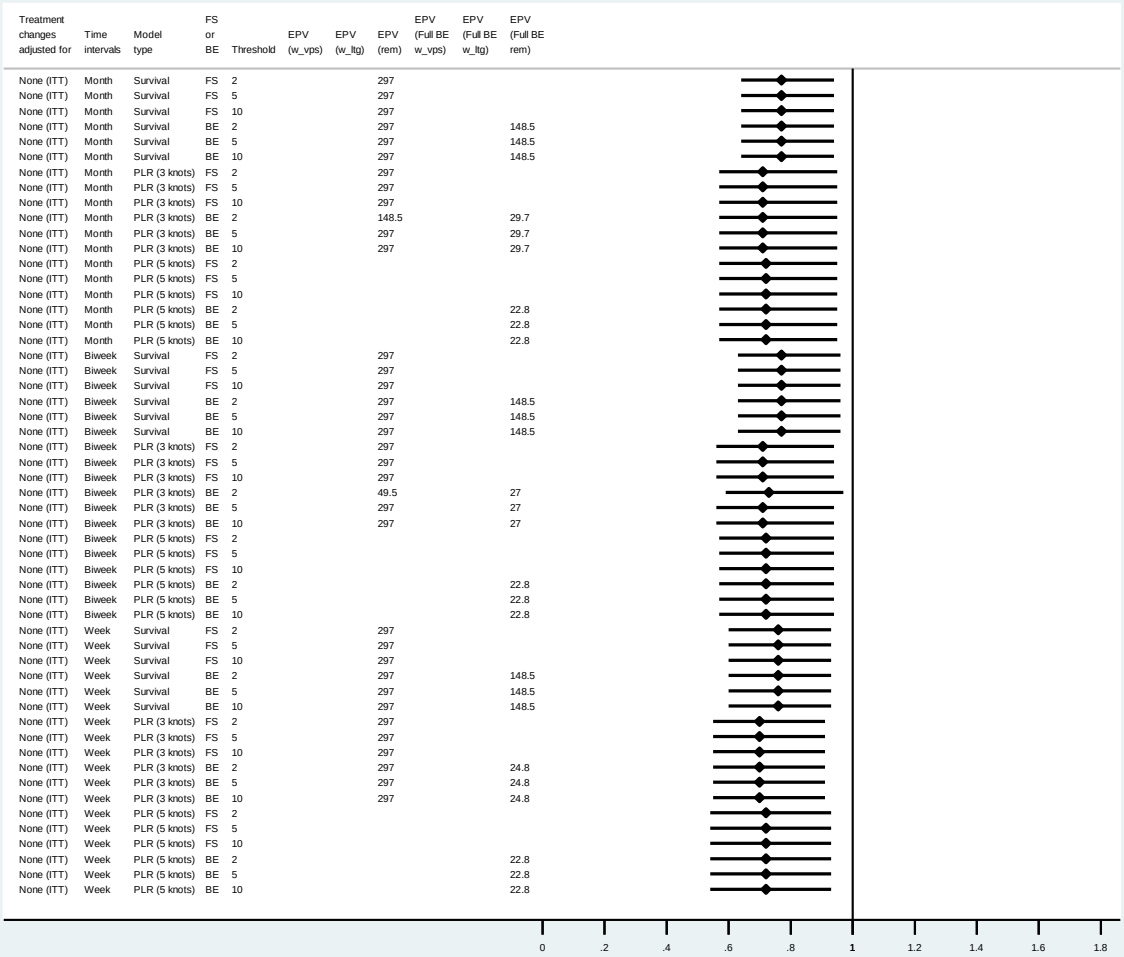


Figure 11 Treatment effects (95% BS CIs) from PLR and Cox models without adjustment for treatment changes (ITT)

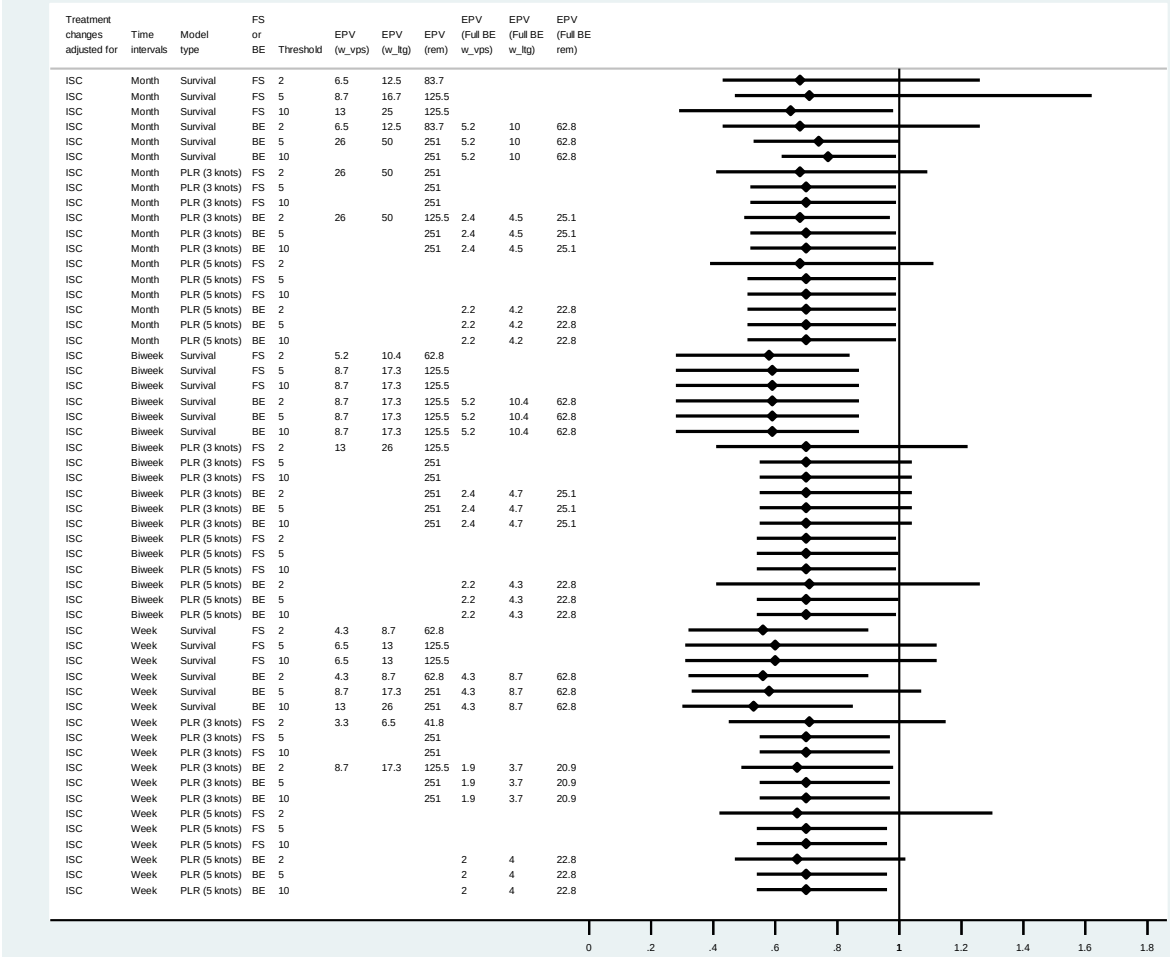


Figure 12 Treatment effects (95% BS CIs) from PLR and Cox models, with adjustment for treatment changes due to ISC alone

