Practical Considerations in Designing and Analyzing Cross-over Clinical Trials

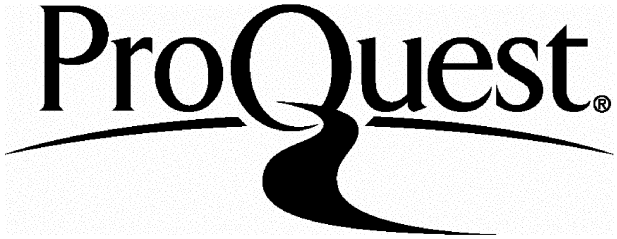
D. N. Lambrou

Supervisor: Professor S.J.Senn

A dissertation submitted in fulfillment
of the requirements for the degree of
Doctor of Philosophy
of the
University of London

Department of Statistical Science
University College London

June 2001



To my parents Nikolaos and Theodora
for their patience and understanding

Abstract

The problem of carry-over in cross-over trials has driven research activity for many decades. Depending on the assumptions made concerning carry-over effect in the 2x2 design, the CROS or the PAR estimator is selected for estimating treatment effect. The two stage procedure, selecting CROS with probability p and PAR with probability 1-p, achieves lower power and higher type I error-rate when compared to CROS. A corrected scheme, which achieves the nominal type I error-rate, proves inferior to alternative schemes regarding power and Mean Square Error Estimation rate. When baseline measurements are included in the analysis of the 2x2 design, a three-stage procedure emerges with similar properties to the two-stage one.

The optimum plan for designing a cross-over study in families with more than two periods and/or sequences, depends on the assumptions made for the carry-over effects and the optimality criterion chosen. Best plans for two treatments, when model mis-specification occurs in both the systematic and/or random part of the model assumed to have generated the observed data, are derived. When three or more treatments are compared, optimum cyclic plans are chosen under a wide range of assumptions concerning carry-over activity.

Acknowledgements

I would like to express my gratitude to my principal supervisor Prof S. J. Senn for his invaluable support and guidance during the completion of that project. I would like also to thank my subsidiary supervisor Prof T. Fearn for his suggestions that made this thesis richer both in content and in form.

This project would have never been completed without the financial support of the Greek Postgraduate Scholarship Body (IKY). I would also like to thank the Department of Statistical Science for the friendly working environment they created.

Finally my deepest gratitude goes to my family for their understanding and financial support.

Contents

List of Tables

List of Figures

Chapter 1

Introduction: Practical Issues in Cross-over Clinical Trials

1.1 Preliminaries

In a cross-over study each patient acts as his own control by trying all available treatments. As in all other types of clinical studies, patients are followed-up for some pre-specified time period and data are collected on them at pre-defined time points within that period. In an ideal world all patients would join the study at the same calendar date and follow-up measurements would be taken at identical time points after the entry date. In addition for cross-over studies switches to alternative therapies should be scheduled at similar time windows for each patient. This is rarely the case though. Each patient has his own trial history. The entry date defines time zero for each patient. The interval between two consecutive treatment periods, if it exists, constitutes the wash-out period for cross-over studies. At wash-out intervals, baseline and other background information is typically collected in order to assess patient's physical condition before entering the next treatment phase.

Occasions exist where there is no standard therapy on the market, and a clinical trial is set up in order to provide the population at risk with such therapy. In that case the control group will receive no active treatment, or equivalently they receive placebo. If the baseline characteristics of the active and the control group are similar at wash-out periods, then any statistically important difference be-

tween the two groups during the next treatment phase can be attributed to the effect of treatment under study. For ethical purposes all participants in a clinical study are on concomitant therapies prescribed by others (GP) and participants are usually advised to avoid certain medications that may prevent the treatment showing its effect. The way specific medications may interact with treatments under study in a cross-over trial, may generate research questions of interest to the medical community.

In what follows, reference will be made to a set of guidelines prepared during the International Conference on Harmonisation (ICH) of technical requirements for registration of pharmaceuticals for human use. These guidelines have been prepared by the appropriate ICH working group and has been subject to consultation by the regulatory authorities.

1.2 Phases of drug development

A typical drug-development exercise involves several phases of clinical research before the drug hits the market. Much thought is devoted to the design of the various phases, since poorly designed and conducted trials can offer misleading findings, in sharp contrast with current scientific knowledge. The real purpose of well-planned trials is to influence clinical practice to an appreciable extent.

1.2.1 Phase I

This is the area in pharmaceutical research where cross-over designs enjoy wide applicability. Phase I studies help the scientific community to understand the biological activity of a test compound on the human body. A small number of volunteers receive the new therapy so that the dose-range expected to be studied in later phases can be determined (see ICH E8 guidelines). From the pharmacology point of view, unacceptable doses can lead to toxicity problems, which in turn cause adverse reactions. In some experiments compounds are tried on animals first and then the maximum animal tolerated dose is extrapolated to humans. The design adopted to determine the boundaries of toxicity are simple step-up/step-down schemes, i.e. a cross-over study with a special design.

Patients start with a quite low dose which gradually increases until toxicity is observed. Modern Bayesian design theory can provide the experimenter with sampling schemes that if used properly, can allow the location of the maximum tolerable dose to be assessed accurately (see Atkinson [1] or Pilz [70]). Once the data have been collected a dose response curve is fitted and then the maximum tolerable dose is determined by solving the dose-response equation with respect to dose for a given value of the response. There are other study-types in Phase I, where the cross-over design has been used successfully. Some are listed below:

- bio-availability studies where the level of drug absorbed by the body at various doses is considered
- bio-equivalence studies for the comparison of two formulations in terms of safety and efficacy
- pharmacokinetic (PK) studies, where drug absorption, distribution and elimination around the body is the main concern
- pharmacodynamic (PD) studies, where the relationship between the drug concentration at the site of action with pharmacologic response, is evaluated. These studies are useful indicators for early determination of the safety and the efficacy profile of the compound under study (see ICH E8 guidelines)
- interaction studies, where the extent to which the PK profile of the drug under study is affected by the presence of other drugs is the focus of interest
- safety studies, where maximum tolerable dose is established. Animal studies may be relevant to that type of clinical trial

Note that in a typical Phase I study no formal sample size evaluation takes place. The number of participants can vary from 12 up to 50 depending on resources and type of compound under study.

1.2.2 Phase II

Based on the phase I results, at phase II an initial assessment of drug effectiveness, but also of drug safety is established. Usually, phase I studies provide the

experimenter with a range of acceptable doses. If the dose administered is lower than the lower limit of that range then the drug is completely ineffective, while doses beyond the upper limit may cause toxicity and increase the possibility of adverse events being present. Phase II studies offer the opportunity to decide more accurately the dose(s) that are worth further attention and might be studied in subsequent phases. This assessment can be carried-out using a cross-over experiment. The number of participants in that phase ranges from small to moderate.

Apart from determining optimum dose for a specific compound, the sponsor is in a position to assess patient's responsiveness to competing therapies or to make comparisons with baseline status (see ICH E8). In both scenarios the use of cross-over trial is appropriate. It is worth noting that the analysis variable(s) during that period, may not be the same as the analysis variable(s) in later phases. Also the study-population at that stage are selected by narrow criteria (see ICH E8), though the Phase II population may have different characteristics compared to population recruited in phase III. This trial period is nothing more than an exploratory phase, which gives the sponsor the opportunity to determine clinical queries worth pursuing at later stages.

1.2.3 Phase III or Parallel versus Cross-over design

The knowledge accumulated from the two previous phases is used for the design of that phase, where the effectiveness of the new compound is firmly established, but in addition knowledge on safety is also collected, so that the role of the new therapy in clinical practice is fully evaluated. Usually phase III involves long term studies, since a deeper understanding of what affects recovery from the disease and of what disease complications the patient will suffer from, needs to be clarified. Obviously, phase III studies need to be of sufficient size and follow-up measurements are taken at carefully selected time points, so that the therapeutic activity of the test compound is clearly demonstrated. According to ICH E8 guidelines, at phase III sponsors may explore dose-response relationships, drug's use in wider populations or drug's effectiveness at different states of disease. How appropriate are cross-over plans for running a Phase III study? Since detailed

evaluation of the new therapy requires long term surveillance, cross-over plans where the test compound is observed at a large number of successive treatment periods may be appropriate. This type of cross-over design may prove problematic though, especially if the alternative therapy is placebo. This is one reason for favoring parallel studies during that stage of drug development.

The preparation needed to set-up a phase III study is quite enormous. The investigators involved need to have minimal knowledge regarding the safety of the therapy and have the necessary infrastructure to run the study. In addition, since phase III studies are usually of appreciable magnitude and of high cost, sponsors should have convincing evidence of the therapy's effectiveness to warrant the effort and expenses involved. From a cost-benefit perspective, cross-over plans are economic solutions, since fewer participants will need to be recruited compared to a parallel group study, in order to detect a pre-defined treatment difference. Regulatory authorities on the other hand, require firm evidence of the new therapy's effectiveness based on data derived from relatively large study populations, since this evidence is used for marketing approval (see ICH E8 document). So, from a regulator's perspective a parallel group study is more appropriate for use in Phase III programs.

Furthermore, the timing of running the phase III study is crucial for the success of the medical program. If the standard therapy has been in use for many years and has been widely accepted as efficacious for some indications by the scientific community, then as long as a newly discovered therapy achieves a remarkable improvement on the same condition, the phase III study should commence as soon as possible. On the other hand, if ongoing research continuously improves the standard therapy, then by the time a long phase III program ends, the proposed therapy will be outdated. It is under the second scenario that cross-over trials may be of some use to sponsors. If a quick comparison of the current standard therapy versus the new treatment is needed, then a cross-over trial with a limited number of participants can be set up to provide sponsors with the necessary answers.

In all phases of drug development a document that describes the objectives, design and procedures the investigator should follow during the course of the trial,

must be prepared. This document is the clinical protocol and defines a set of rules that facilitates communication between all working parties involved in the study. The protocol should be signed-off before recruitment starts and only minor updates may be allowed while the study is ongoing. The protocol in any phase III study, usually contains information about the clinical study-background, clinical objectives, primary and secondary analysis variable(s), study populations, sample size assumptions, inclusion/exclusion criteria, baseline examination, follow-up assessments, data analysis strategies (interim, final, stopping rules) and any other information that affects the running of the study. The protocol of a cross-over study may look more complicated, since special preparations may need to be undertaken during a wash-out interval before the patient enters the next treatment period. Finally note that for drug approval purposes, different studies are run with their own specific objectives. Each study's objective is described in a separate protocol. During a FDA hearing meeting, a document that describes the common features of the studies as well as the contribution made by each study separately should be prepared (see ICH E9 guidelines).

In all clinical phases, Phase III included, the primary question the investigators are most interested in, should be defined. This question is usually stated in a hypothesis testing format and is usually by taking measurements on the primary variable (endpoint). Based on these measurements inference is drawn for the population parameter of interest. As it is stated in the ICH E9 guidelines *the primary endpoint should be the variable capable of providing the most clinically relevant and convincing evidence directly related to the primary objective of the trial* . There may be secondary questions of a statistical rather than a clinical nature, closely related to the primary one. The secondary question(s) are tackled by collecting measurements on the secondary variables which are *either supportive measurements related to the primary objective or measurements of effects related to the secondary objectives* (ICH E9). A good example of a secondary question for a cross-over study might be the presence of carry-over, i.e. the persistence of the current treatment activity to subsequent treatment intervals. Other examples from the same field concern presence of time trends with treatment, especially in multi-period cross-over designs. A typical example of a primary question drawn

from parallel group trials, is the reduction in mortality rate caused by the new therapy. A secondary question of interest then might be how risk factors causing death differ between the competing therapies. Another type of secondary question, relevant to both cross-over and parallel group trials, concerns treatment effectiveness across different sub-groups. Methodological issues arise if lots of statistical tests are performed on various sub-groups, since some of these tests will incorrectly show a statistically important treatment effect. This is the issue of multiplicity, and is usually tackled by making appropriate adjustments (e.g. Bonferonni) at the significance level the various tests are performed. Note that studying treatment effect across sub-groups is only appropriate when these sub-groups are adequately represented in the study population. The ICH E9 guidelines suggest that in most trials sub-group analysis or a statistical model that include interactions should be exploratory and any conclusion of treatment efficacy based solely on sub-group analysis should be avoided.

Although the primary and secondary efficacy questions in phase III are clearly specified, the same does not hold for the safety aspect of the trial. Recall that safety information is usually collected at Phase II, where a cross-over design may be used. Additional safety information, like adverse events and other laboratory measurements are collected during Phase III under a parallel design scheme. Most of the compounds tested at phase III using a parallel study, have already demonstrated safety during phase II using a cross-over study. This is a reason why safety comparisons are dealt less formally at phase III. Although the efficacy part of any study is a well-control experiment, adverse events or other safety measurements are of an observational nature. Adverse events occur in an unpredictable way, what causes them is unknown and their relation to the treatment, if any, is often difficult to understand.

Different types of primary response variables may be encountered in practice. The most common one, met in parallel studies rather than in cross-over ones, is the incidence of a specific event. The incidence of an event is a dichotomous variable, i.e. in statistical terms a factor with two levels. This type can be easily extended to factor variables with more than two numerical levels, which can be ordered or nominal. A third type of response is the continuous one, widely used in hyperten-

sion and asthma cross-over trials. The use of a single outcome variable to answer the primary question is favored in all types of clinical experiments (see ICH E9), since in the situation where inconsistent results are provided by the analysis of more than one outcome variable, interpretation of trial findings becomes difficult. In parallel group trials, combining events to make up a response variable is a typical practice, especially when component events rarely occur. As stated in the ICH E9 document, this approach addresses the multiplicity problem without requiring adjustment to the Type I error rate. Difficulties with that practice arise when component-event analysis, if at all possible, give different results compared to the combined-event one. An hierarchy of the component events, established in advance, could be the answer to the problem. Aggregation of measurements in cross-over studies is a within-subject process, and usually occurs only when repeated observations are made within a given treatment period. Aggregation of measurements across subjects is not commonly met in the cross-over literature. Cross-over trials are not appropriate for diseases where the primary measurement is death, diseases where a long treatment period is needed, diseases where the effect of treatment is irreversible or diseases where gradual deterioration in patient's health is observed. In some clinical trials, the therapeutic ability of the tested compound is quantified by obtaining measurements closely related with the drug activity at the site of action (receptor). For example in asthma trials peak expiratory flow, a measurement of lung function, is compared between competing therapies. In the majority of clinical studies, either cross-over or parallel group ones, instead of studying the clinical endpoint of most interest another response variable, called a surrogate variable, with strong predictive ability for the primary clinical endpoint, is measured. A good example is HIV trials, where monitoring the incidence of AIDS is commonly replaced by measuring the change in CD4 cell-counts. According to the ICH E9 guidelines surrogate variables can only be used, if the biological plausibility of the surrogate and the clinical outcome has been demonstrated, or if there is conclusive evidence from epidemiological studies that the prognostic value of the surrogate variable on the clinical outcome is high, or other trials have shown that treatment effects on the surrogate measurement correspond to effects on the clinical outcome. If a surrogate variable

is used, investigators have to make sure that the surrogate measurements can be taken accurately and reliably, without the need for expensive equipment and highly trained staff. Finally, trial participants should feel comfortable with the procedures undertaken during the measurement process.

Phase III results are the main information submitted to regulatory authorities for licensing a drug to the market. Usually these results may change the current clinical practice and long-term surveillance of the proposed therapy is absolutely a necessity. Parallel group studies are favored by sponsors and regulators during this stage.

1.2.4 Phase IV

During phase III the investigator assess not only the clinical benefit of the new therapy on the population at risk, but also any unwanted effects. This information helps regulatory bodies to decide under what circumstances the new therapy should be recommended for use. The cost of the proposed therapy to the general public is a further dimension of the decision making process. The general public will not be willing to pay for highly expensive agents, especially when they are of limited clinical benefit compared to existing treatments. Currently cost evaluation is not an integral part of the marketing approval process, though ICH E9 guidelines suggest that Phase IV studies are useful for optimizing drug's use in a subject-level but also in society.

The cost of treatment to the general public should not be the only factor to be considered for licensing or not a compound. Improvements on the quality of life of study participants is another dimension that regulatory bodies should consider. A cross-over trial can be used for assessing improvements in various quality of life dimensions between the standard and the newly proposed therapy, since patients try all available treatments at least once. There are various dimensions to the "quality of life" concept though. To begin with, the individual's ability to perform daily life activities (e.g. bathing, dressing) is referred to as the "physical" dimension. Next comes the "psychological" component, referring to emotional and mental well-being. The new treatment may cause side effects such as depression or anxiety affecting in a negative way a participant's daily life.

Finally, there is the "social" component, i.e. the person's willingness to participate in family or other social activities, maintenance of any working obligations at a satisfactory level and the way one interacts with the community in general. Further dimensions of quality of life, of secondary importance, include the effect of treatment on the cognitive abilities of a participant (memory, recognition, e.t.c), sleep patterns, pain related to specific physical activities, failure to form and maintain personal relationships e.t.c.

It has to be mentioned that such life-quality assessments may not only be collected at phase IV. On some occasions they might be the primary response variable in a phase III program, where a parallel group study is used for assessing the primary question. Personally, I have been involved in a parallel group phase III study for comparing a newly form compound against placebo for stroke patients. The primary outcome variable was the Barthel index, a measure of physical functioning and independence. On the other hand, if the primary outcome at phase III is of a clinical nature (e.g. a new anesthetic for use in surgery), then quality of life measurements may be collected at a post-surgical phase (phase IV). A cross-over plan might be used to that purpose.

Medical life-quality data related to either financial or personal costs, are collected either from interviews or questionnaires sent by post. Questionnaires have the advantage of being a cost-effective data collection process and it is also more likely to derive answers to sensitive questions. However, face-to-face interviews tend to provide investigators with complete information that can be used for further analysis. Special attention needs to be given to accurate collection of life-quality data, since in the majority of clinical studies investigators collect cautiously all clinical information relevant to the treatment under study, but this is unlikely for non-clinical data. Quality of life data are not used in regulatory submissions and that is why investigators are often careless in collecting them. If a cross-over design is used for conducting a phase IV trial, then information from the subject with incomplete set of measurements will contribute to the overall treatment assessment. This would have not be the case if a parallel design had been used instead.

A score is usually attached to each dimension of health life-quality data. For

example, the physical functioning score could be the sum of scores on various daily activities. The same principle applies to scores in other dimensions. This is an example, where methodologies for the analysis of ordinal categorical data in cross-over trials become relevant. A review of that literature is given in the next chapter. From the statistical perspective, difficulties arise in interpreting differences between sub-groups on a given scale. For example, does the observed change in the physical functioning score reflect a clinically important improvement in a participant's life? Lack of interpretation may lead to difficulties in evaluating the size of a trial, if the physical functioning score is the primary outcome. In summary, lot's of research effort needs to be placed in incorporating health related quality of life measurements smoothly into current clinical trial practice.