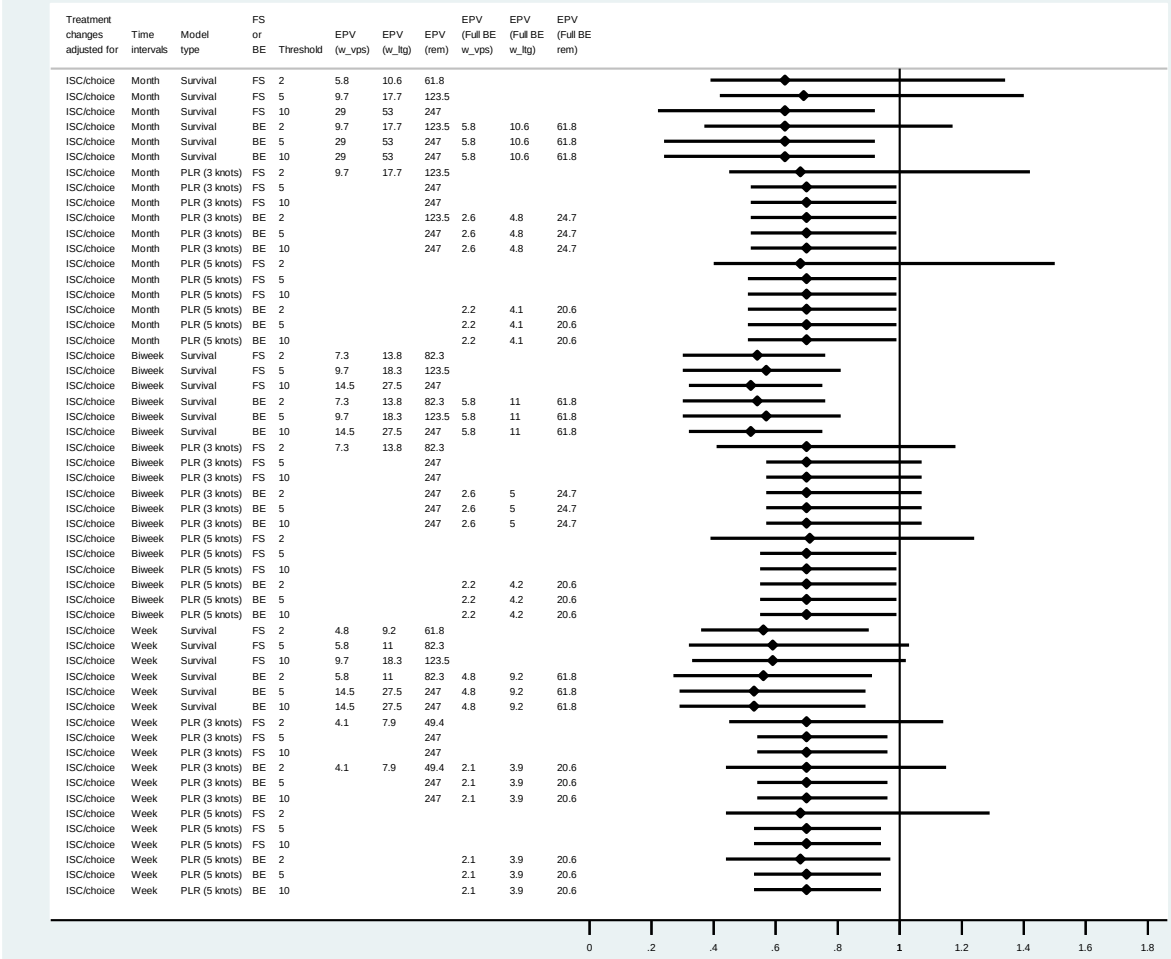


Figure 13 Treatment effects (95% BS CIs) from PLR and Cox models, with adjustment for treatment changes due to ISC or personal choice

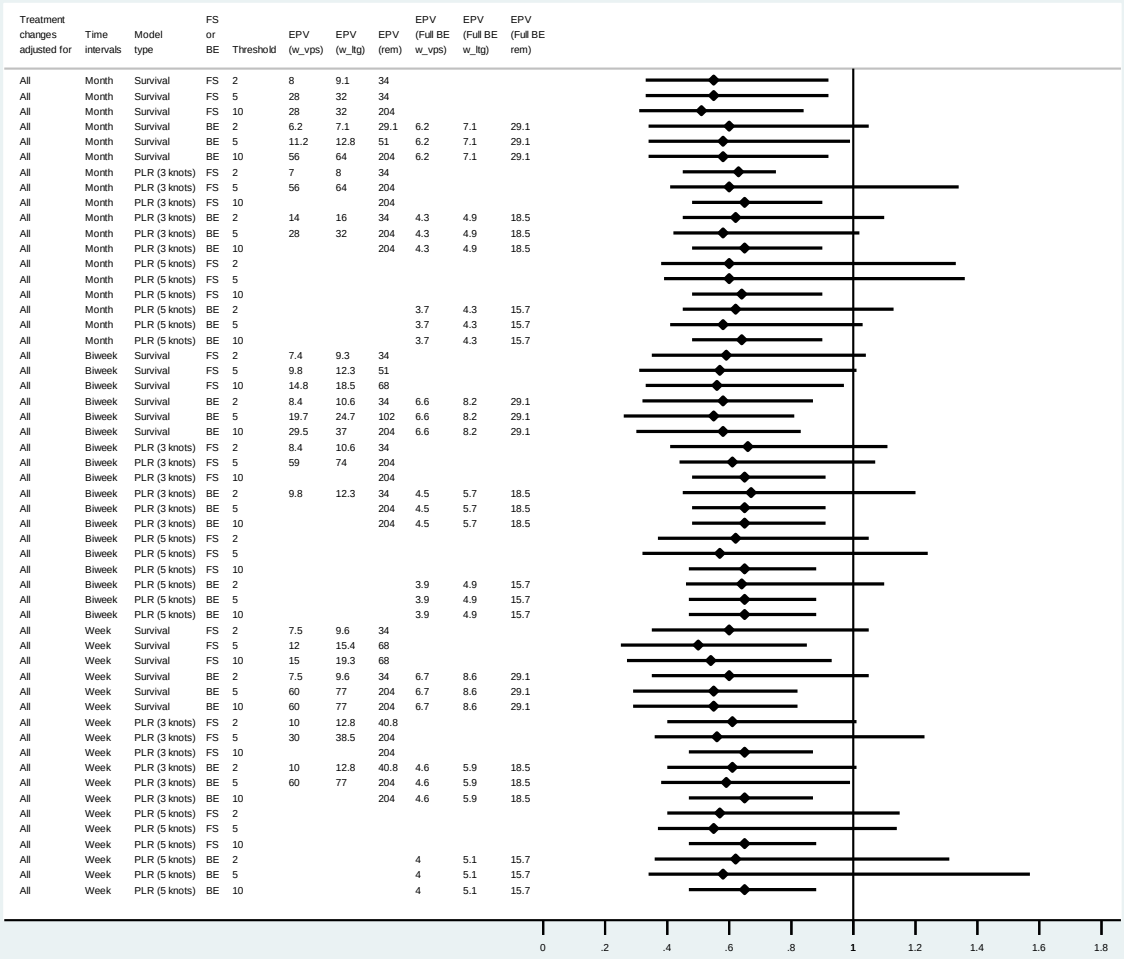


Figure 14 Treatment effects (95% BS CIs) from PLR and Cox models, with adjustment for all treatment changes

Appendix C: Comparison of IPCW and RPSFTM

Table 58 EPV ratios of WD model, according to treatment change scenarios and TVCs

WD models	PLR (0 knots)		PLR (3 knots)		PLR (5 knots)		Cox	
	VPS	**LTG**	**VPS**	**LTG**	**VPS**	**LTG**	**VPS**	**LTG**
Seizure only model								
Number of variables	2	2	4	4	6	6	1	1
Treatment changes: EPV ratio								
None (ITT)	-	-	-	-	-	-	-	-
ISC only	13	25	6.5	12.5	4.3	8.3	26	50
ISC/choice	14.5	26.5	7.3	13.3	4.8	8.8	29	53
All	28	32	14	16	9.3	10.7	56	64
Seizures and dose model								
Number of variables	3	3	5	5	7	7	2	2
Treatment changes: EPV ratio								
None (ITT)	-	-	-	-	-	-	-	-
ISC only	8.7	16.7	5.2	10	3.7	7.1	13	25
ISC/choice	9.7	17.7	5.8	10.6	4.1	7.6	14.5	26.5
All	18.7	21.3	11.2	12.8	8	9.1	28	32
Seizures, dose and AEs model								
Number of variables	4	4	6	6	8	8	3	3
Treatment changes: EPV ratio								
None (ITT)	-	-	-	-	-	-	-	-
ISC only	6.5	12.5	4.3	8.3	3.3	6.3	8.7	16.7
ISC/choice	7.3	13.3	4.8	8.8	3.6	6.6	9.7	17.7
All	14	16	9.3	10.7	7	8	18.7	21.3

Table 59 EPV ratios of WO model, according to treatment change scenarios

WO models		PLR (0 knots)		PLR (3 knots)		PLR (5 knots)		Cox	
		VPS	LTG	VPS	LTG	VPS	LTG	VPS	LTG
	Number of variables	2	2	4	4	6	6	1	1
	Treatment changes: EPV ratio								
	None (ITT)	148.5	74.3	49.5	297	148.5	74.3	49.5	297
	ISC only	125.5	62.8	41.8	251	125.5	62.8	41.8	251
	ISC/choice	123.5	61.8	41.2	247	123.5	61.8	41.2	247
	All	102	51	34	204	102	51	34	204

Table 60 Treatment only models

Treatment only models:	Month			Fortnight			Week		
	Treatment effect	95% CI	95% CI (BS)	Treatment effect	95% CI	95% CI (BS)	Treatment effect	95% CI	95% CI (BS)
No adjustment for treatment changes (ITT)									
PLR (0)	0.679	(0.516, 0.894)	(0.527, 0.906)	0.673	(0.512, 0.887)	(0.524, 0.901)	0.673	(0.512, 0.885)	(0.516, 0.916)
PLR (3)	0.714	(0.551, 0.924)	(0.565, 0.945)	0.707	(0.548, 0.912)	(0.561, 0.943)	0.704	(0.545, 0.908)	(0.547, 0.909)
PLR (5)	0.717	(0.553, 0.931)	(0.567, 0.950)	0.719	(0.556, 0.929)	(0.569, 0.944)	0.721	(0.559, 0.930)	(0.544, 0.928)
Cox	0.774	(0.616, 0.973)	(0.637, 0.944)	0.770	(0.613, 0.968)	(0.632, 0.956)	0.768	(0.611, 0.966)	(0.616, 0.940)
Adjusted for treatment changes due to ISC only									
PLR (0)	0.699	(0.513, 0.953)	(0.522, 0.989)	0.668	(0.491, 0.909)	(0.403, 1.111)	0.699	(0.513, 0.952)	(0.532, 0.985)
PLR (3)	0.699	(0.521, 0.936)	(0.522, 0.996)	0.701	(0.527, 0.934)	(0.546, 1.035)	0.702	(0.527, 0.934)	(0.549, 0.968)
PLR (5)	0.695	(0.517, 0.934)	(0.514, 0.994)	0.700	(0.523, 0.936)	(0.536, 0.998)	0.701	(0.525, 0.935)	(0.535, 0.960)
Cox	0.773	(0.602, 0.992)	(0.621, 0.993)	0.769	(0.598, 0.987)	(0.642, 1.000)	0.767	(0.597, 0.986)	(0.618, 0.962)
Adjusted for treatment changes due to ISC/choice only									
PLR (0)	0.703	(0.515, 0.961)	(0.514, 1.007)	0.673	(0.493, 0.919)	(0.397, 1.206)	0.703	(0.515, 0.961)	(0.537, 0.961)
PLR (3)	0.702	(0.522, 0.942)	(0.518, 0.990)	0.704	(0.527, 0.939)	(0.568, 1.070)	0.704	(0.527, 0.939)	(0.539, 0.957)
PLR (5)	0.696	(0.516, 0.938)	(0.510, 0.996)	0.701	(0.523, 0.940)	(0.548, 0.992)	0.701	(0.524, 0.938)	(0.532, 0.939)
Cox	0.774	(0.601, 0.990)	(0.623, 0.990)	0.770	(0.598, 0.991)	(0.638, 0.996)	0.768	(0.597, 0.989)	(0.624, 0.956)
Adjusted for all treatment changes (ISC/choice and UAEs)									
PLR (0)	0.644	(0.457, 0.909)	(0.456, 0.886)	0.602	(0.427, 0.849)	(0.412, 1.232)	0.641	(0.455, 0.905)	(0.449, 0.881)
PLR (3)	0.646	(0.465, 0.899)	(0.477, 0.899)	0.647	(0.469, 0.891)	(0.480, 0.911)	0.646	(0.470, 0.890)	(0.471, 0.867)
PLR (5)	0.644	(0.462, 0.899)	(0.476, 0.904)	0.649	(0.468, 0.900)	(0.474, 0.881)	0.648	(0.469, 0.897)	(0.471, 0.876)
Cox	0.736	(0.577, 0.973)	(0.590, 0.929)	0.731	(0.553, 0.965)	(0.581, 0.887)	0.728	(0.551, 0.962)	(0.552, 0.905)

Note that none of these models contain time-varying covariates and therefore no weighting is applied.