1.3 Selecting an appropriate population

The information presented in this section is relevant for both parallel and cross-over studies.

In the majority of clinical experiments, the compound under study is working for the population it has been tested. Participants randomized into the treatment phase (study sample) are a subset of the study population, i.e. patients that baseline characteristics obtained but failed to enroll into treatment phase for various reasons. The study population is in turn a subset of a wider population consisting of patients with the medical condition under study, but not eligible to enter the trial. Generalizing results found on the study sample to the study population is legitimate, as long as the study sample is a representative sub-sample of the study population (see ICH E9). The eligibility criteria that separate the study population from the population with the medical condition under study present, should not be excessively restricted, since difficulties in getting sufficient number of participants will arise. On the other hand, if loose inclusion/exclusion criteria are set beforehand, inappropriate participants may be admitted into the study, the sample size will rapidly increase but the probability of observing the primary response outcome will decrease. ICH E9 guidelines suggest that a confirmatory

trial may be helpful for selecting the patient population for which the drug will eventually be indicated.

The entrance criteria are easier to set if the mechanism of action of treatment is known to some extent and the investigator is able to identify a relatively homogeneous population likely to respond to that treatment. On the other hand, treatment efficacy should be demonstrated on a study population where members may differ on one or more aspects of the medical condition under study (e.g. severity of disease). In that case, a heterogeneous group of participants will be collected. In large clinical trials it is more likely to have an heterogeneous rather than a homogeneous group of patients. Cross-over trials have a distinct advantage compared to a parallel design for comparison of treatment effectiveness across various sub-groups in an heterogeneous population. This is because within a sub-group, say males, treatment effect is assessed more precisely since within patient information is utilized. This results in a more accurate assessment across sub-groups (males vs females). It has to be mentioned though that even for a cross-over trial, if treatment effect within sub-group(s) is of interest, the number of patients recruited to adequately power such a study can be enormous.

1.4 Procedures needed to be followed before a study starts

Once the experimenter selects the study-design the next step is to assign the chosen study-population to the various sequences of the chosen cross-over design. The allocation process should be unpredictable, so that experimental bias is avoided. Experimental bias simply means that the decision to randomize or not a subject in a given treatment sequence depends upon this sequence. Obviously, no randomization scheme can guarantee perfect balance on other risk/prognostic factors, but the larger the study is, the more likely the imbalance issue to be resolved for various factors (see ICH E9 guidelines).

There are various ways to randomize patients into a cross-over study. The simplest scheme assigns participants to the various sequences with equal probability. The real advantage of that scheme is ease of implementation, though in groups

with different demographic or other characteristics not adequately represented, substantial imbalance across sequences may occur. Block randomization ensures that imbalance will not be large at any time during the randomization process. The idea is to split the number of eligible subjects into blocks of size equal to a multiple of the number of sequences of the cross-over plan adopted and then within each block equal allocation of patients to sequences occurs. The main advantage with block randomization is that each sequence will be approximately equally represented, if the trial is terminated early for any reason, or type of participants changes during recruitment (e.g. males recruited earlier than females). The investigators though should be blinded to the block size or, if that is not possible, the block size should vary as recruitment continues. As ICH E9 guidelines suggest block sizes should be sufficiently small to avoid possible imbalance, but should be sufficiently large to avoid predictability of treatment sequences towards the end of the randomization process within a block. Blocking maintains balance representation of the various sequences and it is usually taken into consideration in the statistical analysis.

Stratified randomization involves performing sequence randomization within strata defined by selected prognostic or risk factors. Usually the chosen factors are expected to correlate highly with the primary response variable. Simple or blocked sequence randomization is performed within each stratum, although the blocking strategy is usually preferred so that less sequence imbalance occurs in strata with fewer participants. Obviously, as the number of risk factors of interest increase and the levels within factors grow, the number of strata expands rapidly. Only important risk factors should be chosen, so that the number of strata is kept to a minimum (see ICH E9 document). Factors used to perform a stratified randomization should be included in any statistical model thereafter, but one should keep in mind that these factors affect estimates of between subject contrasts rather within subject comparisons. A special example of a study where stratified randomization occurs is the multi-center trial. In that case ICH E9 guidelines recommend that several whole blocks of treatment sequences should be assigned to each center, while randomization procedures should be organized centrally.

Modern randomization schemes have been proposed, though they are of limited

practical use. A famous one, originally suggested by Efron (see [10]), assigns treatments within sequences sequentially as the trial progresses. Assignment of the next treatment regime is based upon previous treatment assignments for that subject, but not on his responses already observed. The allocation probability p to group A (or B) within a sequence is adjusted continuously, so that next treatment assignment is more likely to occur to the treatment group with fewer past appearances on that subject. Randomization strategy should be taken into consideration during the analysis, otherwise the p-value reported will be slightly larger than if the correct analysis was performed. More advanced adaptive randomization schemes make use even of the past responses collected on a subject, in order to decide the treatment allocation in the next period. The play-the-winner rule assigns a subject to the same treatment group as in the previous period, if that treatment has been successful on the previous period; otherwise the participant is assigned to the other treatment group. These schemes were motivated by ethical concerns, since one may wishes to maximize the number of times a patient receives the superior treatment. A major obstacle in implementing these schemes is that response may no be immediately available and it is not yet clear to the statistical community how to take into account the randomization process, in the analysis.

One of the main pre-cautions taken to reduce bias is to keep both patients and investigators blinded to treatment. Most of the efficacy trials are double-blinded ones. If investigator ignores the treatment a patient is receiving, then he is expecting to act in a similar way regardless of the treatment the patient is receiving within a treatment period. Double-blind trials are usually more difficult to carry-out than trials where a simpler blinded scheme is adopted. The key to truly blind a study is to have medications with similar appearance. This may not be possible, unless interference with the treatments occurs to an appreciable extent. The technique of double dummies is then used, where placebos with similar appearance to the products under study are administered simultaneously with the treatments (ICH E9 glossary). Both investigators and patients may try to discover drug's identity. For cross-over studies where patients try both medications, matching drug appearance is crucial, since patients can make their own

comparisons. Appearance is one of the many characteristic that agents should be matched, taste and weight are two others. Finally, a procedure should always be in place to unblind quickly for any individual at any time, if that is necessary. In summary, both randomization and blinding contribute to the quality of the collected data and validate conclusions drawn from the analysis of the study. As ICH E9 guidelines suggest randomization and blinding should be normal features of any controlled clinical trial intended to be included in a marketing application.

1.5 Quality Recruitment

This section contains material that is applicable to both parallel and cross-over studies.

In any clinical study, obtaining sufficient number of participants within a reasonable time period is the key for successful completion of the program as a whole. First of all, the time the recruitment period lasts should be set well in advance. Investigators must make every effort to enroll participants in a timely fashion. Extending the recruitment beyond the originally planned period increases costs and decreases participant's and investigator's morale. Inadequate planning, failure to start on time and under-estimating the importance of factors that may have accelerated the recruitment process if considered promptly, are a few of the primary reasons for recruitment failure.

Realistic estimates of the potential number of participants can only be made by tracking hospital or physicians records. Making the trial publicly known through scientific meetings or media campaigns may increase participation rates. If data sources concerning recruitment are difficult to obtain, then a pilot study (or confirmatory study as mentioned in the ICH E9 guidelines) can be set-up to provide valuable information on best recruitment techniques and yield estimates of potential participants.

There are various strategies to recruit subjects to a clinical study. The strategy chosen, usually depends upon the type of the trial (single or multi-center), the length of the available time and the general setting. The first step in a traditional recruitment process is to identify groups of potential participants in hospitals,

patients of physicians or employees in various organizations. After passing an initial screening test, patients are formally invited to undertake a further eligibility evaluation. An alternative strategy is to bypass the initial testing process and directly invite patients into the program. Sponsors should always remember that techniques achieving high recruitment rates within a geographical area, may completely fail in doing so in other areas. Modifications to the recruitment strategy should be made where necessary. For cross-over studies the length of wash-out period between successive active treatment periods should be carefully chosen so that higher drop-out rates are avoided.

If recruitment is delayed, reasons should be identified why this is the case. In multi-center studies, cites that perform poorly can learn from cites where recruitment performance is excellent. Graphs showing actual recruitment compared to originally planned are useful tools for identification of potential problems. If a center cannot contribute enough participants then it is highly likely to drop-out from the study. For cross-over studies there might be specific treatment intervals where withdrawal rates are high. The knowledge of that information may result in improving the design of future cross-over trials.

One way of tackling lagged recruitment is by relaxing inclusion/exclusion criteria. This will increase the study-population, but the incidence rate of the primary outcome in the new participant-type may not be as large as in the original participants. ICH E9 guidelines suggest that changes in the inclusion/exclusion criteria may be appropriate when knowledge from outside the trial or from interim analyses indicate that this is the right course of action. Another viable solution is to extend the recruitment time or add more recruiting sites. Obviously this increases the overall study-cost and it will inevitably delay publication of study-results. A further approach to the problem is to recycle potential patients, i.e. giving persons who are interested in participating in the study a second chance. Sometimes, accepting a smaller number of patients is the right course of action. Reducing sample size deliberately, has the effect of lowering the power of the study. If on the other hand, treatment effect is higher than originally anticipated, this solution will provide comparable power. On the other hand ICH E9 guidelines suggest that if sample size calculations have been performed using uncertain in-

formation, a revised sample size may be calculated using modified assumptions. This change though, should be documented both in a protocol amendment and in the final study report.

1.6 Collecting Quality Data

Problems in data collection for cross-over studies can be of several sorts. Examples include incorrect data, missing data or data with greater variability than expected. It is essential that inferences from the study are based on accurate and valid data. Key data, like baseline characteristics, primary and secondary outcome measures, should be error-free. Missing data usually arise from inability of physicians or participants to complete questionnaires. Missing data are commonly found in late follow-up measurements, since as the trial progresses participants fail to meet the standards of adherence as established by the investigator. This point is especially relevant to multi-period cross-over studies. Patients with incomplete information still contribute to the overall assessment of various terms in cross-over studies. It has to be noted that the higher the percentage of missing data the less credible are the conclusions drawn from the study. Universally accepted methods for handling missing data cannot be recommended, though ICH E9 guidelines suggest that methods of dealing with missing values should be pre-defined in the protocol and the sensitivity of the results of analysis to the method of handling missing values should be examined.
Incorrect data, on the other hand, are not easily recognized. They usually arise as measurements obtained by clinical staff or technicians using a different definition than the one described in the protocol. Once the error has been spotted, feedback to the personnel responsible for collecting the data should be given immediately, so that the correct value identified and entered into the database. Incorrect data usually appear in a statistical analysis as outliers. The definition of an outlier is arbitrary. Characterization of a value as an outlier should be justified both medically and statistically (ICH E9 document). Regulator bodies favors testing the influence of outliers on the final results, by performing at least two analyses; one with the actual values and another one which eliminates or reduces the outlier

effect. For cross-over studies the presence of outliers generates spurious interactions. Jones and Kenward (see [39]) gives an example where an incorrect value for a subject generated a statistically significant treatment by period interaction, which in the 2x2 case is equivalent to the carry-over effect.

The majority of clinical trials are repeated measurements studies. The variability between repeated assessments on a subject can be of systematic or random nature. In cross-over studies systematic variation can be attributed to different treatments assigned at different time points, while random variation may represent the physical condition of the patient, errors due to the instrument used for the measurement or errors of the clinical staff responsible for data collection. Clinical staff get more experienced with trial procedures as study progresses and this accounts for intra-observer variability. However, depending on level of knowledge and expertise, people will perform the same task differently within the same working environment. This will account for inter-observer variation. Inconsistent behavior of the same clinician or of clinical staff working in the same team, may alert to the need for thorough checking of the collected data.

Certain steps have to be taken in order to minimize the collection of poor quality data. A manual of operations is usually prepared for any clinical trial, where detailed description of participant's visit and the procedures followed during these visits can be found. Questionnaire forms should always derive the key information, being well-organized and have a logical sequence. Standardization of interviewing techniques, laboratory tests and other procedures are crucial to the success of any large study. Finally a typical technique to reduce variability is to repeat the assessment, if at all possible. For example, blood pressure could be measured twice and the average reported.

Monitoring the areas most important for the study, is the key action to obtain high-quality data. Clinical staff, on a regular basis, should receive reports of weaknesses or errors blinded to treatment. Personal experience, suggests that date of event(s) is unlikely to be the same, if reported in two different forms. In follow-up assessments, especially in cross-over designs, it may be the case that missing or late participant visits may be associated with the treatment administered. If that is the case, then the final conclusions drawn from the study will

be biased. Laboratory measurements are good examples where extreme values can be mistakenly recorded. Laboratories should make sure that equipment has been tested, been well-calibrated and appropriate adjustments in scales have been made where necessary. Finally, auditing sites may improve data quality and trial conclusions.

What really can make a difference in collecting quality data, is the adherence of the study-participants to the protocol. Participants may not be willing to be compliant with the study procedures for various reasons, for example they experience unpleasant side effects, or compliance with protocol requires changes in their daily lives, or they may mis-interpret instructions given to them, or their health deteriorates during the study-course regardless the treatment group they have been assigned to. Obviously, shorter studies has greater advantages over longer ones. Also, hospital-based trials tend to have less non-adherence problems than home-based ones. In addition, keeping dose-regimen as simple as possible helps in the derivation of complete data.

1.7 Monitoring large studies

This section concern monitoring of Phase III studies, which are not conducted using a cross-over design, as has already been mentioned. For purposes of completeness though some account of my personal involvement in such studies will be given.

The credibility of a trial is enhanced if the persons who monitor the efficacy and safety variables have no formal involvement with either the participants or the investigators. Data monitoring requires collection and processing of the relevant information in a timely fashion, otherwise monitoring would be of limited value if carried out at a stage where the majority of the data have been collected. Investigators cannot have the monitoring responsibility, since they may discover that treatment A is more effective than treatment B, while participants are still enrolled into the study. Interim analysis results are used to decide whether to continue, terminate or modify the design of an ongoing study. I have been personally involved in the safety and executive committees of an ongoing diabetes study and

the format used during these meetings involves an open and a closed session. In the open session, recruitment status, data quality and other issues that affect the outcome of the trial are considered. In the closed session, baseline characteristics, primary and secondary outcome variables, adverse events and other safety measurements are compared by treatment groups. In the closed session, key members of the committee only decide continuation or premature termination of the study, based on careful review of the interim analysis results presented to them. According to the ICH E9 guidelines, the monitoring committee is responsible for setting operating procedures and maintain records of all its meetings, while the role of each member of that committee (sponsor staff inclusive) should be clearly defined.

An issue that needs to be resolved is how the results of any interim analysis will be presented to the members of the committee. Early in the trial, where the two treatments are expected to be equally efficient (or inefficient), there is no reason to identify the two groups in each table or figure of the report. When one of the two competing therapies show its superiority the committee-members should have full knowledge of the group identities. Usually annual reviews of study-progress suffice to resolve any issues, while in other occasions meetings are scheduled when a specific proportion of the outcome variable has been observed (e.g. 25% of deaths). From a statistical perspective, if the null hypothesis of no difference between the two treatment groups is tested at the same level of significance using accumulated data, then the probability of incorrectly rejecting the null will be higher than the nominal level. Group sequential methods, described by Jennison and Turnbull (see [35]), where the number of interim looks is taken into consideration for setting the significance level at each look as data accumulate, ensure that the overall significance level for the trial remains at the desired level.

The decision to terminate/continue a study will be based on various factors. The extent to which the new therapy is beneficial is one determinant factor. The incidence of serious adverse events in the two treatment groups may force early termination for safety reasons (see ICH E9 guidelines). If in one of the latest interim looks it becomes clear that it is impossible to see a beneficial effect if the

trial continues to the end, then terminating the study has financial advantages for the sponsor (see ICH E9 guidelines). Finally, logistical problems not foreseen during the design phase, may suggest that study continuation is not feasible. In the interim analysis results, possible differences of various prognostic factors at baseline between the two treatment groups should be considered. In addition the impact of missing data on the analysis should also be evaluated. Secondary response variables should be analyzed along with the primary ones. Consistency of results across dominant sub-groups or across different centers should be examined. The decision to terminate a study should not be based on unexpected results in small sub-groups.

1.8 Concluding remarks - Thesis outline

In this chapter I have tried to summarize my three-year involvement in designing and analyzing clinical trials. A lot of topics have not been discussed. For example, different type of designs that can be used to run a study, different sample size formulas the statistician can use depending on the type of the primary response variable, the issue of using baseline measurements as part of the response or as covariate trying to explain variability in the response, setting significance levels for repeating testing, sub-group analyses, comparison of multiple primary response variables, meta analysis, multi-center trials, reporting and interpretation of trial results and many more.

This thesis concerns cross-over studies. A full review of the cross-over literature is provided in the next chapter. In chapter three, the 2x2 cross-over design is studied in depth. Properties of treatment effect estimates under different carry-over assumptions are presented and a newly proposed treatment estimate is studied. The two-stage procedure is discussed in detail and properties of a corrected two-stage scheme are presented. A trial in asthma is then analyzed using both Frequentist and Bayesian methodology. The use of baselines as part of the response leads to a three stage scheme for comparing two treatments, the properties of which are fully evaluated. The baseline measurements enrich the assumptions that can be made for carry-over term(s) and analysis of the same 2x2 trial in

asthma with baselines now incorporated is presented. Bayesian and frequentist analysis when baselines are considered as covariates are also discussed. The impact of covariates on the cross-over trials is also evaluated in some depth. Finally, a non-linear model, where carry-over is modeled as a proportion of treatment effect is presented and the same trial in asthma is analyzed using that model. A model selection exercise using the AIC criterion is performed, for comparing linear and non-linear approaches.

In chapters four and five attention focuses on selecting the best design for running a cross-over study under different assumptions concerning residual effects. In chapter four, clinical justification for the carry-over assumptions made is presented. Assumptions in both the systematic and random part of the model are reviewed and optimal plans are presented for comparing two treatments for design families with limited number of periods and sequences. The impact of model mis-specification in designing a cross-over study is fully evaluated. More specifically, the model used to design the study may be different from the one used to perform the analysis. Optimum plans, where not only the systematic but also the random part of the model is mis-specified, are given. Finally, analysis of a cross-over study with seven treatments where carry-over effects depend on the type of treatments administered in the current and previous period is also presented. An account of the design literature for repeated measurements studies concludes the chapter. In the fifth chapter, optimum plans for the comparison of two and more than two treatments are presented under different carry-over assumptions. Cross-over plans with moderate number of periods and sequences are studied using an optimality criterion widely encountered in practical applications. When three or more treatments are compared, cyclic designs are only considered. Finally, optimum plans for the non-linear model studied in chapter four are derived. In the final chapter, conclusions of the whole thesis are presented and future research directions are given.