Table 61 Treatment and seizure models

Treatment + seizures models:	Month			Fortnight			Week		
	Treatment effect	95% CI	95% CI (BS)	Treatment effect	95% CI	95% CI (BS)	Treatment effect	95% CI	95% CI (BS)
No adjustment for treatment changes (ITT)*									
PLR (0)	0.679	(0.516, 0.894)	(0.527, 0.906)	0.673	(0.512, 0.887)	(0.524, 0.901)	0.673	(0.512, 0.885)	(0.516, 0.916)
PLR (3)	0.714	(0.551, 0.924)	(0.565, 0.945)	0.707	(0.548, 0.912)	(0.561, 0.943)	0.704	(0.545, 0.908)	(0.547, 0.909)
PLR (5)	0.717	(0.553, 0.931)	(0.567, 0.950)	0.719	(0.556, 0.929)	(0.569, 0.944)	0.721	(0.559, 0.930)	(0.544, 0.928)
Cox	0.774	(0.616, 0.973)	(0.637, 0.944)	0.770	(0.613, 0.968)	(0.632, 0.956)	0.768	(0.611, 0.966)	(0.616, 0.940)
Adjusted for treatment changes due to ISC only									
PLR (0)	0.662	(0.486, 0.902)	(0.453, 1.256)	0.668	(0.491, 0.909)	(0.403, 1.111)	0.665	(0.490, 0.904)	(0.384, 1.349)
PLR (3)	0.678	(0.505, 0.909)	(0.411, 1.089)	0.681	(0.511, 0.908)	(0.406, 1.250)	0.676	(0.507, 0.900)	(0.413, 1.170
PLR (5)	0.676	(0.501, 0.911)	$(0.386, 1.112)_{127}$	0.687	(0.511, 0.923)	(0.387, 1.256)	0.689	(0.514, 0.925)	$(0.376, 1.122)_{1}$
Cox	0.610	*(0.458, 0.814)*	(0.187, 0.916)	0.609	*(0.456, 0.814)*	(0.139, 0.907)	0.658	*(0.502, 0.862)*	(0.392, 1.146)
Adjusted for treatment changes due to ISC/choice only									
PLR (0)	0.669	(0.490, 0.914)	(0.477, 1.386)	0.673	(0.493, 0.919)	(0.397, 1.206)	0.673	(0.494, 0.916)	(0.364, 1.076)
PLR (3)	0.684	(0.509, 0.919)	(0.399, 1.155)	0.686	(0.514, 0.916)	(0.419, 1.485)	0.680	(0.510, 0.907)	(0.376, 1.209)
PLR (5)	0.681	(0.504, 0.920)	$(0.372, 1.111)_{104}$	0.691	(0.513, 0.930)	(0.378, 1.248)	0.692	(0.515, 0.931)	(0.376, 1.219)
Cox	0.628	*(0.474, 0.831)*	(0.222, 0.917)	0.627	*(0.473, 0.831)*	(0.265, 0.907)	0.663	*(0.507, 0.868)*	(0.396, 1.083)
Adjusted for all treatment changes (ISC/choice and UAEs)									
PLR (0)	0.605	(0.428, 0.854)	(0.417, 1.141)	0.602	(0.427, 0.849)	(0.412, 1.232)	0.592	(0.420, 0.832)	(0.367, 1.107)
PLR (3)	0.604	(0.431, 0.847)	(0.406, 1.343)	0.609	(0.440, 0.843)	(0.437, 1.072)	0.591	(0.427, 0.820)	$(0.379, 0.989)_{1}$
PLR (5)	0.599	(0.426, 0.844)	$(0.385, 1.362)_{23}$	0.605	(0.433, 0.846)	(0.402, 1.089)	0.596	(0.426, 0.834)	(0.371, 1.066)
Cox	0.581	*(0.441, 0.767)*	(0.340, 0.921)	0.529	*(0.395, 0.710)*	(0.161, 0.892)	0.552	*(0.416, 0.732)*	(0.286, 0.819)

* Note that none of these models contain time-varying covariates and therefore no weighting is applied; therefore these models correspond with those presented in Table 60.

CIs in italics indicate invalid normal Cox regression CIs, while subscripts indicate the number of bootstrapped samples (out of a total of 200) which failed to converge due to overfitting.

Table 62 Treatment, seizure count and dose models

Treatment + seizures + dose models:	Month			Fortnight			Week		
	Treatment effect	95% CI	95% CI (BS)	Treatment effect	95% CI	95% CI (BS)	Treatment effect	95% CI	95% CI (BS)
No adjustment for treatment changes (ITT)*									
PLR (0)	0.679	(0.516, 0.894)	(0.527, 0.906)	0.673	(0.512, 0.887)	(0.524, 0.901)	0.673	(0.512, 0.885)	(0.516, 0.916)
PLR (3)	0.714	(0.551, 0.924)	(0.565, 0.945)	0.707	(0.548, 0.912)	(0.561, 0.943)	0.704	(0.545, 0.908)	(0.547, 0.909)
PLR (5)	0.717	(0.553, 0.931)	(0.567, 0.950)	0.719	(0.556, 0.929)	(0.569, 0.944)	0.721	(0.559, 0.930)	(0.544, 0.928)
Cox	0.774	(0.616, 0.973)	(0.637, 0.944)	0.770	(0.613, 0.968)	(0.632, 0.956)	0.768	(0.611, 0.966)	(0.616, 0.940)
Adjusted for treatment changes due to ISC only									
PLR (0)	0.665	(0.488, 0.907)	(0.471, 1.267)	0.665	(0.490, 0.904)	(0.417, 1.172)	0.666	(0.491, 0.902)	(0.421, 1.278)
PLR (3)	0.673	(0.500, 0.906)	$(0.431, 1.374)_{2}$	0.677	(0.508, 0.903)	(0.432, 1.207)	0.674	(0.505, 0.898)	(0.391, 1.189)
PLR (5)	0.673	(0.497, 0.911)	$(0.391, 1.280)_{127}$	0.681	(0.506, 0.916)	$(0.421, 1.260)_{1}$	0.686	(0.510, 0.922)	$(0.398, 1.298)_{3}$
Cox	0.491	*(0.246, 0.981)*	(0.021, 1.930)	0.527	*(0.333, 0.833)*	(0.021, 0.909)	0.605	*(0.400, 0.915)*	(0.196, 17.318)
Adjusted for treatment changes due to ISC/choice only									
PLR (0)	0.668	(0.488, 0.913)	(0.450, 1.396)	0.667	(0.490, 0.909)	(0.409, 1.197)	0.671	(0.493, 0.912)	(0.409, 1.246)
PLR (3)	0.677	(0.501, 0.915)	$(0.420, 1.331)_{1}$	0.681	(0.509, 0.910)	(0.403, 1.302)	0.677	(0.507, 0.905)	(0.380, 1.332)
PLR (5)	0.676	(0.498, 0.919)	$(0.408, 1.263)_{106}$	0.683	(0.507,0.921)	$(0.408, 1.323)_{1}$	0.687	(0.510, 0.926)	$(0.382, 1.420)_{2}$
Cox	0.492	*(0.259, 0.933)*	(0.008, 1.206)	0.543	*(0.350, 0.844)*	(0.026, 0.930)	0.603	*(0.400, 0.908)*	(0.245, 5.652)
Adjusted for all treatment changes (ISC/choice and UAEs)									
PLR (0)	0.599	(0.423, 0.850)	(0.387, 1.224)	0.600	(0.424, 0.848)	(0.404, 1.197)	0.588	(0.417, 0.829)	(0.374, 1.182)
PLR (3)	0.593	(0.420, 0.838)	$(0.409, 1.422)_{4}$	0.603	(0.435, 0.836)	$(0.429, 1.575)_{1}$	0.584	(0.421, 0.810)	$(0.402, 1.261)_{1}$
PLR (5)	0.583	(0.408, 0.833)	$(0.415, 1.466)_{22}$	0.598	(0.426, 0.838)	(0.417, 1.385)	0.587	(0.419, 0.822)	(0.376, 1.567)
Cox	0.388	*(0.272, 0.552)*	(0.022, 0.690)	0.497	*(0.370, 0.669)*	(0.163, 0.934)	0.529	*(0.399, 0.700)*	(0.260, 0.809)

* Note that none of these models contain time-varying covariates and therefore no weighting is applied; therefore these models correspond with those presented in Table 60.
CIs in italics indicate invalid normal Cox regression CIs, while subscripts indicate the number of bootstrapped samples (out of a total of 200) which failed to converge due to overfitting.

Table 63 Treatment, seizure count, dose and AE models

Treatment + seizures + dose + AEs models:	Month			Fortnight			Week		
	Treatment effect	95% CI	95% CI (BS)	Treatment effect	95% CI	95% CI (BS)	Treatment effect	95% CI	95% CI (BS)
No adjustment for treatment changes (ITT)*									
PLR (0)	0.679	(0.516, 0.894)	(0.527, 0.906)	0.673	(0.512, 0.887)	(0.524, 0.901)	0.673	(0.512, 0.885)	(0.516, 0.916)
PLR (3)	0.714	(0.551, 0.924)	(0.565, 0.945)	0.707	(0.548, 0.912)	(0.561, 0.943)	0.704	(0.545, 0.908)	(0.547, 0.909)
PLR (5)	0.717	(0.553, 0.931)	(0.567, 0.950)	0.719	(0.556, 0.929)	(0.569, 0.944)	0.721	(0.559, 0.930)	(0.544, 0.928)
Cox	0.774	(0.616, 0.973)	(0.637, 0.944)	0.770	(0.613, 0.968)	(0.632, 0.956)	0.768	(0.611, 0.966)	(0.616, 0.940)
Adjusted for treatment changes due to ISC only									
PLR (0)	0.657	(0.482, 0.895)	(0.455, 1.264)	0.658	(0.485, 0.895)	(0.411, 1.068)	0.638	(0.470, 0.867)	(0.410, 1.139)
PLR (3)	0.661	(0.491, 0.890)	$(0.429, 1.253)_{2}$	0.668	(0.501, 0.891)	(0.407, 1.054)	0.644	(0.482, 0.860)	(0.415, 1.182)
PLR (5)	0.658	(0.486, 0.892)	$(0.389, 1.323)_{126}$	0.669	(0.497, 0.900)	(0.405, 1.077)	0.648	(0.482, 0.873)	$(0.416, 1.278)_{3}$
Cox	0.386	*(0.168, 0.884)*	(0.002, 1.661)	0.389	*(0.222, 0.681)*	(0.0007, 0.886)	0.417	*(0.249, 0.699)*	$(0.003, 1.750)_{1}$
Adjusted for treatment changes due to ISC/choice only									
PLR (0)	0.661	(0.483, 0.904)	(0.445, 1.433)	0.661	(0.485, 0.901)	(0.384, 1.146)	0.645	(0.474, 0.879)	(0.446, 1.112)
PLR (3)	0.667	(0.494, 0.902)	$(0.435, 1.343)_{1}$	0.672	(0.502, 0.899)	(0.428, 1.176)	0.650	(0.486, 0.871)	$(0.432, 1.166)_{1}$
PLR (5)	0.664	(0.488, 0.903)	$(0.441, 1.470)_{104}$	0.672	(0.498, 0.907)	(0.395, 1.097)	0.654	(0.484, 0.882)	$(0.421, 1.269)_{3}$
Cox	0.386	*(0.187, 0.796)*	(0.002, 1.131)	0.414	*(0.248, 0.692)*	(0.001, 0.964)	0.429	*(0.267, 0.689)*	(0.014, 2.316)
Adjusted for all treatment changes (ISC/choice and UAEs)									
PLR (0)	0.584	(0.410, 0.834)	(0.375, 1.094)	0.592	(0.414, 0.846)	(0.357, 1.160)	0.570	(0.398, 0.816)	(0.298, 1.514)
PLR (3)	0.576	(0.407, 0.816)	$(0.401, 1.204)_{4}$	0.584	(0.414, 0.823)	$(0.383, 1.128)_{1}$	0.551	(0.386, 0.786)	(0.342, 1.155)
PLR (5)	0.564	(0.394, 0.808)		0.570	(0.398, 0.817)	(0.357, 1.154)	0.541	(0.374, 0.783)	(0.326, 1.172)
Cox	0.161	*(0.077, 0.334)*	(0.0002, 0.414)	0.226	*(0.140, 0.367)*	(0.000002, 0.619)	0.315	*(0.214, 0.464)*	(0.008. 0.706)

* Note that none of these models contain time-varying covariates and therefore no weighting is applied; therefore these models correspond with those presented in Table 60.
CIs in italics indicate invalid normal Cox regression CIs, while subscripts indicate the number of bootstrapped samples (out of a total of 200) which failed to converge due to overfitting.

Table 64 Treatment only models

	PLR (no splines)	PLR (3 knots)	PLR (5 knots)	Cox	RPSFTM
No adjustment for treatment changes (ITT)					
Month	0.68 (0.53, 0.91)	0.71 (0.57, 0.95)	0.72 (0.57, 0.95)	0.77 (0.64, 0.94)	
Fortnight	0.67 (0.52, 0.90)	0.71 (0.56, 0.94)	0.72 (0.57, 0.94)	0.77 (0.63, 0.96)	0.79 (0.64, 0.98)
Week	0.67 (0.52, 0.92)	0.70 (0.55, 0.91)	0.72 (0.54, 0.93)	0.77 (0.62, 0.94)	
Adjusted for ISC treatment changes only					
Month	0.70 (0.52, 0.99)	0.70 (0.52, 0.996)	0.70 (0.51, 0.99)	0.77 (0.62, 0.99)	
Fortnight	0.67 (0.40, 1.11)	0.70 (0.55, 1.04)	0.70 (0.54, 0.998)	0.77 (0.64, 1.00)	0.88 (0.78, 0.99)
Week	0.70 (0.53, 0.99)	0.70 (0.55, 0.97)	0.70 (0.54, 0.96)	0.77 (0.62, 0.96)	
Adjusted for ISC/choice treatment changes only					
Month	0.70 (0.51, 1.01)	0.70 (0.52, 0.99)	0.70 (0.51, 0.996)	0.77 (0.62, 0.99)	
Fortnight	0.67 (0.40, 1.21)	0.70 (0.57, 1.07)	0.70 (0.55, 0.99)	0.77 (0.64, 0.996)	0.77 (0.61, 0.98)
Week	0.70 (0.54, 0.96)	0.70 (0.54, 0.96)	0.70 (0.53, 0.94)	0.77 (0.62, 0.96)	
Adjusted for all treatment changes					
Month	0.64 (0.46, 0.89)	0.65 (0.48, 0.90)	0.64 (0.48, 0.90)	0.74 (0.59, 0.93)	
Fortnight	0.60 (0.41, 1.23)	0.65 (0.48, 0.91)	0.65 (0.47, 0.88)	0.73 (0.58, 0.89)	0.76 (0.58, 0.98)
Week	0.64 (0.45, 0.88)	0.65 (0.47, 0.87)	0.65 (0.47, 0.88)	0.73 (0.55, 0.91)	

Table 65 Treatment and seizure models

	PLR (no splines)	PLR (3 knots)	PLR (5 knots)	Cox	RPSFTM
No adjustment for treatment changes (ITT)					
Month	0.68 (0.53, 0.91)	0.71 (0.57, 0.95)	0.72 (0.57, 0.95)	0.77 (0.64, 0.94)	
Fortnight	0.67 (0.52, 0.90)	0.71 (0.56, 0.94)	0.72 (0.57, 0.94)	0.77 (0.63, 0.96)	0.79 (0.64, 0.98)
Week	0.67 (0.52, 0.92)	0.70 (0.55, 0.91)	0.72 (0.54, 0.93)	0.77 (0.62, 0.94)	
Adjusted for ISC treatment changes only					
Month	0.66 (0.45, 1.26)	0.68 (0.41, 1.09)	0.68 (0.39, 1.11)	0.61 (0.19, 0.92)	
Fortnight	0.67 (0.40, 1.11)	0.68 (0.41, 1.25)	0.69 (0.39, 1.26)	0.61 (0.14, 0.91)	0.88 (0.78, 0.99)
Week	0.67 (0.38, 1.35)	0.68 (0.51, 0.90)	0.69 (0.38, 1.12)	0.66 (0.39, 1.15)	
Adjusted for ISC/choice treatment changes only					
Month	0.67 (0.48, 1.39)	0.68 (0.40, 1.16)	0.68 (0.37, 1.11)	0.63 (0.22, 0.92)	
Fortnight	0.67 (0.40, 1.21)	0.69 (0.42, 1.49)	0.69 (0.38, 1.25)	0.63 (0.27, 0.91)	0.77 (0.61, 0.98)
Week	0.67 (0.36, 1.08)	0.68 (0.38, 1.21)	0.69 (0.38, 1.22)	0.66 (0.40, 1.08)	
Adjusted for all treatment changes					
Month	0.61 (0.42, 1.14)	0.60 (0.41, 1.34)	0.60 (0.39, 1.36)	0.58 (0.34, 0.92)	
Fortnight	0.60 (0.41, 1.23)	0.61 (0.44, 1.07)	0.61 (0.40, 1.09)	0.53 (0.16, 0.89)	0.76 (0.58, 0.98)
Week	0.59 (0.37, 1.02)	0.60 (0.38, 0.99)	0.60 (0.37, 1.07)	0.55 (0.29, 0.82)	