Chapter 2

Cross-Over Trials - A Review

2.1 Types of Clinical Trials

The most common type of clinical trial is the randomized control study, where participants are assigned randomly to a treated or a control group. Most tests used for the analysis of this experiment, like the t-test, can be justified on a randomization argument only, without further assumptions needed to be made on the measured variables. The majority of clinical investigators feel that patients should receive the newly proposed therapy, regardless if that therapy has demonstrated its effectiveness in real life situations. These investigators will not be willing to participate in a trial. From an ethical point of view, investigators who are in doubt about which therapy is superior can possibly participate in the study to settle the question.

In some studies, randomization does not take place. Participants are assigned to the two treatment groups without use of a random allocation scheme. For example, data on the success of a new surgical procedure will only be collected at the institution the new method was applied. Results will then be compared with patients in other hospitals, where a more traditional medical care was implemented. In these studies, patients in the two groups are matched by key characteristics. Matching on some of the important prognostic factors may be impractical, while evaluation of the impact of other equally important characteristics on the outcome response may not be possible.

Another well-known type of study is the withdrawal ones, where patients are

taken off therapy in order to evaluate the duration of benefit of the treatment. Study-population consists of patients who have experienced a treatment benefit for several years.

The purpose of setting-up a factorial design is to evaluate three or more treatments in one experiment. This will reduce the cost and the effort required to compare competing therapies in separate experiments. The only disadvantage with this type of study is the possibility of interactions being present, i.e. treatment A has a lower response when administered in conjunction with treatment B rather than with treatment C. The power for testing for interactions is always lower than for testing main effects. A factorial study adequately powered to detect interactions would require number of participants equal to the sum of the participants of the separate studies. ICH E9 guidelines mention another example of a factorial design; the dose-response trial. In this type of study, a number of m doses of drug A (placebo inclusive) and similarly a number of n doses of the alternative therapy B (placebo inclusive) are selected. Patients are randomized in one of the mxn possible treatment groups. The data collected are used to give an estimate of the response surface and then an appropriate combination of doses of A and B is identified for clinical use. In other trials the basic sampling units are groups rather than individuals. In these plans, called cluster randomization designs, a whole group of individuals (e.g. center) is randomized to one of the two treatment groups. Types of clinical trials, where cross-over design is used extensively, have been presented in the previous chapter. For an extensive review see Senn [80].

2.2 Cross-Over Plans

The 2x2 cross-over trial will be properly studied in the next chapter, though some account of the existing literature will be presented in this chapter as well. In this type of trial each participant receives some or all of the competing therapies. The order in which treatments are administered to participants is randomized. This type of trial has some appeal to the medical community, since each participant is used more than once and comparisons between different treatments

are individual-based. The carry-over effect, i.e. treatment activity in the current period persisting in subsequent periods, has played a key role in evaluating the usefulness of a cross-over plan in medical practice. As ICH E9 guidelines suggest, when a cross-over design is used it is important to avoid carry-over by allowing for sufficiently long wash-out periods. Loss of subjects can be an additional problem in using a cross-over study. An area where the 2x2 design has been successfully applied is to demonstrate the bio-equivalence of two formulations of the same medication.

2.2.1 Parallel vs Cross-over design

Brown (see [3]) was the first to compare the 2x2 cross-over design (2 measurements per participant) with a parallel group study (1 measurement per participant) in terms of cost-effectiveness. He assumes that the model generating the data for the cross-over experiment, contains a term for the mean, period, treatment and carry-over effect, while the subject effect is taken as random. If n subjects are randomized in each sequence of the cross-over experiment, while m in each group of the parallel study, then the two treatment estimates derived from the two trials will be equally efficient, if the following relationship is satisfied between the sample sizes:

$$n = (1-\rho)\frac{m}{2} \tag{2.1}$$

where ρ is the correlation between measurements on a subject in the cross-over experiment. Now, let S_0 be the cost of recruiting a new participant and S_1 the cost of treating and measuring the patient in a given period. Then the relative cost of the cross-over relative to the parallel group study is:

$$R = (1-\rho)\frac{1+2S_1/S_0}{1+S_1/S_0} \tag{2.2}$$

From that equation, Brown concludes that if recruiting a patient is more costly than obtaining follow-up measurements and/or there is large between subject variability, then cross-over is a more economic solution compared to the parallel group study. Brown's approach was interesting, since the two designs were compared on economic rather than statistical grounds.

2.2.2 Multi-stage procedures in the 2x2 case

The problem with the carry-over in the 2x2 case is usually tackled by proposing multi-stage procedures. Two of them, Grizzle's (see [30]) and Kenward-Jones's (see [42]) are studied in the next chapter. Lehmacher (see [55]) suggests another two stage scheme, which is a modification of Grizzle's proposal. More specifically, if τ and λ are the treatment and carry-over effects respectively, then the joint hypothesis $H_0 : \tau = \lambda = 0$ is tested at level a using Hotelling's T^2-test. If it is significant then four separate hypotheses about treatment, carry-over, bias of treatment estimate $(\tau - \lambda/2)$ and second period difference $(\tau - \lambda)$, are all simultaneously tested at level a. Lehmacher argues that his multiple test procedure preserves the nominal level of significance for treatment difference, but the power of the scheme was not evaluated in the original paper. Lehmacher's approach can be seen as an updated version of Willan's statistic (see [91]), where the maximum of two treatment estimates (CROS, PAR) is used for testing treatment difference at half of the nominal significance level. Jones and Lewis (see [40]), by using a simulation based approach, compares the power of Grizzle's, Willan's and Lehmacher's procedures and concludes that Grizzle's is the best while Lehmacher's the worst. The Type I error rate is not reported for any of the above schemes though. Willan (see [91]) argues that his procedure achieves the nominal significance level, and his treatment estimate compares favorably to the CROS estimate in terms of power and MSE, when carry-over is a small proportion of the treatment effect.

2.3 The 2x2 case with baselines

2.3.1 Baselines as part of the response

In the 2x2 case a "run-in" and a "wash-out" period are included and measurements are collected during these intervals. On other occasions, a wash-out period is not possible and the baseline measurements obtained in the run-in interval can be used as covariates in the analysis. Four repeated measurements can be collected on each subject and Kenward with Jones (see [42]) provide an extensive

account of the covariance structure that can be assumed on the vector of observations on a particular subject.

One of the special forms this structure can take, is a stationary first order autoregressive (AR(1)), where correlation between repeated measurements depends on their distance in time, plus a random subject effect. The length of the washout interval can be incorporated into that structure. Kenward with Jones did a detailed investigation on what particular structure could be recommended for future use, by analyzing data on 2x2 cross-over studies where baseline measurements were available. Unfortunately no particular structure emerged and in a few cases none of the structures considered fit the data particularly well. Kenward with Jones go even further proposing a three stage procedure for the analysis of cross-over data with baselines. Jones and Lewis (see [40]) using a simulation based approach studies the properties of the three stage procedure without reporting the Type I error rate. An analytical approach is used instead in the next chapter to study that procedure and the Type I error rate is also evaluated. The interesting point in Kenward and Jones's work is that GLS estimates of parameters are equivalent to their OLS counterparts, appropriately adjusted for baseline readings.

2.3.2 Baselines as a covariate

Chi (see [6]) discusses the recovery of inter-block information in cross-over trials, without restricting the arguments to the 2x2 case only. He considers the simple carry-over model where subject is taken as a random effect. This model in a matrix notation can be partitioned into a fixed and a random part as follows:

$$Y = X\beta + C\xi + e \tag{2.3}$$

where β contains overall mean, period, treatment and carry-over effects, while ξ is the vector of patient (or block) effect. In both the fixed and/or the random part baseline measurements can be easily incorporated. Chi derives the GLS estimate as a combined estimate of intra and inter-block analysis. More specifically, the intra-block analysis offers as the following solution:

$$\hat{\beta}_{intra} = \left(X^T \left(I - C(C^TC)^{-1}C^T\right) X\right)^{-1} X^T \left(I - C(C^TC)^{-1}C^T\right) Y \tag{2.4}$$

with

$$\hat{\Sigma}_{intra} \propto \left(X^T \left(I - C(C^TC)^{-1}C^T\right) X\right)^{-1}. \tag{2.5}$$

The inter-block analysis offers the following estimate

$$\hat{\beta}_{inter} = \left(X^TCC^TX\right)^{-1} X^TCC^TY \tag{2.6}$$

with

$$\hat{\Sigma}_{inter} \propto \left(X^TCC^TX\right)^{-1} \tag{2.7}$$

If $\hat{\beta}_{intra}$ and $\hat{\beta}_{inter}$ are independent, the combined analysis gives the following estimate of β:

$$\hat{\beta}_{GLS} = \left(\hat{\Sigma}^{-1}_{intra} + \hat{\Sigma}^{-1}_{inter}\right)^{-1} \left(\hat{\Sigma}^{-1}_{intra}\hat{\beta}_{intra} + \hat{\Sigma}^{-1}_{inter}\hat{\beta}_{inter}\right) \tag{2.8}$$

which is the same as:

$$\hat{\beta}_{GLS} = \left(X^T\hat{var(Y)}^{-1}X\right)^{-1} X^T\hat{var(Y)}^{-1}Y \tag{2.9}$$

Chi concludes that when missing data are available, recovering the inter-block information may not be worth while, although he does not report any conditions, under which recovering such information may prove beneficial.

Senn (see [75]), discusses the use of baselines in asthma trials. He argues against the idea of correcting for baseline by subtracting the baseline measurement from the measurements obtained during the trial's active treatment period. This action has the benefit of reducing the variability in the analysis variable, though it creates spurious correlation between the analysis variable and the baseline reading. Senn seems to favor an analysis of covariance method, where instead of subtracting the baseline measurement an estimated fraction of it is subtracted.

2.4 Simple extensions of the 2x2 design

Laird et al (see [50]) derive treatment effect estimates for two period designs by extending the number of sequences. Balaam's design is a four-sequence cross-over plan (AA, BB, AB, BA), where carry-over effect can be estimated using within subject information and a treatment by carry-over interaction can also be

considered. If n subjects are allocated in each sequence group, then the treatment estimate has variance

$$\text{var}(\tau) = \frac{\sigma^2(1-\rho^2)}{4n}\frac{1}{1+(1-\rho^2)} \tag{2.10}$$

where σ and ρ are estimated as shown in the previous section.

Koch's design is a six-sequence two-period plan for the comparison of three treatments. Two of them are active compounds (labeled A, B) while the third is the standard therapy, S. Interest is focused on the comparison of the two active treatments. So, if n subjects are allocated in each one of the sequence groups (AS, SA, BS, SB), then nm allocated in each one of AB and BA. This is an incomplete block design where sequences are not equally replicated. The simple carry-over model assumed throughout may not be appropriate for this design, especially if the standard therapy is placebo. Laird et al derive expression for the efficiency of the contrast $\tau_A - \tau_B$, although he notes that Koch's may not be an appropriate design for efficiently estimating carry-over effects.

Ebbutt (see [9]) was one of the first to analyze data on three-period cross-over plans for comparing two treatments. He considers a design with two sequences (ABB/dual) and a design with four sequences (ABB, ABA, duals). The two sequence design has been proved to be universal optimal for estimating treatment effect (Laska et al, see [52]), irrespective if one includes carry-over effects or not in the model, and whether or not baseline measurements are available. Ebbutt defends the four-sequence plan on the grounds that treatment by period interactions are now estimable, although he does not include such a term into his model. In addition it would be more difficult for the investigators to break the randomization code, if the four sequence plan is used. In the four-sequence design, n subjects are allocated in each sequence group, while $2n$ subjects are assigned in each sequence of the two-sequence plan. The simple carry-over model with fixed subject effects is assumed throughout. The two plans are roughly equally efficient for estimating treatment and carry-over differences, though the two-sequence plan has the additional advantage that treatment and carry-over estimates are orthogonal to each other. A more detailed investigation of these and other plans, under various model assumptions is provided in the next chapter.

Hafner et al (see [32]) analyze data on two group of mice under different experimental conditions. From each group, equal number of mice are randomly allocated to one of the two sequences of the following cross-over plan: ABA/dual. Hafner et al assume random mouse effects, and his analysis is based on transforming the $3x1$ vector of original responses to a new response, by multiplying it with a $3x1$ vector of coefficients of an appropriately selected within-subject linear function. The summary response, is then analyzed using a typical ANOVA method or a Wilcoxon rank sum test. The interesting point about that work, is that more than one linear function could be available for estimating the same effect and the most efficient one should be used. Senn and Hildebrand (see [81]) considers a similar approach to that of Hafner et al, by analyzing a three-period three-treatment cross-over trial in asthma. The 3x1 vector Y_{ij} of the j^{th} patient in the i^{th} sequence is modeled as follows:

$$Y_{ij} = \mu 1_{3x1} + s_{ij} 1_{3x1} + P + T_i \tau + \epsilon_{ij} \tag{2.11}$$

where μ the overall mean, s_{ij} the random subject effect, P the period matrix, T_i the treatment matrix and ϵ_{ij} the error vector. Contrasts, defined on each subject, estimating treatment or other effects of interest can be expressed as:

$$Z_{ij} = C^T T_i^T Y_{ij} \tag{2.12}$$

All six possible treatment sequences are used in this study. Treatment effect is estimated orthogonally to the carry-over one for this design.

The problem of demonstrating equivalence between a reference (R) and a test product (T) has a long history. Vuorinen and Turunen (see [89]) propose a three-stage procedure for bioequivalence assessment using the two period cross-over model. As usual, the Type I error rate and the power of the proposed scheme are not reported. The model assumed, allows not only means but also variances to depend on the treatment administered at a specific period. For a subject who has been randomized in either treatment-sequence, it is assumed that

$$y_T \sim N\left(\mu_T = \mu + \tau_T, \sigma_T^2 = \sigma_S^2 + \sigma_{eT}^2\right) \tag{2.13}$$

$$y_R \sim N\left(\mu_R = \mu + \tau_R, \sigma_R^2 = \sigma_S^2 + \sigma_{eR}^2\right) \tag{2.14}$$

where σ_S^2 is the intersubject variance, while $\sigma_{eT}^2, \sigma_{eR}^2$ are the intrasubject ones for the two therapies. The correlation between measurements on the same subject

will be $\rho = \sigma_S^2/(\sigma_R\sigma_T)$. The three stage scheme is based on hypothesis testing procedures for three key parameters: $\theta = \mu_T/\mu_R$, $\eta^2 = \sigma_T^2/\sigma_R^2$ and ρ. If $\theta \in (0.80, 1.25)$, $\eta^2 \in (0.70^2, 1.43^2)$ and $\rho > 0.5$ then one can claim individual bioequivalence. If the hypothesis on ρ is rejected, then one can claim population bioequivalence. If both the hypothesis on ρ and η^2 are rejected, then one claims average bioequivalence. All tests are performed at level a. The parametric version of the scheme is based on appropriately defined t-statistics for performing the various tests, while the non-parametric one is based on the Mann-Whitney statistic.

Shumaker and Metzler (see [86]) criticize the above scheme, by arguing that there are no data to question the average bioequivalence criteria set by FDA. In addition, the three stage scheme requires defining the range of η^2 and ρ and these choices usually are not based on scientific knowledge. To prove the point, Shumaker and Metzler analyze data on a four period two-treatment study, where the design RTTR/TRRT has been used. Two analysis variables are considered; area under the curve (AUC) and maximum concentration (CMAX). Surprisingly enough, the authors consider the four-period plan as two replicates of the two-period cross-over (RT/TR) design. The main reason for doing so, is to assess more accurately the within and between-subject variances, though the original four period plan is used to that purpose as well. They conclude that average bioequivalence criteria set by FDA, would have provided us with identical results compared to the individual based criteria, i.e. three stage scheme.

2.5 Bayesian approaches

2.5.1 2x2 case with baselines

Grieve (see [26]) performed the Bayesian analysis for the two-period cross-over design without baseline measurements initially. Inclusion of baselines raised the question of how period effect is modeled, since four measurements are collected per subject (Grieve, see [27]). Some authors modeled the period effect in the run-in and first treatment period using a common term (Willan and Pater, see [92]). Similarly, in the early years, it was assumed that carry-over from the first

treatment period to the wash-out interval is the same as the carry-over from first treatment to second treatment period (Chi, see [7]). Grieve follows Kenward and Jones's model, where four distinct terms are used to describe period effects and two terms are used for modeling residual effects (λ and θ). In Grieve's analysis the four measurements per subject have assumed to follow a multivariate normal distribution with a common uniform covariance matrix. Following Box and Tiao (see [2]) an ignorance prior for the model parameters is assumed, i.e.:

$$p(\mu, \gamma, \pi_1, \pi_2, \pi_3, \tau, \lambda, \theta, \sigma^2, \rho) \propto \frac{1}{\sigma^2(1-\rho)(1+3\rho)} \tag{2.15}$$

where γ is the sequence effect and ρ the intra-subject correlation coefficient. Grieve initially derives the conditional distribution of the mean parameters given the variance components. From this result, the conditional distribution of each mean parameter given the other mean parameters and the variance components can be easily evaluated. These distributions are the building blocks of an MCMC scheme, which is implemented in the next chapter. Marginal posterior densities of τ, λ and θ turn-out to be t-distributions, appropriately shifted and scaled. This model ($M2$) assumes that θ and λ are unrestricted. Grieve, considers three further possibilities: $\theta = 0$ (model M_{11}), $\lambda = 0$ (model M_{12}) and $\lambda = \theta = 0$ (model M_0). Posterior distribution of the treatment effect is also derived for the three new models and a cross-over trial in angina is analyzed for all four possibilities. A model-selection exercise using the Bayes factor approach is performed, and Grieve concludes that the models M_{12} and M_0 are the most likely to have generated the observed data. In other words, inclusion of the first-order carry-over term is feasible, while presence of the second order carry-over effect is unlikely.

2.5.2 2x2 case with missing data

Grieve (see [28]) develops the Bayesian approach to take account of missing data in the 2x2 cross-over trial without baseline measurements. For sequence i, where $i = 1, 2$, n_i subjects have complete data, n_{i1} subjects have data for the first period only and n_{i2} subjects have data for the second period only. Data are assumed to be missing at random. Uniform within-subject covariance matrix is assumed as before. Grieve derives the conditional distribution of the treatment

and carry-over effect given the variance components and the posterior density of the variance components. For the derivation of the posterior distribution for τ, σ^2 can be easily integrated out from $p(\tau|\sigma^2, \rho, data)$, but numerical methods need to be employed in order to get rid of ρ.