Table 66 Treatment, seizures and dose models

	PLR (no splines)	PLR (3 knots)	PLR (5 knots)	Cox	RPSFTM
No adjustment for treatment changes (ITT)					
Month	0.68 (0.53, 0.91)	0.71 (0.57, 0.95)	0.72 (0.57, 0.95)	0.77 (0.64, 0.94)	
Fortnight	0.67 (0.52, 0.90)	0.71 (0.56, 0.94)	0.72 (0.57, 0.94)	0.77 (0.63, 0.96)	0.79 (0.64, 0.98)
Week	0.67 (0.52, 0.92)	0.70 (0.55, 0.91)	0.72 (0.54, 0.93)	0.77 (0.62, 0.94)	
Adjusted for ISC treatment changes only					
Month	0.67 (0.47, 1.27)	0.67 (0.43, 1.37)	0.67 (0.39, 1.28)	0.49 (0.02, 1.93)	
Fortnight	0.67 (0.42, 1.17)	0.68 (0.43, 1.21)	0.68 (0.42, 1.26)	0.53 (0.02, 0.91)	0.88 (0.78, 0.99)
Week	0.67 (0.42, 1.28)	0.67 (0.39, 1.19)	0.69 (0.40, 1.30)	0.61 (0.20, 17.32)	
Adjusted for ISC/choice treatment changes only					
Month	0.67 (0.45, 1.40)	0.68 (0.42, 1.33)	0.68 (0.41, 1.26)	0.49 (0.01, 1.21)	
Fortnight	0.67 (0.41, 1.20)	0.68 (0.40, 1.30)	0.68 (0.41, 1.32)	0.54 (0.03, 0.93)	0.77 (0.61, 0.98)
Week	0.67 (0.41, 1.25)	0.68 (0.38, 1.33)	0.69 (0.38, 1.42)	0.60 (0.25, 5.65)	
Adjusted for all treatment changes					
Month	0.60 (0.39, 1.22)	0.59 (0.41, 1.42)	0.58 (0.42, 1.47)	0.39 (0.02, 0.69)	
Fortnight	0.60 (0.40, 1.20)	0.60 (0.43, 1.58)	0.60 (0.42, 1.39)	0.50 (0.16, 0.93)	0.76 (0.58, 0.98)
Week	0.59 (0.37, 1.18)	0.58 (0.40, 1.26)	0.59 (0.38, 1.57)	0.53 (0.26, 0.81)	

Table 67 Treatment, seizures, dose and AEs models

	PLR (no splines)	PLR (3 knots)	PLR (5 knots)	Cox	RPSFTM
No adjustment for treatment changes (ITT)					
Month	0.68 (0.53, 0.91)	0.71 (0.57, 0.95)	0.72 (0.57, 0.95)	0.77 (0.64, 0.94)	
Fortnight	0.67 (0.52, 0.90)	0.71 (0.56, 0.94)	0.72 (0.57, 0.94)	0.77 (0.63, 0.96)	0.79 (0.64, 0.98)
Week	0.67 (0.52, 0.92)	0.70 (0.55, 0.91)	0.72 (0.54, 0.93)	0.77 (0.62, 0.94)	
Adjusted for ISC treatment changes only					
Month	0.66 (0.46, 1.26)	0.66 (0.43, 1.25)	0.66 (0.39, 1.32)	0.39 (0.002, 1.66)	
Fortnight	0.66 (0.41, 1.07)	0.67 (0.41, 1.05)	0.67 (0.41, 1.08)	0.39 (0.0007, 0.89)	0.88 (0.78, 0.99)
Week	0.64 (0.41, 1.14)	0.64 (0.42, 1.18)	0.65 (0.42, 1.28)	0.42 (0.003, 1.75)	
Adjusted for ISC/choice treatment changes only					
Month	0.66 (0.45, 1.43)	0.67 (0.44, 1.34)	0.66 (0.44, 1.47)	0.39 (0.002, 1.13)	
Fortnight	0.66 (0.38, 1.15)	0.67 (0.43, 1.18)	0.67 (0.40, 1.10)	0.41 (0.001, 0.96)	0.77 (0.61, 0.98)
Week	0.65 (0.45, 1.11)	0.65 (0.43, 1.17)	0.65 (0.42, 1.27)	0.43 (0.01, 2.32)	
Adjusted for all treatment changes					
Month	0.58 (0.38, 1.09)	0.58 (0.40, 1.20)	0.56 (NA)	0.16 (0.0002, 0.41)	
Fortnight	0.59 (0.36, 1.16)	0.58 (0.38, 1.13)	0.57 (0.36, 1.15)	0.23 (0.000002, 0.62)	0.76 (0.58, 0.98)
Week	0.57 (0.30, 1.51)	0.55 (0.34, 1.16)	0.54 (0.33, 1.17)	0.32 (0.008, 0.706)	