Grieve, then investigates the value of recovering missing data information in the 2x2 case. To that purpose, he evaluates the variance of the treatment effect estimate under three scenarios: missing values included in the analysis, completely ignore patients with missing data and finally assume that data have been available on all participants. It turns-out that missing data play an important role for drawing inference for the carry-over rather than for the treatment difference. A Bayes factor approach is implemented for choosing between two competing models, the one with carry-over term (M_1) against the model with no carry-over term (M_0). Grieve concludes that Bayes factor when missing data are considered, is close to the Bayes factor when data are available on all participants.

2.6 Frequentist Missing Data Solutions

Frequentist approaches to the missing data problem, includes the work of Patel (see [68]), who argues that taking into consideration patients with incomplete data in the 2x2 case enhances the power of the test for various interactions, like the treatment by period one, which is equivalent to the carry-over effect for the 2x2 design. Only second period data are allowed to be missing in Patel's work. Patel also assumes that no more than 40% of the study-participants are allowed to have missing values in the second period. The first period measurements, for patients with missing second period data, have the same mean and variance as the first period data in the complete cases. If u is the vector of first period data for complete and incomplete cases and v the second period data for the complete cases, then the likelihood function can be written as $f(u)g(v|u)$. MLE is used to estimate treatment and carry-over effects under the following scenarios: subjects with incomplete data are included in the analysis, subjects with incomplete data are discarded. Simulation is performed to compare the empirical with the nominal Type I error rate and to evaluate the power of the test statistics pro-

posed for the carry-over and the treatment effect, under these scenarios. These statistics have a t-distribution with appropriate number of degrees of freedom. Simulations assumed a small number of study-participants, up to 20, and a positive intra-subject correlation coefficient. When incomplete cases are included, Patel concludes that the nominal and empirical Type I error rate agree closely and do not seem to depend on sample size. In addition, the power of the test for carry-over is higher compared to the power of the test where only complete pairs are used. For the treatment effect this phenomenon is less evident.

More recent work on missing data, includes that of Richardson and Flack (see [71]), who use the design ABB/BAA to compare a newly proposed imputation approach with other established methodologies. Richardson and Flack follow closely Little and Rubin (see [57]) or Schafer (see [72]) in defining missing data mechanisms. One of them is MCAR where missing observations are a random sub-sample of the originally planned sample. A refinement of that scheme, MAR, is one where the missingness depends on the already observed values but not on the missing values themselves. Another possibility where missing mechanism depends on the missing values but not on the already observed values, NMAR-1, is also considered. Finally, both observed and missing values could drive the drop-out mechanism; the NMAR-2 type of missingness. Four analysis methods are compared; Complete Case analysis (CC), Maximum Likelihood (ML) of complete and incomplete cases, Residual Draw method with one and three imputed values (RD1, RD3). The residual draw method imputes conditional predictive mean with additional noise. Richardson and Flack consider sample sizes of appreciable magnitude, up to 80. The percentage of missing data in the second period is always lower than the third period one and may depend on the treatment administered. Compound symmetry or AR(1) structure is assumed for the within-subject covariance structure.

Richardson and Flack conclude that bias of treatment effect estimate is always lower than the carry-over estimate over all analysis methods. The CC method has the worst performance in terms of bias for both treatment and carry-over effect over all missing data mechanisms. The other three analysis methods are comparable in terms of bias. In terms of variance estimation, empirical signifi-

cance levels and power, the RD-3 and ML give comparable results closer to the nominal levels and recommended by the authors as the appropriate methods for analyzing cross-over designs with missing data. The authors, finally, recognize the importance of becoming aware of the missing data mechanism, although they suggest that investigators have information which can help in identifying reasons for patient dropout.

Another approach of analyzing cross-over experiments with missing data is the one suggested by Jones and Kenward (see [39]). Separate estimates from the complete and incomplete data for the parameter of interest are first derived and then combined using the inverse of the estimated variances as weights.

2.7 Categorical Data

Binary data are modeled by Jones and Kenward (see [38]) by using a log-linear model, where within-subject dependence is taken into consideration. Their methodology can easily be extended, when the primary outcome is categorical. Suppose that a cross-over study in s sequences and p periods is used to compare t treatments. Assume the primary response is categorical with c category levels. For the number of subjects (m_{ijk}) who fall in the k^{th} response category within the i^{th} sequence in the j^{th} period, the following log-linear model can be considered:

$$\text{mean} + \text{category} + \text{sequence} + \text{period} + \text{treatment} + \text{carry-over} \qquad (2.16)$$

In this model successive responses on the same sequence are independent from each other. By introducing appropriate interaction terms, associations between adjacent cells in the same sequence are generated. The revised model looks as follows:

$$\begin{aligned} &\text{mean} + \text{category} + \text{sequence} + \text{period} + \\ &\text{terms at the sequence by period level} + \text{period by period interactions} \end{aligned} \qquad (2.17)$$

The first term in the second line of the equation above is composed of terms like treatment and carry-over effects, while the "period by period" interaction term introduces associations between observations taken on the same sequence. This is simply a log-linear model where the joint distribution of the sequence's counts

is considered. From the above equation marginal probabilities can be derived, though their analytical expressions are awkward. Jones and Kenward (see [39]) discuss the association between log-linear model for cross-over data with log-linear models for contingency tables. Essentially, data within a sequence-period cell, can be classified according to the levels of various outcome variables (e.g. response category). Observed marginal totals can be fixed by fitting the corresponding term in a log-linear model, as illustrated in McGullaph and Nelder (see [66]) or Everitt (see [11]). Jones and Kenward (see [39]) conclude that tests concerning the statistical significance of various terms can be seen as a special case of a likelihood ratio statistic.

Instead of collapsing individual-based data into counts, an alternative approach would be to model subject-based data directly. Conditionally upon the subject effect, s_{ik}, measurements on the same subject are independent. If it is assumed that:

$$\begin{aligned} logit\,(p_{ijk} = l_j|s_{ik}) \quad &= \quad \text{intercept for category } l_j + s_{ik} + \\ &\qquad \text{effects from cell (i,j) for subject k} \end{aligned} \tag{2.18}$$

where l define the categories of the response variable, then from the conditional independent assumption, the following relationship holds:

$$p(y_{i1k} = l_1 \ldots y_{ipk} = l_p|s_{ik}) = p(y_{i1k} = l_1|s_{ik}) \ldots p(y_{ipk} = l_p|s_{ik}) \tag{2.19}$$

Now, if a probability function $g(s)$ is assumed for the subject effect, then the joint distribution of the data-vector for a specific subject is as follows:

$$p(y_{i1k} = l_1 \ldots y_{ipk} = l_p) = \int p(y_{i1k} = l_1 \ldots y_{ipk} = l_p|s_{ik})g(s)\,ds \tag{2.20}$$

Ordered categorical data can be easily incorporated into the log-linear modeling framework. To that purpose, a regression is used on the category scores, where higher order terms (quadratic, cubic e.t.c) can be included in the model. Ezzet and Whitehead (see [12]) use a random effects approach to model ordinal data in the 2x2 design. Ezzet and Whitehead explain how an ordinal variable can be derived by discretizing a continuous latent variable which follows a logistic distribution. Ezzet and Whitehead illustrate the subject-based model by analyzing data for the comparison of two inhalation devices, using a four-category ordinal

scale response. In their discussion, the authors point out that treating the outcome variable as continuous may lead to biased results and the real treatment effect may not be recovered.

A third way to analyze categorical data, is by modeling linear contrasts or directly marginal probabilities in period j of sequence i. Under this approach recovery of the within-subject dependence structure is impossible, unless higher order joint probabilities are modeled as well. Fidler (see [15]) illustrates an approach where a model with six terms is used for the analysis of binary data in the 2x2 design: sequence, period, treatment, carry-over, overall success probability and correlation between responses on the same subject. McNemar's and Gart's tests are special cases of Fidler's model.

2.8 Other types of cross-over data

2.8.1 Multivariate Data

The analysis of the 2x2 design from a multivariate perspective was first presented by Zimmermann and Rahlfs (see [93]). The authors argue that the multivariate approach has certain advantages over the univariate one, since simultaneous testing of hypothesis of interest are possible, while restrictive assumptions on the within-subject covariance structure can be avoided. The authors assume a simple carry-over model with a general within-subject covariance structure. A simultaneous hypothesis concerning treatment and carry-over effect is first performed. This hypothesis is usually rejected at conventional significance levels and this leads to a test for examining the importance of carry-over difference. If carry-over effect is shown to be different from zero, then only first period data are used for testing treatment effect, while in the case where residual effect is statistically unimportant then all four cell means are used for drawing inference for treatment effect. This work is extended in the case of the cyclic design (ABC,BCA,CAB) where carry-over effect in the third period represents residual effect from the second and the first period. A similar multi-stage procedure to the 2x2 case for testing treatment effect is proposed. The authors conclude that for the 2x2 design the univariate approach gives identical results to the multivari-

ate one. For multi-period designs different hypothesis can be tested under the two approaches, though the multivariate procedures has the advantage that less restrictive assumptions are imposed. A hybrid procedure is finally recommended which uses the advantages of both approaches.

In multi-period multi-sequence cross-over trials a single outcome variable is usually of interest. There are occasions however, where two or more outcome variables may be observed within a treatment period. This is simply a cross-over design with multivariate observations and can be analyzed using standard multivariate techniques, see Mardia et al [60] or Kraznowski [46]. Grender and Johnson (see [24]) discuss an example of a cross-over trial with a bivariate response, where the effect of caffeine on stress reactions was studied by measuring systolic and diastolic blood pressures before performing a task and after administering caffeine or placebo. An adequate wash-out period was allowed in this study.

Let y_{ijk} be the response vector of the k^{th} subject within the j^{th} period who has been randomized in the i^{th} sequence. The model can be expressed in matrix notation as follows:

$$E\left(y_{ijk}\right) = \mu + \pi_j + \tau + \lambda \tag{2.21}$$

where $\mu, \pi_{\mathrm{j}}, \tau, \lambda$ are vectors corresponding to the overall mean, period, treatment and carry-over effects. Note that Grender and Johnson do not include a sequence effect, since even in the multivariate 2x2 case that term is confounded with the carry-over effect. The above model can be presented in a concise form as follows:

$$E\left(Y\right) = A\phi \tag{2.22}$$

where each row of Y corresponds to responses of each individual. All multivariate cross-over hypothesis can be written in the form:

$$C\phi M = 0 \tag{2.23}$$

and an appropriately constructed F-test can be used to test the hypothesis above. Surprisingly enough Grender and Johnson propose a multi-variate analogue of the two stage procedure to test the hypothesis of treatment effect. Obviously the defficiencies of that scheme are well known, when one outcome variable is measured in each period, but it is my view the same defficiences will be evident in the

multivariate case as well. A common covariance matrix Σ is assumed for the observations taken on a subject across responses.

Grender and Johnson extend the above work to accommodate analysis of two or more responses taken repeatedly across time. For example, in the previous trial, diastolic and systolic blood pressure can be measured more than once within a period. In such circumstances, the interaction of time with period, treatment and carry-over should be tested and if not important then one can average responses over time points and use the analysis outlined above. More specifically, Grender and Johnson propose a three stage procedure, where the time by carry-over interaction is tested first, followed by a test of no carry-over differences. Based on the outcome of the test for carry-over, either data from both periods are used or only the first period data considered for analysis purposes. The Wilks likelihood ratio criterion, which transforms to a F-statistic for the 2x2 cross-over case, is used for the hypothesis testing of various effects.

In a subsequent paper (see [25]), Grender and Johnson fit polynomial models for a 2x2 cross-over design where several responses are measured within a period, repeatedly over time. With such data, a multi-variate test of equality of means at time points within sequence by period cells, is first performed. This hypothesis is usually rejected at conventional levels of significance and the next step is try to claim parallelism of mean profiles across groups defined by the sequence by period cells. If that hypothesis accepted, then averaging response(s) across time is the way forward. However, if the parallelism hypothesis is rejected, then the aim might be to discover how mean profiles across sequence by period groups differ. To that purpose, a polynomial model can be fitted to subject specific data. The estimated parameters of the polynomial model are subsequently analyzed using appropriate techniques for cross-over plans. Grender and Johnson illustrate the technique by fitting second order polynomials in a study which investigates the effect of eating onions on triglyceride levels of patients with heart disease.

2.8.2 Survival Data

The analysis of survival data in the cross-over literature is one of the areas that has been under-developed. France et al (see [20]) are one of the few authors who

describe different approaches for the analysis of survival data in the 2x2 case. The trial used to illustrate the methods concerns treatment of angina pectoris. A well established therapy is compared to the combination of that therapy with a newly proposed treatment. There is a 4-week run-in period followed by two active four-week treatment periods. No wash-out interval is allowed. France et al initially analyze the data using standard methods of analysis, like analysis of variance or Wilcoxon rank sum test. None of the above methods, take into account the correct underlying data-distribution or the censoring mechanism. Because of that, treatment effect estimates derived from these methods are biased. France's et al survival method follows closely Cox's proportional hazards regression model. Each patient has a separate baseline hazard function appropriately shifted to allow for treatment and period effects. France et al do not include a carry-over term in their analysis. For the i^{th} patient the hazard function is:

$$h_i(t, \text{period}, \text{treatment}) = h_{0i}(t)exp(\beta_1 * \text{treatment} + \beta_2 * \text{period}) \qquad (2.24)$$

The treatment and period effect can be estimated by maximization of the partial likelihood and depends only on the number of treatment preferences in the two sequence groups. Treatment A is preferred to treatment B, if the survival time on A is longer than that on treatment B. If $n_{1A}, n_{1B}, n_{2A}, n_{2B}$ are the number of preferences of A and B in the two groups, then

$$\hat{\beta}_1 = ln\sqrt{\frac{n_{1B}n_{2B}}{n_{1A}n_{1B}}} \qquad (2.25)$$

A similar, though more elaborate, expression holds for the variance of $\hat{\beta}_1$. France et al, are in a position to extent their methodology to a three-period cross-over trial. They do not manage though to extend their work to include cross-over designs with unlimited number of sequences/periods and for comparison of more than two treatments.

Feingold and Gillespie (see [14]) propose an alternative method for the analysis of cross-over survival data, easily applied to many different experimental designs under various censoring mechanisms. In Feingold and Gillespie's method, each observation is replaced by a score and then standard statistical techniques applied (e.g. ANOVA) to the derived scores. Feingold and Gillespie use the Gehan

score, defined differently for censored and uncensored observations. An alternative way of analysis is to use medians or other quantiles obtained from the survivor curve from each sequence/period combination and then apply the CROS weights to these summary statistics. Feingold and Gillespie seem to favor the average-quantile statistic, which is simply the average distance of each survivor curve from the origin calculated over a quantile range where all survivor curves are defined. The asymptotic variance of that statistic is difficult to calculate and bootstraping could be used to that purpose. Both approaches are illustrated with an example of the protective value of two types of helicopter passenger immersion suits. Simulation methodology is used to compare score transformation method (ST) with France's et al (FLK) procedure. Overall the ST method seems to perform better in terms of power and bias for treatment effect estimation regardless of the censoring rate. The analysis of time failure cross-over data can be carried out using standard survival analysis software, where treatment, period and carry-over effect are time dependent covariates.

2.8.3 Classical and modern non-parametric approaches

For many years the use of classical non-parametric procedures, like Wilcoxon rank sum test, have dominated the analysis of cross-over data where the distributional assumptions (e.g. normality) have been violated. This approach is illustrated in Koch (see [45]) for the 2x2 design. Koch assumes the simple carry-over model with a random subject effect. For estimating treatment effect, the within subject differences are calculated and the Wilcoxon rank sum test is applied to these differences. Similarly for testing the hypothesis of no residual effects, the within subject sums are first evaluated and then the Wilcoxon test is applied to these sums.

McHugh and Gomez-Marin (see [67]) examine and compare a randomization model for analyzing the 2x2 cross-over design with the simple carry-over model. An additivity assumption is then introduced in the randomization model and a new comparison with the simple carry-over model is performed. The randomization model assumes that the test and reference products can be tried only on a finite population of size N. Conceptually each of the N experimental subjects can

be allocated to any one of the four possible sequence by period combinations, generating a hypothesized $4N$ responses, which can be used to describe the observed responses. It turns out that the treatment estimator based on the randomization argument alone, has a distinct different variance from the treatment estimator under the simple carry-over model. The additivity assumption, introduced next, simply assumes that the $4N$ conceptually responses are composed of a subject, a treatment and a carry-over effect. The results of the randomization model with the additivity condition are comparable to the results of the simple-carry-over model, as far as the precision of the treatment effect estimate is concerned.

Tsai and Patel (see [88]) were one of the first to apply modern non-parametric approaches to the analysis of the 2x2 trial with baselines. Tsai and Patel implement these methods to a 2x2 design that includes a placebo run-in period and a wash-out interval of adequate length between the two active treatment periods. Baseline measurements are taken both during the run-in and wash-out periods. Tsai and Patel, consider a similar approach to that of Jones and Kenward for the management of carry-over effects. A test for the significance of the residual effect from the first treatment period to the wash-out interval is first performed, by taking the differences between the two baseline measurements and then applying a Wilcoxon rank sum test to the derived data from the two sequence groups. Where Tsai and Patel's work differs from conventional approaches, is the way they test for residual effect from the first to the second treatment period, and the way they test for treatment effect. Before testing for treatment and carry-over effects, Tsai and Patel remove the effect of baselines. Data from first and second treatment period are modeled separately. A linear regression is performed, with treatment period data as response and corresponding baseline measurement as covariate. These models are not fitted by minimizing the sum of the squares of the difference between the response and its expected value, but rather a slightly complicated function of that difference is optimized. Robust linear fit minimizes for each period j, where $j = 1, 2$, the following function:

$$\sum_{i,k} \phi\left(\left(y_{ijk} - \alpha_j - \beta_j x_{ijk}\right)/\sigma_j\right) \tag{2.26}$$

Note that a common regression coefficient is assumed for both sequences within a period. The function $\phi(x)$ is Huber's function (see [34]), and parameter estimates

are derived by solving a system of nonlinear equations simultaneously. The above equation also implies a different variance parameter for the two period groups. An alternative way to identify a relationship between two continuous variables, is by fitting a locally weighted robust regression curve. Cleveland (see [8]) was the first to introduce the idea, which allows us to use neighborhood points of a given point (x, y) to obtain a fitted value for y. With these points, a weighted least squares fit is performed, where the weighted function is symmetric about x and decreases to zero as the distance from x increases. As before, this method is applied separately to the data from the two periods.

Using either of the above approaches, a pair of residuals (r_{i1k}, r_{i2k}) can be calculated for each subject, and the hypothesis of no carry-over effect from first to second treatment period or of no treatment difference is based on these residual pairs. A Wilcoxon rank sum test is applied to the sets $(r_{111}+r_{121}, \ldots, r_{11n_1}+r_{12n_1})$ and $(r_{211}+r_{221}, \ldots, r_{21n_2}+r_{22n_2})$ for carry-over testing, where n_1 and n_2 are number of subjects randomized to the two sequence groups. For the comparison of treatments a Wilcoxon rank sum test is performed on the differences $r_{i1k} - r_{i2k}$.

2.8.4 Poisson Data

There is an extensive literature covering generalized linear model approaches for modeling purposes in repeated measures settings (see [56]). One of the few papers paying special attention to the analysis of count cross-over data is the one by Layard and Arvesen (see [54]). The authors suggest that cross-over experiment should be avoided if it is thought that carry-over effects could occur. So, a Poisson distribution is assumed for the count data, while a log-link relates the mean of that distribution to the linear predictor. The linear predictor contains terms for subject, period and treatment only. Layard-Aversen's analysis conditions upon subject totals. In this way testing for drug by period interactions is not feasible, though for tackling this problem they recommend an alternative procedure based on a t-test for appropriately transformed patient data. They illustrate their approach using two examples, where a 2x2 design was used to run the trials.

The authors extend their methods to multi-period designs using the 3x3 Latin

square for illustrative purposes. The approach based on conditioning on the subject total suffers from the fact that pairwise comparisons among treatments are not easily performed. An alternative route where data are first appropriately transformed and then a weighted linear regression is performed on the transformed values, with weights determined beforehand, is recommended.

2.9 Variance Components Estimation

Laird et al (see [50]) propose an interesting method of estimating the variance components, when compound symmetry structure (i.e. random subject effects) is assumed for the responses on a subject, in two period cross-over studies. More specifically, if d_i, s_i denotes the difference and the sum of the two responses on the i^{th} subject, then the following two models are fitted,

$$d = X_d\beta + e_d \tag{2.27}$$

$$s = X_s\beta + e_s \tag{2.28}$$

where X_d and X_s denote the design matrices for the sum and difference vector. From these models two mean square errors, MSE_d and MSE_s, are derived and the covariance/correlation parameters are estimated as follows:

$$\hat{\sigma}^2 = (MSE_s + MSE_d)/4 \tag{2.29}$$

$$\hat{\rho} = (MSE_s - MSE_d)/(MSE_s + MSE_d) \tag{2.30}$$

Laird et al combine the estimates of β derived from equations (2.27) and (2.28) to derive the GLS estimate. Obviously this method generalizes in a straightforward way, when baseline measurements are included as covariates.

Matthews (see [63]) considers the estimation of the dispersion parameters in the general case of a p-period cross-over trial with a continuous outcome, where n subjects are recruited. The model assumed, includes subject, period and treatment effects (all fixed), while carry-over term is not considered. The linear model can be summarized in the following equation

$$y = (I_n \otimes 1_p)s + (1_n \otimes I_p)\pi + T\tau + \epsilon = Z\alpha + \epsilon \tag{2.31}$$

where y is an np-dimensional vector, s, π, τ is the subject, period and treatment effect respectively. The variance matrix of the error-vector ϵ has a block diagonal form

$$W = \sigma^2 I_n \otimes V \tag{2.32}$$

where the matrix V_{pxp} describes the intra-subject correlation structure. This structure takes the form of a stationary first-order autoregressive process, with its $(i,j)^{th}$ element equal to $(1-\rho^2)^{-1}\rho^{|i-j|}$. Matthews removes the nuisance parameters, subject and period terms, by pre-multiplying both sides of the above equation with an appropriate matrix. The model for the transformed response looks as follows:

$$z = \Delta\tau + \epsilon^* \tag{2.33}$$

This model contains only the parameters we are interested in, τ, ρ and σ^2. The author then applies ordinary maximum likelihood and derives an analytic expression for the correlation coefficient ρ. The above approach, called restricted maximum likelihood, is equivalent to integrating out the nuisance parameters from the full likelihood function. Matthews compares the above method with a conditional profile likelihood approach, where a likelihood function containing only the parameter of interest, ρ, can be written down explicitly. Simulation studies are used to compare the two inference methods plus the standard maximum likelihood approach. The designs used are a four-sequence three-period one (ABB, AAB, duals) and a four-sequence four-period one (ABBA, AABB, duals), where 12 subjects are allocated in each sequence. Matthews concludes that both conditional and restricted likelihood approaches perform better than the standard maximum likelihood in terms of bias, though the restricted likelihood approach is to be preferred because it can easily be generalized to the case where intra-subject covariance structure is described by more than one parameter. Standard weighted least squares can be used to estimate τ, the treatment effect, with ρ replaced by its estimate. Uncertainty concerning the estimation of ρ can safely be ignored in our inferences for τ, since ρ and τ are orthogonaly estimated.
Guilbaud (see [31]) estimates variance components in the 2x2 case, assuming that variances under the two treatment regimes are different. Interest centers

on drawing inference for the ratio $\theta = \sigma_A^2/\sigma_B^2$, which measures the relative variability within subjects under the two treatments. Guilbaud derives initially the exact distribution of the following quantity $\gamma = (\sigma_A^2 - \sigma_B^2)/(\sigma_A^2 + \sigma_B^2)$, from which inferences about θ can be made. As before, the key statistic is based on the within-subject sum and difference pair (s_{ik}, d_{ik}), where k indexes subject and i sequence group. The author proves that $(\gamma^* - \gamma)/s^*$ follows a t-distribution on $n-3$ degrees of freedom, where n is the total number of participants recruited in the study. The value of γ^* equals the common slope of two parallel lines fitted to the two sequence groups by ordinary least squares, with the d_{ik} treated as fixed predictor, while the s_{ik} treated as the response. The s^* is simply the standard error of that slope.

2.10 Choosing the right design

2.10.1 Theoretical results on repeated measures designs

Kunert (see [47]) deviates from conventional approaches to identify optimal plans for repeated measurements designs, a special case of which are cross-over plans. Special restrictions are usually imposed on a plan to be optimal under a predefined model. For example, number of treatments should appear equally often in each sequence and period. This work is presented in Ch4.

Kunert proves an orthogonality condition which ensures that the information matrices for the estimation of the same effects in two models are equal. Note that for the two models it is assumed that one of them is nested within the other. The author discriminates the set of parameters the experimenter is primarily interested in (η), from the parameters that are of secondary importance (ξ). The following model is assumed:

$$Y = A\eta + B\xi + \epsilon \tag{2.34}$$

where the error vector has uncorrelated components with common variance. The information matrix for the parameters of interest η is:

$$C = A^T \left(I - B \left(B^T B \right)^{-1} B^T \right) A \tag{2.35}$$

The responses on the nested model are generated according to the following model:

$$Y = A\eta + B_1\phi + \epsilon \tag{2.36}$$

Obviously $B = [B_1|B_2]$, with an information matrix for the parameters of interest η given by an expression similar to the one described in equation (2.35). The information matrices in the two models will be equal if the following orthogonality condition is satisfied:

$$A^T \left(I - B_1 \left(B_1^T B_1 \right)^{-1} B_1^T \right) B_2 = 0 \tag{2.37}$$

So, if one can find an optimum design for the model described by equation (2.36), which at the same time satisfies the orthogonality condition described by equation (2.37), then this design is optimum for the more elaborated model described by equation (2.34).

In Kunert's work simple carry-over is assumed throughout. The author considers cross-over designs where residual effect is allowed even in the first period. From a practical standpoint, this assumption is not as unreasonable as it sounds, since in most clinical trials participants are already on a standard therapy and if that therapy is compared to a newly proposed treatment, then carry-over in the first period may exist. Of course, Kunert also considers cross-over plans with no residual terms in the first period. Optimum results claimed for the family of generalized latin square (GLS) designs, where both number of subjects (n) randomized and number of periods (p) used are a multiple of number of treatments (t) compared, and treatments appear equally often in each sequence and on each period. If these conditions are satisfied, then any design made of sequences where a treatment is followed equally often by all other treatments (including itself) is optimum for the estimation of treatment effects in that family. In that paper, Kunert replaces the strong assumption that number of subjects recruited must be a multiple of number of treatments, with a weaker one that relates the number of times a treatment i is followed by treatment j with the number of times treatments i and j appear in period k. In a similar fashion the assumption that number of periods must be proportional to the number of treatments can also be

replaced with a weaker one. Similar results for efficient estimation of the carry-over effects are also presented.

In a subsequent paper, Kunert (see [48]), deals with the situation where number of periods equals number of treatments ($p = t$). Balance uniform designs (i.e. $n \propto t$, treatments appear equally often in each period/sequence, each treatment is followed equally often by other treatments, but never by itself) are optimum for estimating treatment differences, when the number of subjects equals or is twice as high the number of treatments. Whenever $n = 2t$ this result is true only when more than six treatments are compared. In case where the subjects recruited is a multiple of the number of treatments greater than two, then the efficiency of a balance uniform design is greater than

$$\frac{(t-1)^2 - 2(t-1)t^{-1}}{(t-1)^2 - 2(t-1)t^{-1} + t^{-2}} \tag{2.38}$$

The equation above implies, that as the number of treatment grows balance uniform designs are almost optimum. In that work, Kunert extends these results by providing conditions for efficient estimation of residual effects when $p = t$. In these conditions, it is assumed that $n = t(t-1)$ and the experimenter is in a position to find a uniform balanced design where each pair of treatments appears equally often in the last and second to last period. Then if the last period is replaced with the second to last one, the resulting plan will be optimum for estimation of residual terms.

Kunert (see [49]) extends the results above, in the situation where repeated measurements taken on the same subject are related according to an AR(1) process. As before, number of treatments equal number of periods. In that paper terms for the mean, subject, period and treatment effect are only fitted. Carry-over terms are not allowed. The sum-to-zero parameterization is used for the treatment effect. Kunert's model is exactly the same as the one described in equation (2.31) with a covariance matrix for the error vector given in equation (2.32). Now, if we consider the matrix A_{pxp} with the property $AVA^T = I$, then by pre-multiplying both sides of equation (2.31) with $I_n \otimes A$ the resulting error vactor has uncorrelated components. Let $B = [1_n \otimes A | I_n \otimes A 1_p]$, then the information matrix for

the treatment effect estimates matrix is:

$$C = T^T \left(I_n \otimes A^T\right) \left(I - B \left(B^T B\right)^{-1} B^T\right) \left(I_n \otimes A\right) T \tag{2.39}$$

The family of designs where each treatment appears equally often in each sequence and each period, followed equally often by other treatments (including itself), is considered. In addition, each pair of treatments should appear equally often in the first and last period. Kunert calls this set of plans "Williams design with balance end-pairs". The main conclusion of Kunert's work, is that a Williams design with balance end pairs is optimum for estimation of treatment effects, irrespective of the value of the AR(1) coefficient ρ, over the family of designs where treatments appear equally often in each sequence. Furthermore, Williams designs with balanced end-pairs are optimal for estimating treatment effect over the whole design family under study, when $p = t = 3$. When $p = t > 3$, then Williams design with balanced end-pairs is optimum for treatment effects over the whole design family, only when the AR(1) coefficient ρ is greater than

$$\frac{t - 2 - \sqrt{t^2 - 8}}{2(t - 3)} \tag{2.40}$$

A general note on Kunert's optimality criterion is in order. The ϕ_α optimality criterion has been used throughout. Let $\lambda_i, (i = 1, \ldots, k = rank(C))$ be the non-zero eigenvalues of the information matrix C for the parameters we are interested in. Then for every α, where $0 \leq \alpha \leq \infty$, the following criterion can be defined:

$$\frac{1}{k} \left(\sum_{i=1}^{k} \lambda_i^{-a} \right)^{1/a} \tag{2.41}$$

Kunert's results are valid for all values of α. When $\alpha = 0$ the D-criterion, widely used in subsequent chapters, is recovered. The ϕ_1 criterion corresponds to the well-known A-criterion. Universal optimality, described in Ch4, is a more general concept than ϕ_α optimality, though most of Kunert's results are valid under the universal optimality criterion as well.

2.10.2 Practical results on repeated measures designs

In an impressive review paper, Matthews (see [65]) questions the conventional approaches to identifying optimum plans. His main criticism concentrates on how

appropriate is the simple carry-over model, a model widely used for derivation of optimum designs, and proposes alternative solutions. Before proceeding to that debate, Matthews is critical about considering carry-over terms in the first period. Magda's work (see [59]) is based heavily on this assumption, where a hypothesized pre-period interval with the same treatments administered as in the last period, is used. So, carry-over in the first period is determined by treatments administered in the final period. The data collected during the pre-period interval are not used in the analysis stage, and this is the point where this scenario may sound unreasonable to the practical user of the cross-over trial. The modeling of the carry-over term has been criticized by Fleiss (see [18]). According to Fleiss carry-over plan, a treatment carries-over to all other treatments, except itself. This type of carry-over with some extensions is studied in subsequent chapters. For two treatment comparison, Matthews expresses the Fleiss model as follows:

$$y_{ijk} = \mu + s_{ik} + \pi_j + \tau d_{i,j} + \frac{1}{2}\lambda d_{i,j-1}\left(1 - d_{i,j}d_{i,j-1}\right) + \epsilon_{ijk} \qquad (2.42)$$

where $d_{i,j}$ is the treatment (A or B) administered to sequence i in period j. Assuming independent errors and all other term in the above equation fixed, Matthews derives optimum dual sequence plans for the Fleiss type of carry-over, where unequal number of subjects may be allocated to each sequence pair. The author presents best designs in the three or four-period families and concludes that these designs are highly efficient under the simple carry-over model. Worth noting that in Matthews work, sequences like AAA/BBB or AAAA/BBBB are candidates for inclusion in the proposed plans. The author also points out that since the type of carry-over is not known in the planning stage, designs which are robust to model mis-specification in terms of efficiency and/or bias are worthwhile to be derived. Results to that direction are provided in Ch4.

In a subsequent paper, Matthews (see [64]), studies the problem of how efficient Ordinary Least Squares (OLS) treatment estimate is, when the responses on the same subject are stochastically dependent and this dependence is captured by an unknown parameter ρ. If ρ was known a-priori, then Generalized Least Squares (GLS) would have been fully efficient. The fact that ρ is estimated forces the analyst to use a practical alternative, the empirical GLS (EGLS), which will not always be more efficient than OLS. Matthews points out that the choice

of estimate heavily depends on the design selected to run the trial. Designs, where the OLS estimate is highly efficient even when repeated measures on a subject are correlated, may be preferred for running the trial during the planning stage. Matthews investigation, compares OLS with GLS when ρ is known. If the efficiency of OLS is 90% or more over all plausible values of ρ, then the extra complexity of the GLS analysis may not be justified.

More specifically if ρ is known, then the estimate of the variance for the fixed effects using GLS would be:

$$var(\alpha_{GLS}) = \sigma^2(Z^T W^{-1} Z)^{-1} \tag{2.43}$$

The OLS estimate is simply $(Z^T Z)^{-1} Z^T y$ with variance:

$$var_{true}(\alpha_{OLS}) = \sigma^2(Z^T Z)^{-1}(Z^T W Z)(Z^T Z)^{-1} \tag{2.44}$$

The estimated variance from a typical OLS analysis would be:

$$var_{analysis}(\alpha_{OLS}) = \hat{\sigma}_0^2(Z^T Z)^{-1} \tag{2.45}$$

where $\hat{\sigma}_0^2$ is the OLS estimate of σ^2. Matthews compares the expected values of the variance estimates provided in the last two equations under the GLS model. Although he includes a carry-over effect in his model, attention is focused on how misleadingly the variance of the treatment effect is estimated using OLS, when GLS should be used instead. For the intra-subject covariance structure, either a first order autoregressive or a first order moving average model is assumed. The author concludes that designs where the OLS estimate can be used without much loss of efficiency are (ABB,AAB,duals) and (ABB,ABA,duals) from the four-sequence three-period family, while for the two-sequence four-period family good choices are (ABBA, dual) and (ABBB, dual).

Matthews then questions the sensitivity of not equal allocation of available patients to each sequence. The author illustrates his point using a dose-escalating design for the comparison of three doses (1,2,3) and placebo (P), where the proportion of subjects randomized to each sequence are not necessarily equal. The four-period design used, consists of the following four sequences: P123,1P23,12P3 and 123P. It turns out that the efficiency of equireplicate designs is over 90% and this is true over a wide range of optimality criteria (A-, D-, E-criterion).

A further point raised by Matthews concerns the modeling of period effect. This term is typically incorporated in any modeling exercise. Since every patient has his own trial history, the statistical interpretation of the period effect is unclear. It is my view that data collected on a subject is part of a stochastic process and effects believed to influence patient's response at a given time point should be included as terms in his mean response at that time. An alternative way of incorporating period effects in an analysis, is to describe stochastic dependence on observations taken within subjects or across groups. In that respect, Matthews proposes a solution where period is considered as a random effect. Sensitivity of optimality results when period term is excluded from the model is a potential research direction according to the author. Finally, Matthews raises the point that treatment by subject interaction may be worth investigating from both sponsor's and GP's perspective.

2.10.3 Results on special design families

In a different mode, Pigeon and Raghavarao (see [69]), propose designs for comparing u test treatments $(0, \ldots, u-1)$ with a control x. The authors are interested in efficient estimation of the contrasts of treatment and carry-over effects of the u test treatments versus the control treatment. Designs where the variance matrix of the contrasts of interest is completely symmetric tend to be optimum. Control balance residual effect designs found to have this property and being equally efficient as incomplete block solutions, for estimating contrasts of interest. A control balance residual effect design possesses the following properties: each treatment appears at most once in each subject, control and test treatments occur the same number of times in each period (t_0 times for the control and t_1 times for the test), control treatment occurs with each test treatment in λ_0 subjects and each test treatment occurs with every other test treatment in λ_1 subjects, the prementioned property also holds if the last period is deleted, the ordered pair of treatments (x,i) occur in successive periods in v_0 subjects and the ordered pair (i,j) occur in successive periods in v_1 subjects. Finally, for every treatment pair (θ,ϕ) the number of subjects where θ occurs with ϕ in the last period equals the number of subjects where ϕ occurs with θ in the last period. The author provides

rules for the construction of such a design.