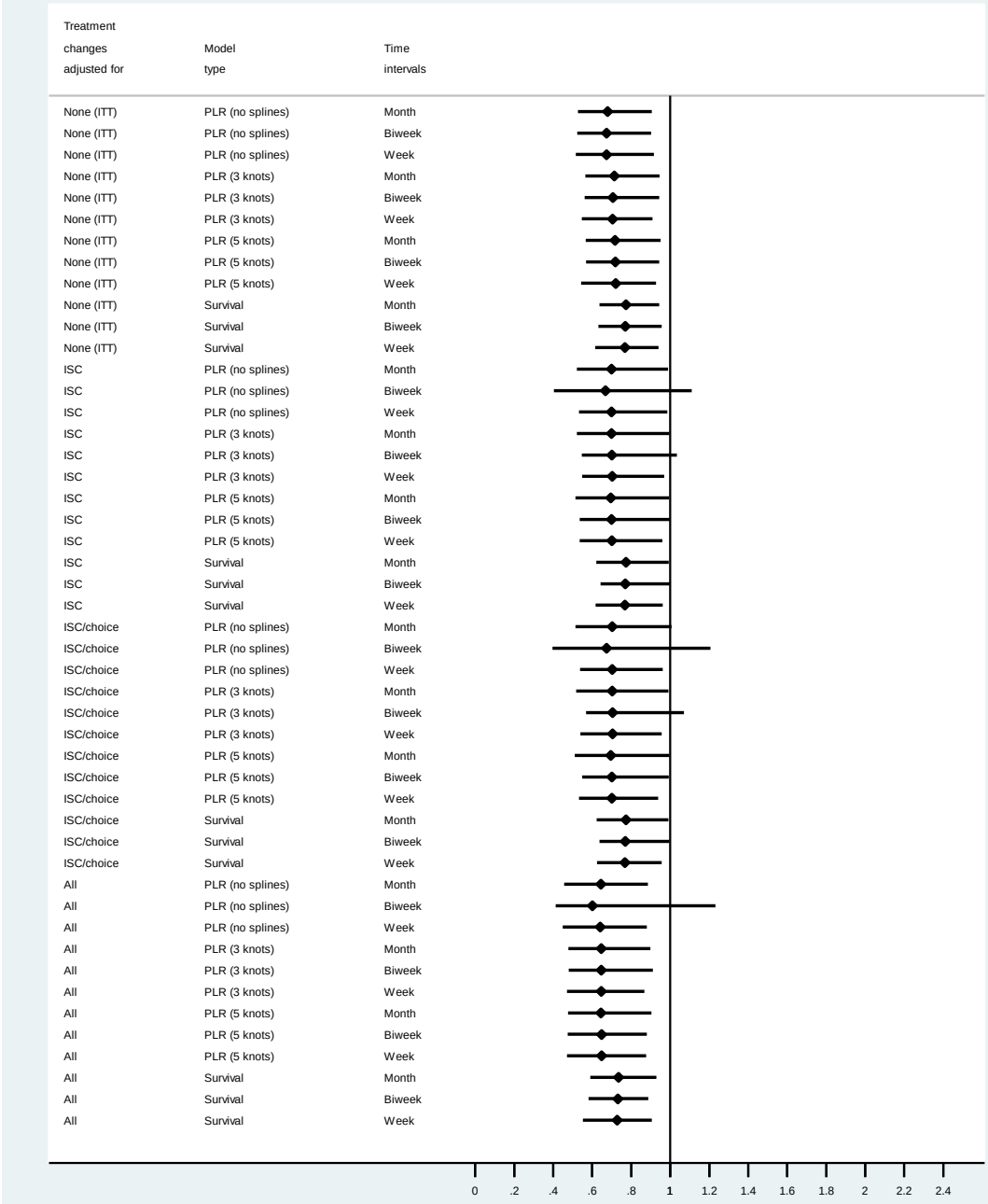


Figure 15 Treatment only models (BS CIs)

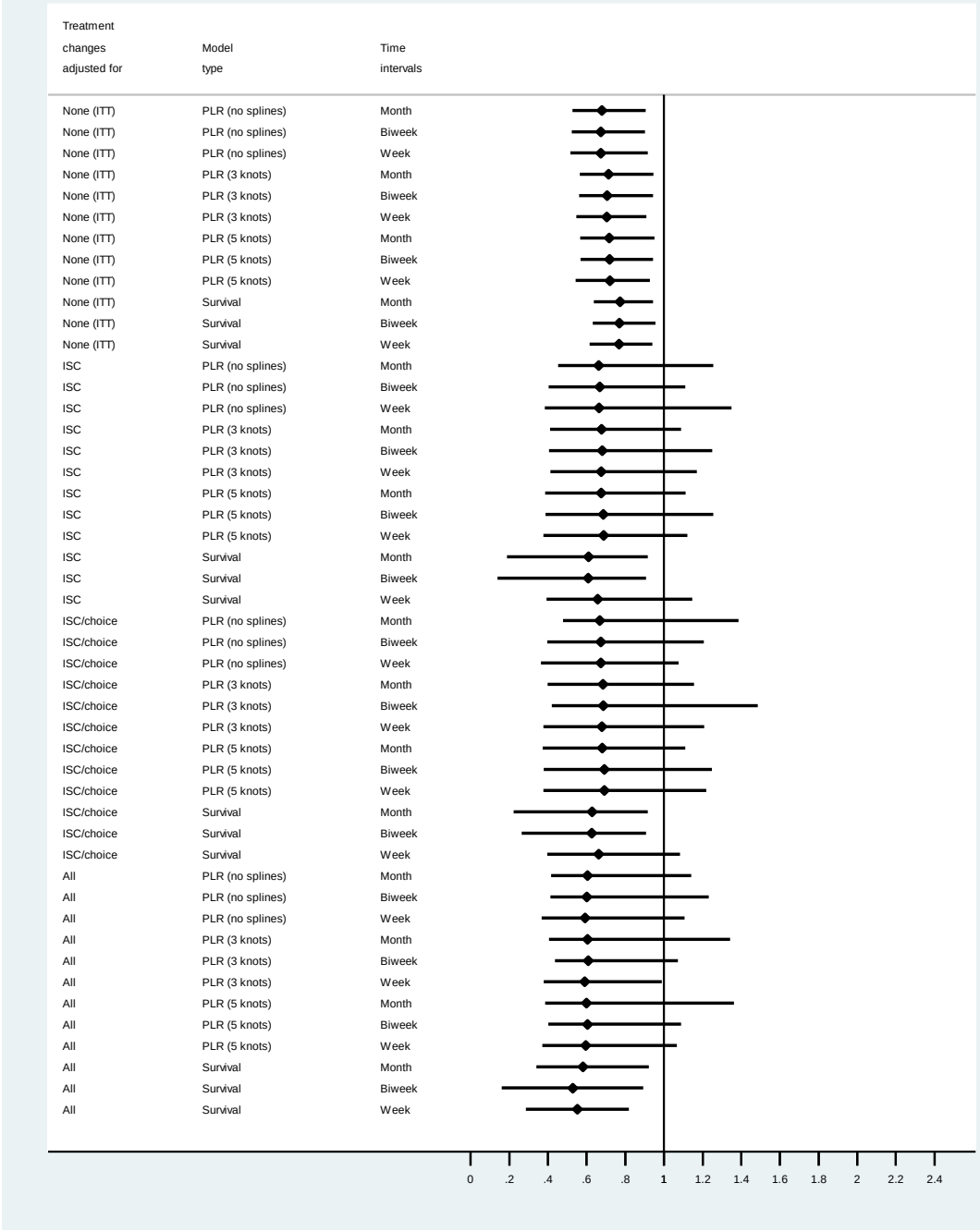


Figure 16 Treatment and seizures models (BS CIs)