An interesting paper, close to that of Matthews, is Lasserre's work (see [53]) on determining optimum plans for the comparison of two treatments when two, three and four period design-families are considered. Three models studied and subject is considered as a random effect throughout. In the first model, overall mean, period and treatment terms are included. The second model contains all terms included in the first plus carry-over of the simple type. The third one is an extension of the second, where treatment by period interaction is also fitted. A first result in Lasserre's work is that the estimate of the period effect is independent from the between subject variability. Variance of treatment effect or other parameters of interest, are presented in terms of the ratio σ_s^2/σ^2 where σ_s^2 is the subject error variance and σ^2 the error variance. For the two-period designs all possible sequences are considered (AA,AB,BA,BB) and if n_{11}, n_{12}, n_{21} and n_{22} subjects allocated in each one of them, the following restrictions must be satisfied $n_{11} = n_{22}$ and $n_{12} = n_{21}$. The same principle applies to three and four-period families. The variance of the treatment effect estimate is minimized under model 1 when equal number of subjects are allocated to the sequences (AB,AB). For the other two models, equal number of patients should be allocated in all four possible sequences. For the three-period two-sequence family anyone of the three possible designs can be recommended for use under model 1. Under the same model, in the three-period four-sequence family the plan (AAB, ABA, duals) is optimum. Surprisingly enough for model 2 the same design estimates efficiently both treatment and carry-over effects: ABB/dual. For model 3, two six-sequence designs are recommended. When four-periods are used, then under model 1 the number of designs the experimenter can choose from to run his study is much higher, compared to the number of plans for the other two models.

Two-period cross-over designs where more than two treatments are compared are studied by Carriere and Reinsel (see [4]). Number of randomized subjects (N) should be a multiple of t or t^2, t being the number of treatments under consideration. The authors derive first the information matrix jointly for the treatment and carry-over effects and then the information matrix for the treatment effect alone, adjusted for all the other terms in the model, carry-over inclusive. Carriere and

Reinsel prove that a two-period design where treatments appear equally often in each period and the number of subjects receiving the treatment pair (i, j) equals the number of subjects receiving the pair (j, i), is optimal for estimating treatment effects among all two-period repeated measurement designs, when $N \propto t^2$. The covariance matrix for any treatment effect contrast is also evaluated.
Up till now we have not touched upon Bayesian methodologies for deriving optimum plans for cross-over trials. The application of Bayesian ideas becomes more relevant when non-linear terms included in the model, see Ch5 for further details. Both Atkinson and Donev (see [1]) or Fedorov and Hackl (see [13]) provide a thorough account of Bayesian experimental design theory. An interesting paper on Bayesian design of experiments for linear models in the presence of variance components is that by Lohr (see [58]). The paper is not related to cross-over experiments particularly, but the ideas presented there can be applied easily to that area as well. Any cross-over model can take the following form:

$$y_{ijk} = \mu_{ij} + s_{ik} + \epsilon_{ijk} \tag{2.46}$$

where $s_{ik} \sim N(0, \sigma_b^2)$ and $\epsilon_{ijk} \sim N(0, \sigma^2)$. The question of interest is identifying best plans for variance component estimation. A reasonable guess is needed for the variance components, in fact for the ratio $\gamma = \sigma_b^2/\sigma^2$, which will be translated to a proper prior distribution for that parameter. Two Bayesian design criteria considered in Lohr's work: the expected value of the log of the determinant of the information matrix is maximized (extension of D-optimality), the average variance of a linear combination of the parameters is minimized (extension of A-optimality). After deriving the Fisher information matrix for the variance components, the number of measurements per sequence is estimated and a further condition needs to be satisfied for the design to be D-optimum. All these conditions depend heavily on the prior distribution for γ. Results between the two criteria differ, when the prior belief is that γ is small.

2.11 Concluding remarks

The majority of the methodologies concerning the analysis of cross-over data has been developed with the 2x2 design in mind. Extensions of these analysis methods

to higher order designs are under-developed. Obviously, there is an extensive literature on the analysis of repeated measurement experiments, a special case of which is the cross-over experiment. Lindsey (see [56]) provides over 50 pages of references for analyzing repeated measurement data. Worth investigating, if extended 2x2 analysis strategies to more general settings lead to techniques already known for analyzing repeated measures.

Concerning the identification of optimum plans, attention has been focused on the simple carry-over model. Results for other carry-over types are provided in subsequent chapters. These results apply to cross-over plans with limited number of sequences and periods. Generalization of these results to higher order designs require further research effort.

Chapter 3

The 2x2 Cross-Over Trial

3.1 Cross-over and Parallel Group Trials

A cross-over trial is an attempt to make a fair comparison between two or more treatments on a group of patients. Patients are divided into two or more groups, and each group receives a sequence of treatments. The time period a trial lasts is divided into sub-periods. At the beginning of each sub-period one and only one treatment is administered to each patient and the effect of the treatment is assessed at the end of it.

The main advantage of conducting a cross-over trial in medical research, is the ability of making treatment estimates based on within-subject measurements (see Senn [77]). This simply means that the variance of the proposed treatment-estimator is lower compared to the one of a parallel group trial. Moreover patients through their measurements, provide their own judgement about the performance of the therapies and it's the combination of these judgements that forms the final picture concerning effectiveness of different treatment regimes. On the contrary if we use a parallel trial to compare therapies, then each group of patients receives only a specific treatment and then comparisons are made between groups to assess treatment-effect. Obviously a significant difference between two treatments in a parallel study, might be caused because of differences in groups and not of any real treatment effect. However randomization arguments exclude the possibility of a significant group effect.

A main disadvantage with the cross-over trial is the carry-over effect. This simply

means that the effect of a treatment in a given period is present at the beginning of subsequent periods. As a result other important effects of interest, like the treatment by period interaction in the 2x2 case, are intrinsically aliased with the carry-over effect. A solution to the problem is to allow for adequate wash-out period between any two active periods, so that the effect of the treatment in the present period is eliminated at the start of subsequent periods. A drawback of this solution is that the time-period a trial lasts is extended considerably, increasing at the same time substantially the possibility of drop-outs.

3.2 The model

A response obtained on a patient participating in a cross-over study at a specific treatment period is affected by a number of factors, some of which are listed below (see Jones and Kenward [39]):

- Physical condition of the patient at the time the measurement has been taken (subject effect).

- Effect of the period in which the measurement was taken. This corresponds to time trend effects, probably affecting the trial as a whole. For example measurements taken in Winter might be substantially lower than measurements taken in Summer, no matter which treatment is administered to the patient. But the statistician should always keep in mind that patients are not recruited simultaneously. For example two patients who have been assigned to the sequence AB might visit the clinic for the first time in different dates. This raises the important question of how the period (time) effect should be defined and modeled, though Matthews (see [65]) proposes several approaches to this query.

- Effect of the treatment given to the patient at that period. This simply counts the improvement (if any) in patient's health by the specific treatment, when this improvement is compared to the normal condition of the patient. Usually the model is over-parameterized if a different term is allowed for each treatment effect. Contrasts of treatments are usually in-

cluded in the model. A standard parameterisation is the one in which each treatment is compared to a standard therapy.

- Effect of treatments administered in previous periods. This is simply the well-known carry-over (residual) effect of previous treatments administered to patients during the course of the study. The presence of carry-over effect, not only biases the estimated treatment effect, but also creates problems on the choice of best design at the early stages of the trial-design phase. As a result, adjusting for those residual effects is of questionable value for assessing the real effect of the current treatment. But the question raised is if carry-over effect is present at all in a well-planned clinical trial.

- Random fluctuation counts for errors which cannot be controlled or explained by the trialist. The effects of explanatory variables that may influence the response but have not been measured during the study period, are also included in that term.

The above additive effects if written in an equation form, give the linear model for cross-over trials described below:

$$y_{ijk} = \mu + s_{ik} + \pi_j + \tau_{d(i,j)} + \lambda_{d(i,j-1)} + \epsilon_{ijk} \tag{3.1}$$

where

- μ : a general overall mean.

- s_{ik} : The effect of subject k in the i sequence group.

- π_j : The period effect.

- $\tau_{d(i,j)}$: Effect of treatment given to sequence i, at period j.

- $\lambda_{d(i,j-1)}$: Residual effect of treatment given to sequence i, at period $j-1$.

Note that carry-over effect lasts only for one period and depends on the preceding treatment and not on the current one. For that reason the model described in (3.1) is called 'simple carry-over' model. It is possible to include higher order carry-over terms in the previous model, but this may create identification problems during the estimation process. Furthermore, it is highly unlikely in practice

that carry-over terms of higher order will be present at all.

Two important statistical points are in order. To begin with, in the classical linear model set-up it is always assumed that $\epsilon_{ijk} \sim N(0, \sigma_W^2)$, where σ_W^2 stands for the within-patient error variance. A question of interest is if the subject effect should be considered as fixed or random (see Chi [6]). From the statistician's perspective, inferences should be drawn for the population with the medical condition as a whole, rather than for the trial participants only. This calls for considering the subject effect as random variable. In addition with random subject effect, the number of parameters needed to describe that effect does not increase with the number of subjects. Finally, variability for which one has no explanatory variables to explain, or for which one wishes to allow without trying to explain can be described by a random subject effect. For all the above reasons, the assumption $s_{ik} \sim N(0, \sigma_B^2)$ seems a realistic one.

Secondly, in the model described in (3.1), no interaction effects were included. In a cross-over trial it is not possible to test all multi-way interactions, although some of them might be of interest in specific settings. For example the treatment effect might be a function of time. In the classical linear model setting this implies a statistically significant treatment by period interaction. Another similar example is the patient by treatment interaction, which simply indicates that the treatment is highly beneficial for one group of patients, but less so for another. The availability of that information is of great interest to pharmaceutical companies, as it would show the sub-population(s) for which the development of a compound is worthwhile. This information is available only at the later phases of drug development.

In the examples that follow the statistical significance of some of the interaction terms will be examined. In most cases it will be concluded that the inclusion of such terms in the model is hardly necessary.

3.3 The 2x2 case

In the simplest of the cross-over designs, where two treatments are tested in two periods, half of the recruited patients are randomly allocated to one of the

two sequence treatment groups; namely AB and BA. The four group by period observed means are sufficient for drawing inference about the treatment effect. The treatment estimator is simply a weighted average of these means; namely

$$\hat{\tau} = w_1\bar{y}_{11} + w_2\bar{y}_{12} + w_3\bar{y}_{21} + w_4\bar{y}_{22} \tag{3.2}$$

where $\sum_{i=1}^{4} w_i = 0$. The first two moments of the four sequence by period means vector, assuming random subject effect, can be expressed for the simple carry-over model as follows:

$$E\begin{pmatrix} \bar{y}_{11} \\ \bar{y}_{12} \\ \bar{y}_{21} \\ \bar{y}_{22} \end{pmatrix} = \begin{pmatrix} 1 & -1 & -1 & 0 \\ 1 & 1 & 1 & -1 \\ 1 & -1 & 1 & 0 \\ 1 & 1 & -1 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \pi \\ \tau \\ \lambda \end{pmatrix} \tag{3.3}$$

and,

$$V\begin{pmatrix} \bar{y}_{11} \\ \bar{y}_{12} \\ \bar{y}_{21} \\ \bar{y}_{22} \end{pmatrix} = \begin{pmatrix} \frac{\sigma_W^2+\sigma_B^2}{n_1} & \frac{\sigma_B^2}{n_1} & 0 & 0 \\ \frac{\sigma_B^2}{n_1} & \frac{\sigma_W^2+\sigma_B^2}{n_1} & 0 & 0 \\ 0 & 0 & \frac{\sigma_W^2+\sigma_B^2}{n_2} & \frac{\sigma_B^2}{n_2} \\ 0 & 0 & \frac{\sigma_B^2}{n_2} & \frac{\sigma_W^2+\sigma_B^2}{n_2} \end{pmatrix} \tag{3.4}$$

In equation (3.4), it has implicitly been assumed that between subject measurements are uncorrelated, so the variance-covariance matrix is a block-diagonal one. Our goal is to determine the weights in (3.2). Standard statistical techniques, like generalized least squares (GLS), provide the following estimates for the treatment difference 2τ:

- If carry-over is not present in the model, then:

$$(w_1, w_2, w_3, w_4)^T = (-\tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}, -\tfrac{1}{2})^T$$

- But if carry-over is included, then:

$$(w_1, w_2, w_3, w_4)^T = (-1, 0, 1, 0)^T$$

It is assumed that $n_i, i = 1, 2$ patients allocated in each one of the two sequence groups, while $n = n_1 + n_2$ is the total number of patients recruited in the trial.

Note that when carry-over term is included in the model, the proposed weights do not depend on the data from the second period.

The above estimators can be written in a more concise form as follows:

$$\text{CROS} = -\frac{1}{2}\bar{y}_{11} + \frac{1}{2}\bar{y}_{12} + \frac{1}{2}\bar{y}_{21} - \frac{1}{2}\bar{y}_{22} \tag{3.5}$$

when carry-over parameter is not included in the model, and

$$\text{PAR} = \bar{y}_{21} - \bar{y}_{11} \tag{3.6}$$

when carry-over parameter is included in the model.

The proposed treatment-estimators as expressed in (3.5) and (3.6) have a simple interpretation: if the trialist can be reasonably confident that residual effects from first period therapy are not present at the start of the second period, then he can use the whole of his data to extract information about the treatment differences; on the contrary if he strongly believes that residual effects are still in existence at the beginning of the second period, then inference about the treatment difference should be based only on the first period data, which are free from any residual effects. In the second case 50% on average of the available information (second period data) is discarded.

In conclusion, if you include a carry-over term, then you should throw away half of your data in order to derive a statistically optimal treatment estimator.

3.3.1 What if CROS is used when we should use PAR in the simple carry-over model

From the discussion so far, it is clear that inclusion of a carry-over term in the 2x2 case leads to a treatment estimator which sacrifices a lot of the available information, in order to retain good statistical properties (MVUE). One may argue that data from both periods should be used to estimate treatment differences, no matter if carry-over effect is included in the model or not. In the more general case, where the carry-over term is included, although PAR estimates unbiasedly the treatment difference 2τ it has higher variance since it is a between-patient estimator. On the other hand CROS is a biased estimator, but with lower variance, since it utilizes within-patient information to estimate treatment effect. Bias and

variance of the proposed estimators are presented below:

$$\text{Bias}_{\text{PAR}} = 0 \qquad \text{V}_{\text{PAR}} = (\sigma_W^2 + \sigma_B^2)(\frac{1}{n_1} + \frac{1}{n_2}) \tag{3.7}$$

$$\text{Bias}_{\text{CROS}} = \frac{\lambda}{2} \qquad \text{V}_{\text{CROS}} = \frac{\sigma_W^2}{2}(\frac{1}{n_1} + \frac{1}{n_2}) \tag{3.8}$$

From (3.7) and (3.8), it is not difficult to evaluate the Mean Square Error (MSE) of the proposed estimators and try to figure out under which circumstances we should use the within-patient estimator CROS instead of the less precise between-patient estimator PAR. Since:

$$\text{MSE}_{\text{PAR}} = (\sigma_W^2 + \sigma_B^2)(\frac{1}{n_1} + \frac{1}{n_2}) \tag{3.9}$$

$$\text{MSE}_{\text{CROS}} = \frac{\sigma_W^2}{2}(\frac{1}{n_1} + \frac{1}{n_2}) + \frac{\lambda^2}{4} \tag{3.10}$$

the required condition for selecting CROS instead of PAR is easily proved to be:

$$\frac{\lambda^2}{2} < (\sigma_W^2 + 2\sigma_B^2)(\frac{1}{n_1} + \frac{1}{n_2}) \tag{3.11}$$

In the special case where equal number of patients are allocated to the two sequence groups, i.e $n_1 = n_2 = n$, the previous condition takes the simpler form:

$$\frac{\lambda^2}{4} < \frac{\sigma_W^2 + 2\sigma_B^2}{n} \tag{3.12}$$

By studying more carefully (3.12) we see that our final decision about which estimator is the best one to be used in the analysis, depends on the magnitude of the unknown residual effect. Generally speaking, the carry-over effect might be expected to be small, smaller than the combination of the within and between patient variance stated in the right-hand side of (3.12). As a result we are more likely to select the CROS estimator instead of the PAR, in real-life situations.

3.3.2 Combining the two estimators - Frequentist approach

The statistical properties of CROS and PAR have been extensively studied in the previous section. A typical medical statistician will be tempted to linearly combine the two estimators in order to improve upon them. Our new treatment effect estimator takes the following form:

$$\hat{\tau}_c = w\text{CROS} + (1 - w)\text{PAR} \tag{3.13}$$

In order to investigate the properties of the combined estimator, the joint distribution of PAR and CROS under the simple carry-over model is needed and it is provided below. It is assumed that equal number of patients are allocated to the sequence groups.

$$\begin{bmatrix} \text{PAR} \\ \text{CROS} \end{bmatrix} \sim N\left(\begin{bmatrix} \tau \\ \tau - \lambda/2 \end{bmatrix}, \frac{1}{n} \begin{pmatrix} 2(\sigma_W^2 + \sigma_B^2) & \sigma_W^2 \\ \sigma_W^2 & \sigma_W^2 \end{pmatrix} \right]$$

Note that $\text{Cov}(\text{PAR}, \text{CROS}) = \text{V}_{\text{CROS}}$, a property which greatly simplifies the expressions for the first and second order moments of $\hat{\tau}_c$. We would like to choose the weight w in (3.13), so that to minimize the MSE of $\hat{\tau}_c$. To that purpose, a new between-patient estimator has to be defined as follows:

$$\text{SEQ} = 2(\text{PAR} - \text{CROS}) = (\bar{y}_{21} + \bar{y}_{22}) - (\bar{y}_{12} + \bar{y}_{11}) \tag{3.14}$$

This estimator plays a key role for drawing inference about carry-over effect, since its expectation is simply that effect, while its variance is:

$$\text{V}_{\text{SEQ}} = \frac{\sigma_W^2 + 2\sigma_B^2}{n} \tag{3.15}$$

There is an easy clinical interpretation for the SEQ estimator: The sum of the responses are calculated for each patient and the averages of those sums are obtained for each sequence group. Those averages are compared between groups, and this difference forms an unbiased carry-over effect estimator. SEQ is being used for testing statistical significance of any residual effects under the simple carry-over model.

The optimal weight for minimizing the MSE of the combined estimator is as follows:

$$\text{w}_{\text{frequentist}} = \frac{\text{V}_{\text{SEQ}}}{\text{V}_{\text{SEQ}} + \lambda^2} = \frac{1}{1 + T_\lambda^2} \tag{3.16}$$

where T_λ is the t-statistic for testing $\lambda = 0$. The above expression has been derived by Jones and Wang (see [41]), who also report a simulation study for a range of values of λ and σ_B^2 which shows that the combined estimator has worse performance in terms of MSE when compared to CROS and the two stage procedure (presented in the next section). Equation (3.16) simply confirms that when the carry-over effect is negligible, more weight is put on the within-patient CROS estimator, while in the unlike alternative scenario of a huge carry-over effect the between-patient PAR estimator gets more credit.

3.3.3 Combining the two estimators - Bayesian approach

Grieve (see [26]) considers in his Bayesian analysis of the simple carry-over model, the problem of model selection. In our set-up, a discrete set of competing models (simple and no carry-over) is proposed and the Bayes factor is used for selecting a single model. For the 2x2 cross-over trial, the Bayes factor is simply the ratio of the marginal likelihood under the no carry-over model, to the marginal likelihood under the simple carry-over model. Grieve evaluates the Bayes factor against a carry-over effect as follows:

$$B_{01} = \sqrt{3n/4}\left[1 + F/\left(2n-2\right)\right]^{-n} \tag{3.17}$$

where, F is the statistic for assessing significance of the carry-over effect, while n is the number of subjects recruited in each sequence. Grieve reports that the maximum value for B_{01} occurs when $F = 0$. Although the observed F value rarely is identical to zero, in the example that will be shortly analyzed, but also in the trial considered by Grieve, the ratio $F/(2n-2)$ does seem to approach zero. If this assumption is made, then the Bayes factor simplifies to:

$$B_{01} \approx \sqrt{3n/4} \tag{3.18}$$

A general use of the Bayes factor is to form posterior estimates of parameters of interest by averaging over a discrete set of quantities derived from posterior distributions under different model assumptions. Grieve implements this idea, by combining CROS and PAR using the following weight on CROS:

$$\mathrm{w}_{\mathrm{bayes}} = \frac{\pi B_{01}}{1 + \pi B_{01}} \tag{3.19}$$

where, π is the prior odds against a carry-over effect. If we are indifferent a-priori to the choice of a model, then we can assume that $\pi = 1$. Substituting equation (3.18) into equation (3.19), the Bayesian weight on CROS takes the following form:

$$\mathrm{w}_{\mathrm{bayes}} = \frac{\sqrt{3n}}{2 + \sqrt{3n}} \tag{3.20}$$

Obviously the Frequentist approach assigns a weight in CROS that results in the smallest possible MSE for treatment effect estimation. The question, how close

is the Bayesian combined estimator when compared to its Frequentist competitor over a range of λ values, is raised. Figure (3.1) reveals that for moderate values of the carry-over difference the two approaches are indistinguishable, while the Bayesian is out-performed considerably by the Frequentist solution for low and high values of λ. The Bayesian solution though, has the distinct advantage that the weight assigned to CROS depends only on the sample size and not on unknown parameters.

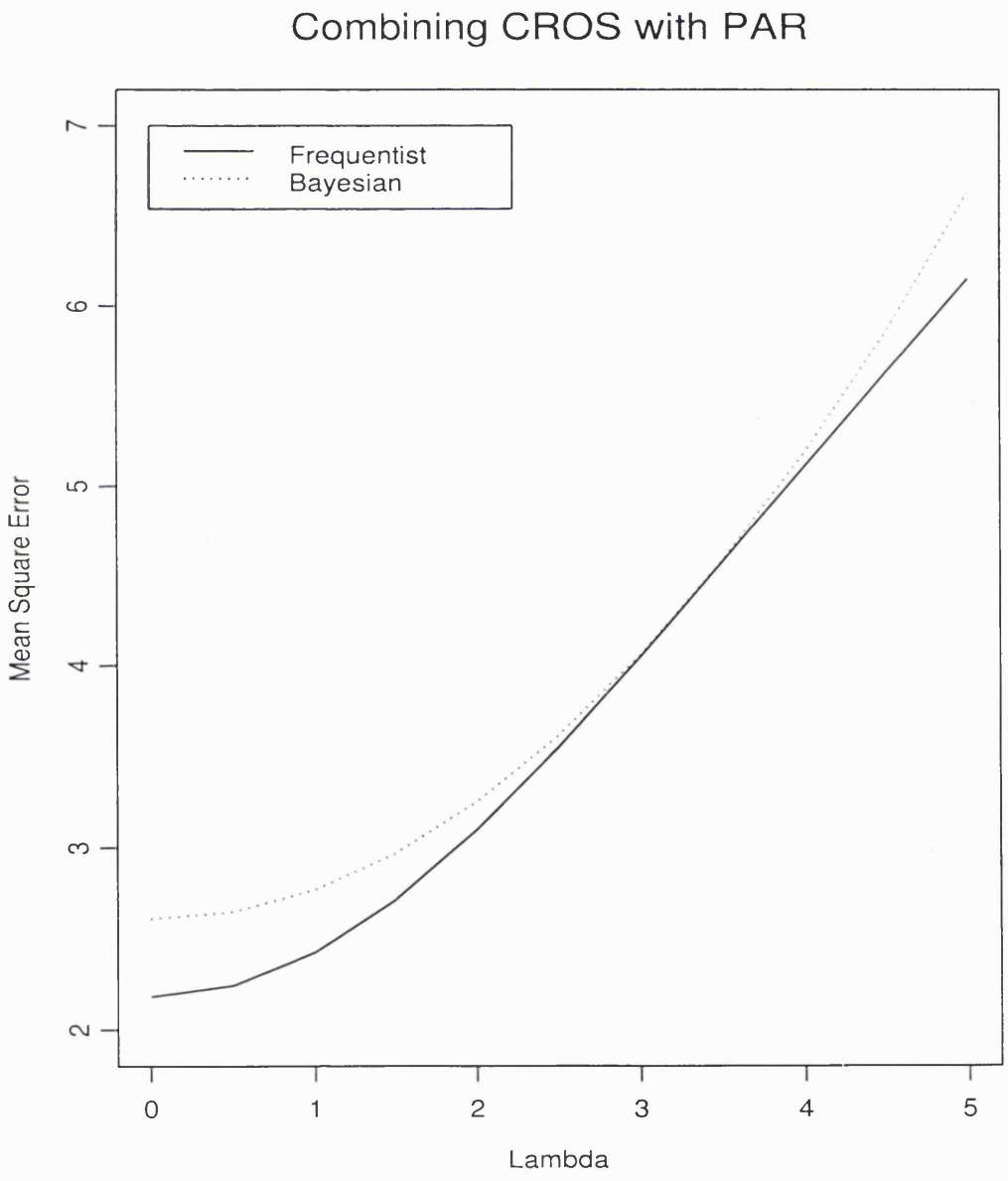


Figure 3.1: Combining the two treatment estimators

3.4 The two stage procedure

In our discussion so far, it is clear that residual effect plays a key role in drawing inferences about the treatment difference. Of course clinical knowledge could rule out carry-over occurring in any appreciable degree, but it is quite unlikely, under either the Frequentist or even the Bayesian point of view, to be in a position to incorporate any knowledge about the residual effect of a treatment without accurate knowledge of the treatment effect itself. Usually such knowledge becomes partially available at the early stages of a clinical trial (Phase I) where drugs are tested on healthy volunteers, but statisticians face difficulty in incorporating that piece of information at the later stages of either planning or analyzing the outcome of a cross-over or a parallel group trial.

A first attempt to tackle this problem was the solution proposed by Grizzle (see [30]). The idea was to test formally for the presence of carry-over effect, rather than relying on subjective opinions provided by medical doctors for its existence. His procedure composed of two stages. At the first stage the significance of carry-over was decided by comparing the means of the two sequences (SEQ estimator). It has to be said that this test for carry-over is under-powered, as noted by Senn (see [74]). At the second stage the treatment effect estimate is based on the information provided about the residual effect at the first stage. Schematically the procedure is displayed in Figure (3.2).

The detailed proposed scheme is as follows:

- **Stage 1**: Use the between-patient carry-over estimator, SEQ, to test the significance of carry-over effect at 10% level.

- **Stage 2**: If test for carry-over is significant use the between-patient PAR estimator to evaluate the extend of the treatment difference at 5% level, otherwise the within-patient treatment estimator, CROS, should be used for drawing inference about treatment, again at 5% level.

A mixed effects model is assumed throughout, since subject effects are considered as random. At first sight the two stage procedure seems to be the correct one for analyzing data from cross-over experiments, since the minimum variance unbiased (MVU) treatment estimator is the recommended one for inferential purposes,

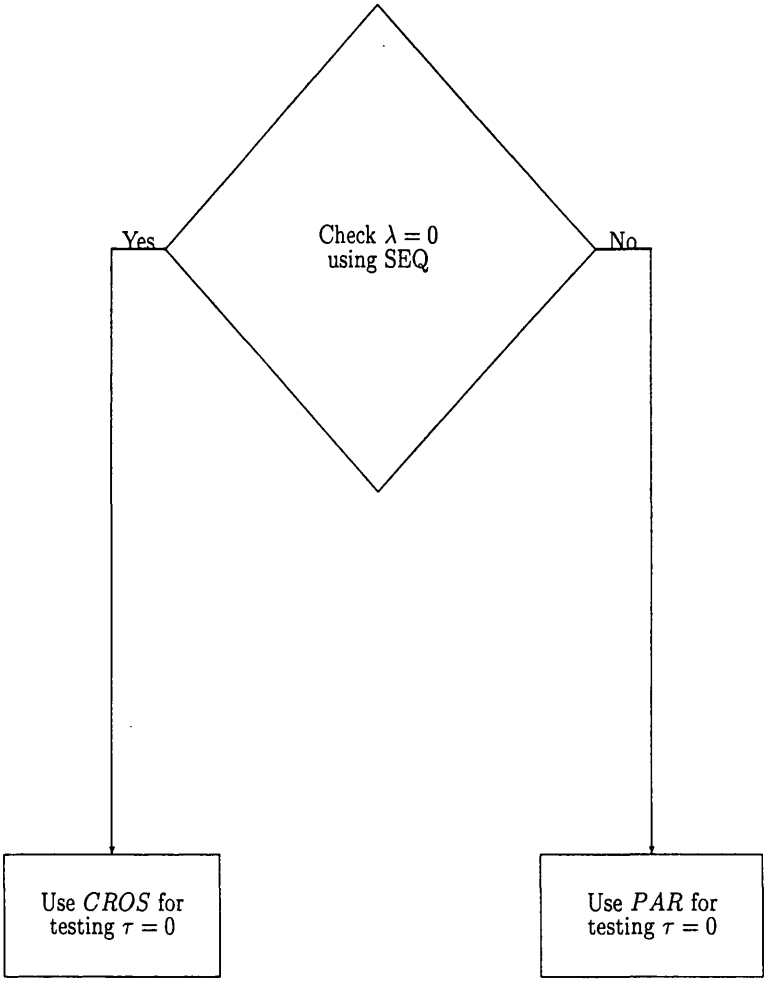


Figure 3.2: Flow diagram of the two stage procedure

irrespective of the significance or not of the residual effect. Before proceeding further it would be helpful at this point to present the marginal distributions of treatment and carry-over estimators, useful for evaluating the performance of the two stage procedure under the simple carry-over model with random subject effects. These are:

$$\text{SEQ} \sim N\left(\lambda, 4\left(\sigma_W^2 + 2\sigma_B^2\right)/n\right) \tag{3.21}$$

$$\text{PAR} \sim N\left(\tau, 2\left(\sigma_W^2 + \sigma_B^2\right)/n\right) \tag{3.22}$$

$$\text{CROS} \sim N\left(\tau - \lambda/2, \sigma_W^2/n\right) \tag{3.23}$$

Evaluation of Type I and Type II error rates of the two stage procedure require the distribution of the treatment estimator used at the second stage of the procedure, conditional on the value taken by the carry-over estimator used at the first stage. These are:

$$\text{PAR|SEQ} \sim N\left(\tau - \lambda/2 + SEQ/2, \sigma_W^2/n\right) \tag{3.24}$$

$$\text{CROS|SEQ} \sim N\left(\tau - \lambda/2, \sigma_W^2/n\right) \tag{3.25}$$

As shown by Freeman (see [21]), the scheme suffers from many deficiencies the consequences of which will be soon demonstrated. The majority of these deficiencies stem from the high correlation between SEQ and PAR. It is not difficult to show that:

$$\text{Corr}\left(\text{PAR}, \text{SEQ}\right) = \sqrt{\left(\sigma_W^2 + 2\sigma_B^2\right)/\left(2\sigma_W^2 + 2\sigma_B^2\right)} \tag{3.26}$$

but,

$$\text{Corr}(\text{CROS}, \text{SEQ}) = 0 \tag{3.27}$$

This simply implies that although the PAR estimator is unbiased for estimating the treatment effect under the simple carry-over model, it is highly biased if carried out having seen the value of the SEQ estimator (see Senn [78]). More specifically:

$$\text{E(PAR|SEQ)} = \tau + \frac{SEQ - \lambda}{2} \tag{3.28}$$

$$\text{V(PAR|SEQ)} = \text{V}_{\text{CROS}} = \frac{\sigma_W^2}{n} \tag{3.29}$$

A further implication of the high correlation between PAR and SEQ, is that if the PAR estimator is chosen at the second stage, then the size of the test for investigating a treatment difference using PAR, should not be set at the conventional 5% level, but at a much lower level, such as 0.5% (see Wang and Hung [90] or Senn [79]). This ensures that the overall Type I error rate of the procedure is kept at the nominal 5% level.

Another way of looking at the deficiency of the two stage procedure is by noting that conditionally on SEQ, CROS and PAR are both biased; the first by an amount of $-\lambda/2$, while the second by $(\text{SEQ} - \lambda)/2$. Note that the conditional distribution of CROS|SEQ is identical to that of CROS since the two estimators are statistically independent. So PAR|SEQ is biased regardless of the presence of carry-over effect, on the other hand CROS (or equivalently CROS|SEQ) is biased only in the presence of carry-over (see Senn [78]). Furthermore CROS and PAR have the same conditional variance given SEQ. As a consequence the treatment estimator with the smallest conditional bias (in absolute terms) should be chosen by the procedure at the second stage. But when the PAR estimator is chosen from the procedure for testing treatment effect, the difference $\text{SEQ} - \lambda$ must be large enough so that the carry-over effect is statistically different from zero. From the discussion above the PAR estimator will have higher conditional bias than its competitor CROS (or CROS|SEQ) and the same conditional variance, but the two stage procedure will select PAR instead of the more efficient CROS. In conclusion PAR is selected when it should not by the two stage procedure.

A thorough investigation of the performance of the two stage procedure was attempted 25 years later after Grizzle proposed this scheme, by Freeman (see [21]). Trialists who had analyzed data from cross-over experiments using the two stage procedure for many years, had implicitly assumed that PAR and SEQ were independent, so that they were incorrectly thinking that the overall Type I error of the procedure was 5%. Because of the high correlation between PAR and SEQ the real Type I error is 8.7% in the absence of any residual effect. In the case where carry-over effect is a small fraction of the treatment effect the CROS estimator is more powerful when compared to the two stage procedure, as will be soon demonstrated. The two stage procedure is superior to the CROS estimator,

in terms of power, only in the unlike case where carry-over effect is a substantial fraction of the treatment effect.

Before attempting a more detailed investigation of the two stage procedure a basic notation will now be introduced. Let a_s stand for the size of the test for carry-over, while a_p and a_c represent the size of the test for treatment when PAR or CROS are used respectively. Moreover $f_{\mathrm{SEQ}}(x)$,$f_{\mathrm{PAR}}(x)$ and $f_{\mathrm{CROS}}(x)$ denote the marginal densities, while $\mathrm{V}_{\mathrm{SEQ}}$,$\mathrm{V}_{\mathrm{PAR}}$ and $\mathrm{V}_{\mathrm{CROS}}$ are the unconditional variances of the estimators indicated in the subscripts. Finally z_a is the value which cuts-off the upper $a\%$ of the standard normal distribution, $\Phi(x)$ is the cumulative density of the same distribution, while c_s,c_p and c_c are the critical values for testing the hypothesis of significance for carry-over or treatment effect and defined as follows: $c_s = z_{a_s/2}\sqrt{\mathrm{V}_{\mathrm{SEQ}}}$, $c_p = z_{a_p/2}\sqrt{\mathrm{V}_{\mathrm{PAR}}}$ and $c_c = z_{a_c/2}\sqrt{\mathrm{V}_{\mathrm{CROS}}}$. More specifically, following Senn (see [78]), according to the plan of the two stage procedure the following treatment estimator is used:

$$\hat{\tau}_{TS} = \begin{cases} \mathrm{CROS}, & if \quad |\mathrm{SEQ}| < z_{a_s/2}\sqrt{\mathrm{V}_{\mathrm{SEQ}}} \\ \mathrm{PAR}, & if \quad |\mathrm{SEQ}| > z_{a_s/2}\sqrt{\mathrm{V}_{\mathrm{SEQ}}} \end{cases}$$

The evaluation of the power of the two stage procedure, requires first the calculation of the power for each arm; the left one which points to the use of CROS and the right one where PAR is used as tool for estimating treatment effect. The unconditional power of each arm is more easily evaluated by considering first the conditional power of each treatment estimator upon the possible values of the carry-over estimator (SEQ) that gave rise to that treatment estimator. Because CROS and SEQ are independent it is obvious that:

$$\begin{aligned} \mathrm{Power}(\mathrm{CROS}|\mathrm{SEQ}) &= \mathrm{Power}(\mathrm{CROS}) = \\ & 1 - \Phi(z_{a_c/2} - \frac{\tau - \frac{\lambda}{2}}{\sqrt{\mathrm{V}_{\mathrm{CROS}}}}) + \Phi(-z_{a_c/2} - \frac{\tau - \frac{\lambda}{2}}{\sqrt{\mathrm{V}_{\mathrm{CROS}}}}) \end{aligned}$$

On the contrary, because PAR and SEQ are highly correlated, in order to work out the power of that arm we have to evaluate first the power of $\mathrm{PAR}|\mathrm{SEQ} = x$, which as a function of SEQ will then be integrated out over the values of SEQ for which PAR is selected at the second stage of the procedure, as follows:

$$\mathrm{Power}(\mathrm{PAR}) = \int_{|x| \geq c_s} f_{\mathrm{SEQ}}(x)\,(\mathrm{Power}\,(\mathrm{PAR}|\mathrm{SEQ} = x))\,dx$$

where,

$$\text{Power}(\text{PAR}|\text{SEQ}=x) = 1 - \Phi\left(\frac{z_{a_p/2}\sqrt{\text{V}_{\text{PAR}}} - (\tau - \lambda/2 + x/2)}{\sqrt{\text{V}_{\text{PAR}|\text{SEQ}}}}\right) + \Phi\left(\frac{-z_{a_p/2}\sqrt{\text{V}_{\text{PAR}}} - (\tau - \lambda/2 + x/2)}{\sqrt{\text{V}_{\text{PAR}|\text{SEQ}}}}\right)$$

By combining all the above we can evaluate the power of the TS procedure as follows:

$$\begin{aligned}\text{Power (TS)} &= \text{prob (CROS is selected) Power (CROS|CROS is selected)} \\ &+ \text{prob (PAR is selected) Power (PAR|PAR is selected)}\end{aligned}$$

where,

$$\text{prob(PAR is selected)} = \text{prob(reject } \lambda = 0) = \text{prob}(|\text{SEQ}| > z_{a_s/2}\sqrt{\text{V}_{\text{SEQ}}})$$

To illustrate the performance of the two stage procedure, suppose that the real treatment difference is $2\tau = 5$, while the carry-over difference can be set at any value in the interval $(0, 2\tau)$. By assuming the within-patient variance to be half the between-patient variance (for example suppose that $\sigma_W^2 = 48$, so that $\sigma_B^2 = 96$) and by requiring the power of the CROS test to be 90%, the number of patients in each sequence group can easily be estimated to be $n = 22$. Those values are the same as those used by Jones and Lewis (see [40]) on their discussion of the usefulness of the cross-over experiments at the third phase of clinical trials. The results comparing the power of the two stage procedure with that of CROS and PAR are presented in Table 3.1.