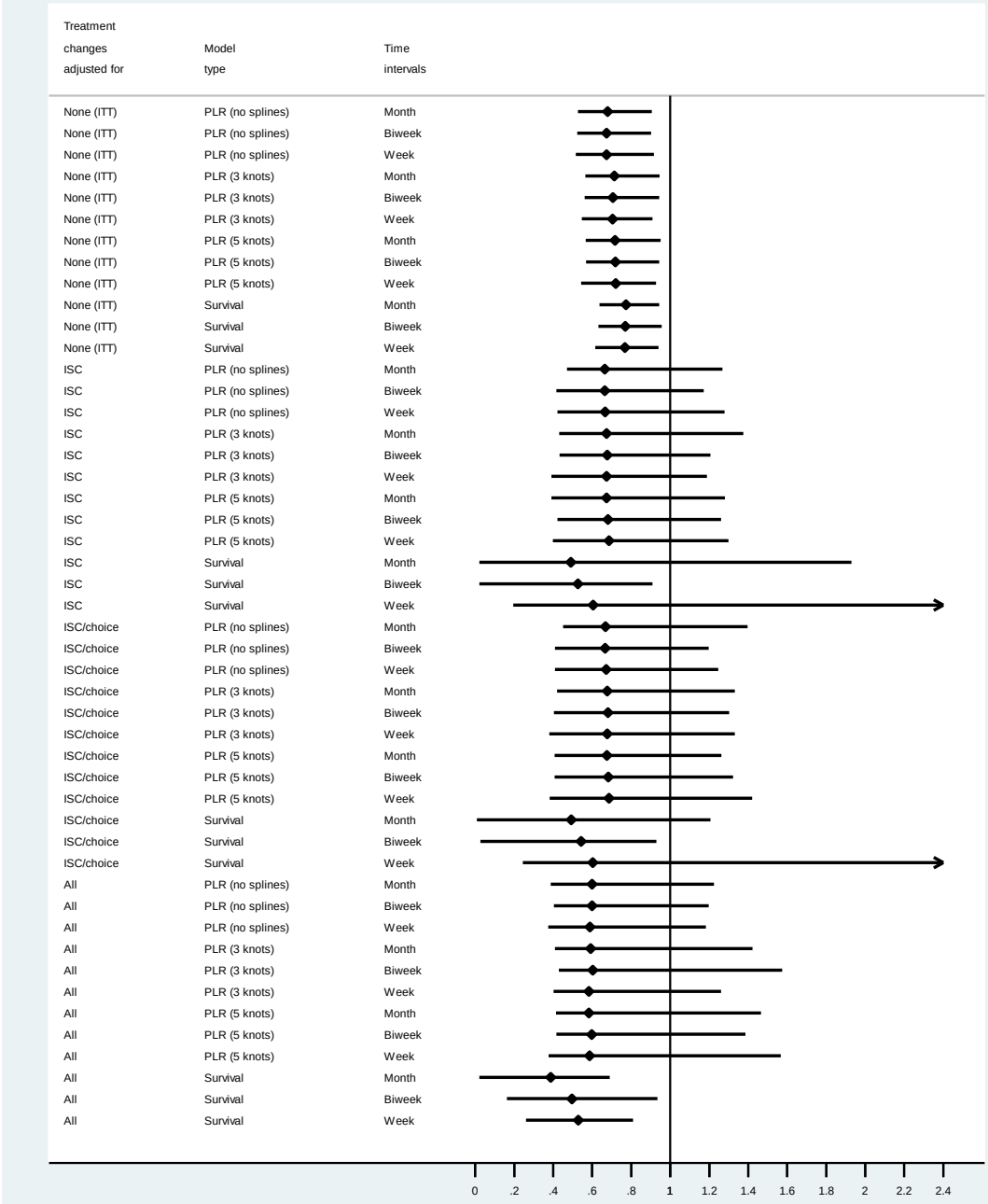


Figure 17 Treatment, seizures and dose models (BS CIs)

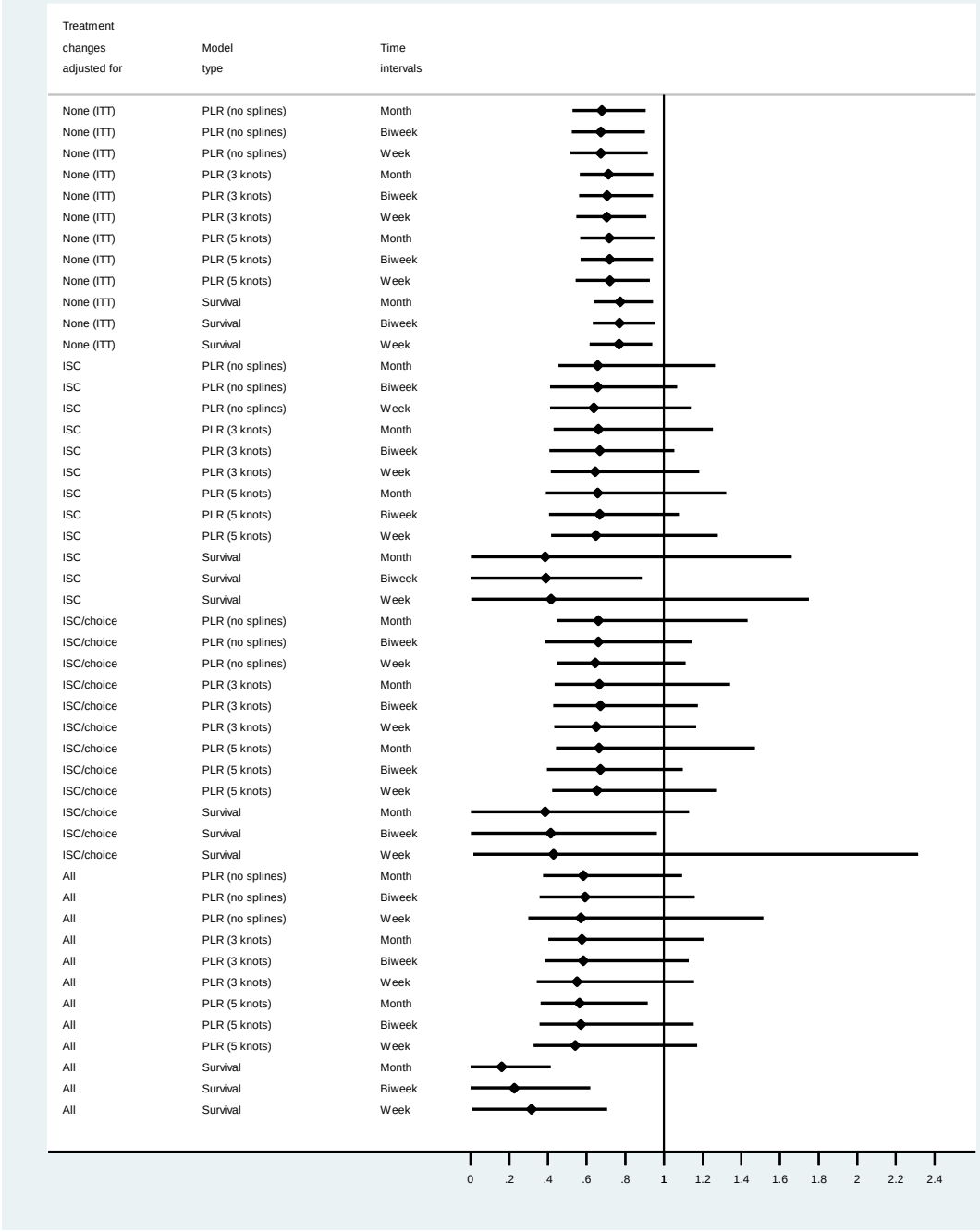


Figure 18 Treatment, seizures, dose and AE models (BS CIs)

Table 68 Accounting for reasons for treatment changes in RPSFTM

	HR (95% CI)	Number (%) of treatment changes adjusted for [% of treatment arm]		Number of ignored (censored) treatment changes (% of total number of IPCW treatment changes, m)		
		VPS (n=235)	LTG (n=217)		VPS (m=108)	LTG (m=105)
No adjustment for switches	0.794 (0.641, 0.983)	0	0	Switches between VPS and LTG (due to ISC, UAEs or choice)	18 (16.7)	35 (33.3)
				Switch to other AED	8 (7.4)	8 (7.6)
				Additions	18 (16.7)	24 (22.9)
				Withdrawals	64 (59.3)	38 (36.2)
Adjusted for switches due to ISC only	0.877 (0.777, 0.990)	9 (8.3) [3.8]	25 (23.8) [11.5]	Switches between VPS and LTG (due to UAEs or choice)	9 (8.3)	10 (9.5)
				Switch to other AED	8 (7.4)	8 (7.6)
				Additions	18 (16.7)	24 (22.9)
				Withdrawals	64 (59.3)	38 (36.2)
Adjusted for switches due to ISC or choice only	0.772 (0.608, 0.981)	11 (10.2) [4.7]	25 (23.8) [11.5]	Switches between VPS and LTG (due to UAEs)	7 (6.5)	10 (9.5)
				Switch to other AED	8 (7.4)	8 (7.6)
				Additions	18 (16.7)	24 (22.9)
				Withdrawals	64 (59.3)	38 (36.2)
Adjusted for all switches (due to ISC, choice or UAEs)	0.756 (0.584, 0.980)	18 (16.7) [7.7]	35 (33.3) [16.1]	Switch to other AED	8 (7.4)	8 (7.6)
				Additions	18 (16.7)	24 (22.9)
				Withdrawals	64 (59.3)	38 (36.2)