As far as the bias is concerned the argument goes as follows:

$$\begin{aligned}E_{\text{TS}} &= \int_{|x|\le c_s} E\,(\text{CROS}|\text{SEQ}=x)\, f_{\text{SEQ}}(x)\, dx \\ &+ \int_{|x|\ge c_s} E\,(\text{PAR}|\text{SEQ}=x)\, f_{\text{SEQ}}(x)\, dx \\ &= (\tau - \lambda/2)\left(\Phi\left(z_{a_s/2} - \frac{\lambda}{\sqrt{\text{V}_{\text{SEQ}}}}\right) - \Phi\left(-z_{a_s/2} - \frac{\lambda}{\sqrt{\text{V}_{\text{SEQ}}}}\right)\right) \\ &+ \int_{|x|\ge c_s} (\tau - \lambda/2 + x/2)\, f_{\text{SEQ}}(x)\, dx \qquad (3.30)\end{aligned}$$

where the independence of CROS and SEQ estimators has been exploited once more. Turning now to the assessment of the variance of the procedure, we apply the following tower property regarding variances:

$$\begin{aligned} V_{\mathrm{TS}} &= V(E(\mathrm{CROS}|\mathrm{SEQ})) + E(V(\mathrm{CROS}|\mathrm{SEQ})) \\ &+ V(E(\mathrm{PAR}|\mathrm{SEQ})) + E(V(\mathrm{PAR}|\mathrm{SEQ})) \end{aligned} \tag{3.31}$$

The above expressions can be evaluated by using the appropriate conditional moments:

$$\begin{aligned} E(V(\mathrm{CROS}|\mathrm{SEQ})) &= \frac{\sigma_W^2}{n} prob(\text{select CROS}) \\ V(E(\mathrm{CROS}|\mathrm{SEQ})) &= 0 \\ E(V(\mathrm{PAR}|\mathrm{SEQ})) &= \frac{\sigma_W^2}{n}(1 - prob(\text{select CROS})) \\ V(E(\mathrm{PAR}|\mathrm{SEQ})) &= \frac{1}{4} V_{\mathrm{SEQ}} \\ &= \frac{1}{4}\left(\int_{|x|\geq c_s} x^2 f_{\mathrm{SEQ}}(x)\, dx - \left(\int_{|x|\geq c_s} x f_{\mathrm{SEQ}}(x)\, dx \right)^2 \right) \end{aligned} \tag{3.32}$$

The tower property of the expectation and variance operators has been used to evaluate the first and second order moments of the two-stage procedure. From Table 3.1 it is clear that the power of both CROS and the two stage procedure (TS) decrease as the carry-over difference increases. The reverse argument is true as far as the bias and variance of the above estimators are concerned. Regarding the PAR estimator one can easily notice that the values of the power, bias and variance do not depend at all on the carry-over difference.

The prime interest to the statistician involved in the analysis of a cross-over clinical trial is always which estimator should be used for the statistical analysis and unfortunately no ultimate decision can be reached towards that end. Note that only in the case when carry-over is 50% or more of the real treatment effect is the two stage procedure superior to the CROS estimator in terms of power. The PAR estimator has always the lower power compared to the other two estimators and should never be used. However the TS procedure has lower bias when compared to the CROS estimator but has substantially higher variance along the whole range of values for the carry-over difference. If one now computes the mean square error (MSE) of the two estimators the CROS estimator will be preferred as

Table 3.1: Properties of the three treatment estimators when $\tau = 5$

	Power			Type 1			Bias			Variance		
λ	CROS	TS	PAR	CROS	TS	PAR	CROS	TS	PAR	CROS	TS	PAR
0.0	0.922	0.881	0.282	0.050	0.087	0.05	0.000	0.000	0.000	2.182	6.974	13.091
0.5	0.895	0.863	0.282	0.053	0.089	0.05	-0.250	-0.139	0.000	2.182	7.018	13.091
1.0	0.861	0.839	0.282	0.063	0.098	0.05	-0.500	-0.278	0.000	2.182	7.149	13.091
1.5	0.820	0.809	0.282	0.080	0.112	0.05	-0.750	-0.414	0.000	2.182	7.368	13.091
2.0	0.772	0.772	0.282	0.104	0.131	0.05	-1.000	-0.546	0.000	2.182	7.672	13.091
2.5	0.718	0.730	0.282	0.135	0.157	0.05	-1.250	-0.673	0.000	2.182	8.059	13.091
3.0	0.658	0.683	0.282	0.174	0.187	0.05	-1.500	-0.794	0.000	2.182	8.527	13.091
3.5	0.594	0.635	0.282	0.220	0.223	0.05	-1.750	-0.908	0.000	2.182	9.070	13.091
4.0	0.528	0.584	0.282	0.273	0.263	0.05	-2.000	-1.013	0.000	2.182	9.684	13.091
4.5	0.460	0.535	0.282	0.331	0.306	0.05	-2.250	-1.110	0.000	2.182	10.363	13.091
5.0	0.394	0.488	0.282	0.395	0.349	0.05	-2.500	-1.196	0.000	2.182	11.100	13.091

having the lower MSE. Once more the PAR estimator has the worst performance in terms of MSE.

Since in most cases any residual effect from previous treatments is negligible, Table 3.1 shows that we should be quite confident in using most of the time the CROS estimator without pre-testing for carry-over effect. However in the unlikely case of a statistically significant carry-over difference the TS procedure is a viable alternative in terms of power but still inferior to CROS in terms of MSE. Note here that the comparison based on the power performance is not a fair one, because the size of the test for treatment for the TS procedure when $\lambda = 0$ is 8.7%, while for the other two alternatives the corresponding figure is 5%. The question of adjusting the size of TS so that the nominal 5% level is achieved, will be discussed in the next section.

3.4.1 Can we improve the two stage procedure?

The answer to that question lies in the percentage of time that PAR is selected by the two stage procedure. The original scheme, as proposed by Grizzle, selects PAR 10% of the time under the null. It is obvious that if TS is modified such that the size of the test for PAR is lowered, then it would be possible to fix the Type

I error rate at the nominal 5% level, but it is not clear at all if that alteration will improve the power of the procedure and make it superior to CROS. One way to adjust the TS procedure is by keeping fixed the size of the test for carry-over and the size of the test for treatment (when CROS is selected) at their original values, but adjusting the size of PAR. To illustrate the idea, suppose that the unconditional size of the test for PAR (a_{PAR}), is set according to the relation:

$$(1 - a_{SEQ})\, a_{CROS} + a_{PAR} \;=\; 0.05 \tag{3.33}$$

This relationship is approximately valid, since if P(SEQ,PAR) defines the probability that PAR and SEQ are jointly significant, then by requiring Type I error rate of the procedure to be 5% the following exact relation holds:

$$(1 - a_{SEQ})\, a_{CROS} + P\,(\text{PAR}, \text{SEQ}) \;=\; 0.05 \tag{3.34}$$

Because P (PAR, SEQ) $\leq a_{PAR}$ we conclude that

$$(1 - a_{SEQ})\, a_{CROS} + a_{PAR} \;\leq\; 0.05 \tag{3.35}$$

where equality holds when PAR and SEQ are perfectly correlated (see Senn [79]). This expression will be close to equality for the sorts of Corr (PAR, SEQ) commonly encountered in practice.

Equation (3.33) implies that the original scheme could be corrected in two different ways. According to the first plan the investigator might wish to test the significance of the treatment difference at the same pre-specified level (say 5%) irrespectively of which treatment estimator is chosen by the procedure at the second stage (i.e $a_{CROS} = a_{PAR}$). This approach requires re-setting the level of carry-over testing, but keeps the sizes of CROS and PAR equal. If we target Type I error at 5% then the permissible range of values for the common size of the test for the treatment difference lies in the interval (2.5%, 5.0%).

An alternative way of amending the procedure requires fixing the size of the test for carry-over difference at the traditional 10% level, while altering simultaneously the sizes of CROS and PAR so that equation (3.33) is satisfied. As it is obvious from that equation an increase in the size of CROS should be accompanied by a decrease in the size of PAR, if that plan is followed. This approach may have implementation difficulties, since the analyst has to decide different significance

Table 3.2: Performance of the corrected two stage procedure

	Type I error					Power				
	Plan 1		Plan 2			Plan 1		Plan 2		
a_{SEQ}	7.6%	14.8%	10.0%			7.6%	14.8%	10.0%		
	a_{CROS}									
λ	2.6%	2.7%	4.5%	5.0%	5.5%	2.6%	2.7%	4.5%	5.0%	5.5%
0.0	0.047	0.049	0.050	0.050	0.050	0.848	0.821	0.870	0.871	0.852
0.5	0.049	0.051	0.053	0.053	0.053	0.818	0.798	0.849	0.850	0.828
1.0	0.055	0.056	0.061	0.062	0.062	0.782	0.768	0.821	0.821	0.797
1.5	0.064	0.066	0.075	0.076	0.078	0.740	0.731	0.785	0.785	0.758
2.0	0.079	0.079	0.094	0.097	0.099	0.692	0.688	0.742	0.741	0.712
2.5	0.098	0.097	0.119	0.123	0.127	0.640	0.641	0.693	0.691	0.659
3.0	0.122	0.119	0.150	0.155	0.160	0.586	0.591	0.638	0.635	0.601
3.5	0.151	0.145	0.185	0.192	0.199	0.531	0.540	0.581	0.575	0.539
4.0	0.186	0.176	0.225	0.233	0.241	0.477	0.489	0.521	0.513	0.475
4.5	0.224	0.209	0.268	0.277	0.285	0.426	0.440	0.462	0.451	0.410
5.0	0.266	0.245	0.312	0.322	0.331	0.380	0.396	0.405	0.391	0.348

levels for the testing of the treatment effect, depending on the treatment estimator chosen at the first stage of the procedure. In that case the maximum value that a_{CROS} can be set at is 5.5%, while the corresponding range of acceptable values for a_{PAR} is from 0% to 5%.

Applying the first correction scheme leads to an improvement of the power of the procedure, as the size for testing the treatment difference decreases, contrary to the Type I error rate which looks to deviate from the desired 5% level. In Table 3.2 both power and Type I error rate for Plan 1 are displayed, when $a_{CROS} = a_{PAR} = 2.6\%$ (or 2.7%). From equation (3.33) it can easily be derived that the size of the test for carry-over should be set at 7.6% and 14.8% respectively. Those values were chosen on the grounds of providing best power values, while keeping the Type I error rate close to 5%.

Moving on now and studying more carefully the performance of the second correction scheme, it can be seen that power initially increases as a_{CROS} varies from

0% to 5% but decreases afterwards. Type I error rate gets closer and closer to 5% as a_{CROS} moves from 0% to 5.5%. Once more the values chosen to illustrate the performance of the second correction plan give the highest power values. It is worth mentioning here that the first correction plan alters the bias and variance of the procedure, while the second one leave them unchanged. Comparison of the two plans performance show that correction Plan 1 is less effective in improving the power of the procedure if the Type I error rate is set at about the same level for both.

In conclusion all attempts to improve the two stage procedure have failed for the whole range of carry-over values. This indicates that this procedure is rather of historical rather than actual value and by no means should be used in the future by the analyst of the cross-over experiment.

3.4.2 Another two-stage procedure

Equation (3.12) implies that the following treatment estimator can be defined:

$$\hat{\tau}_{TS2} = \begin{cases} \text{CROS}, & if \quad |\text{SEQ}| \leq 2\sqrt{(\sigma_W^2 + 2\sigma_B^2)/n} \\ \text{PAR}, & if \quad |\text{SEQ}| > 2\sqrt{(\sigma_W^2 + 2\sigma_B^2)/n} \end{cases}$$

An investigation similar to the one followed for studying the two stage procedure, reveals that the new scheme is worse in terms of power from CROS and the two stage procedure. More specifically, for the most interesting case where $\tau = \lambda = 0$, the Type I error rate is evaluated at 8.4% while the power is as low as 78%. This result shows that another attempt to define the two-stage scheme in a more rational way, has failed.

3.5 A 2x2 trial in asthma

Salbutamol is a well established bronchodilator for patients suffering from moderate or severe asthma. A recently developed bronchodilator, called formoterol, is tested against the old method in a 2x2 cross-over trial conducted on 13 children. The response measurement was peak expiratory flow (PEF). Let A denote formoterol and B salbutamol, respectively. Children were randomized to one of the two sequence groups, such that 7 of them were allocated to the sequence

group AB, and the rest to the dual group BA. After their first visit in the clinic a wash-out period of at least one day followed before their second visit. These data can be found in Grieve and Senn (see [29]). A graphical display of the asthma trial data is presented in figure (3.3). A simple ANOVA analysis where the total Sum of Squares (SS) has been split up into two components, a between and a within patient SS, is shown in Table (3.3).

Overall the new treatment gives higher mean peak expiratory flow than the old one, although the improvement seems to be higher when salbutamol is administer to the patient before formoterol. This indicates that a carry-over effect (or equivalently a treatment by period interaction) may be present. If carry-over effect is there this simply means that the persistence of salbutamol is longer when compared to that of formoterol. A clinical explanation to this phenomenon is that patient's body has been addicted to the old treatment, so that its effect dies out slowly. As a result a longer wash-out period will be needed before the residual effect of salbutamol will have completely disappeared.

But is there a carry-over effect? ANOVA Table (3.3) indicates that the carry-over effect is negligible. The improvement in PEF in favour of formoterol in the sequence BA was twice as much as that in the sequence AB. This was not because of a carry-over effect as the ANOVA table revealed, but it was due to a peculiar observation for subject 13. Particularly for this subject his first active period measurement (treated with salbutamol) was extremely low (only 90), when compared to the mean of PEF from subjects in the same sequence group and at the same period (mean=322). In the same way his second measurement (220), when treated with formoterol, although it looks more similar to the measurements taken on patients in the same sequence group and at the same period (mean=371), it is still substantial lower. The between and within subject studentized residuals are shown in figure (3.4). The outlier values observed in this figure, correspond to subject 13.

In addition, fitting "patient" as a random effect, a typical analysis of the resulting mixed effects model using restricted maximum likelihood (REML) as the method of estimation for the variance components (see Searle [73]), evaluates the between-patient component at $\hat{\sigma}_B = 66.52$ more than two times larger compared

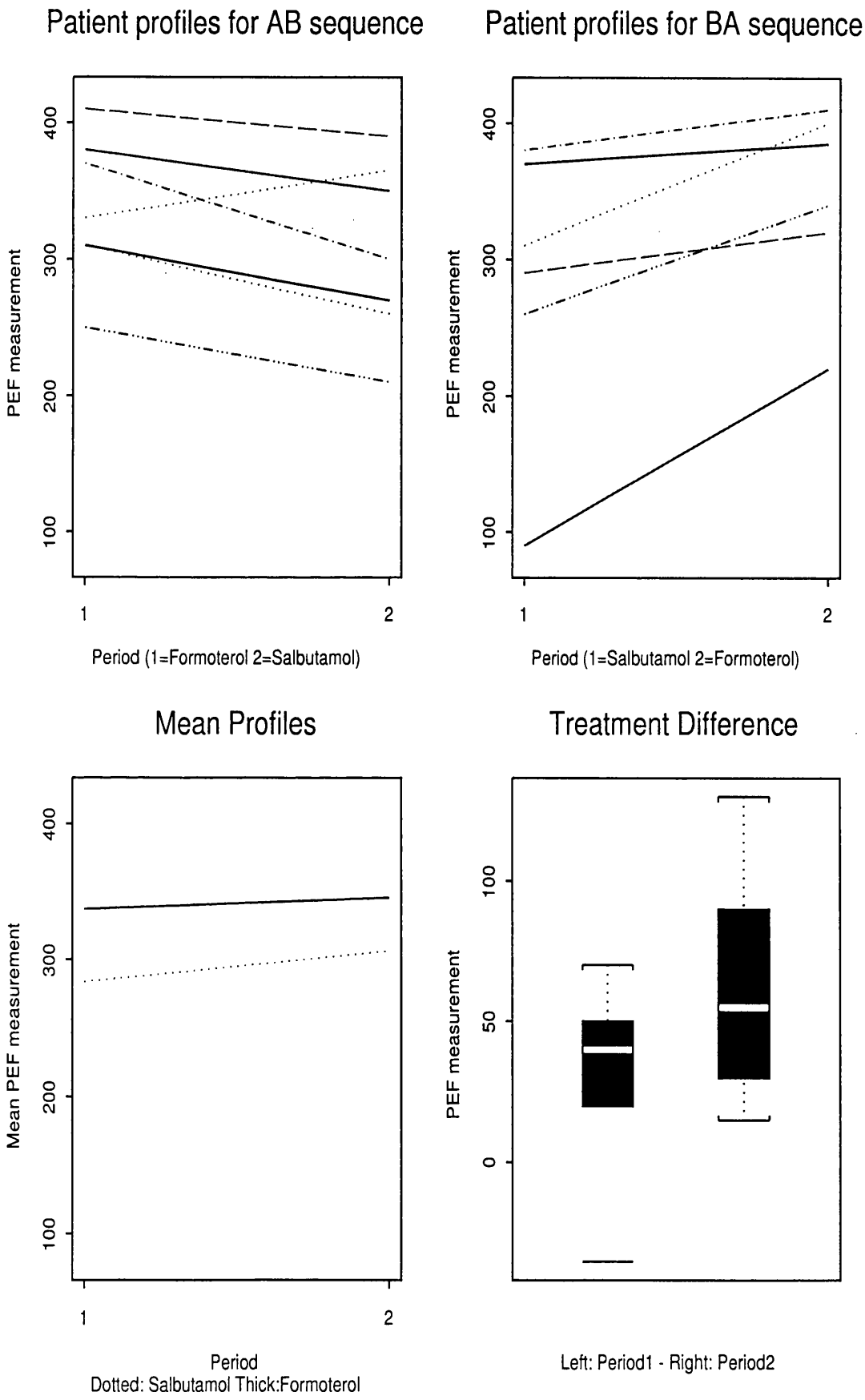


Figure 3.3: Graphical summary of the asthma trial (without baselines)

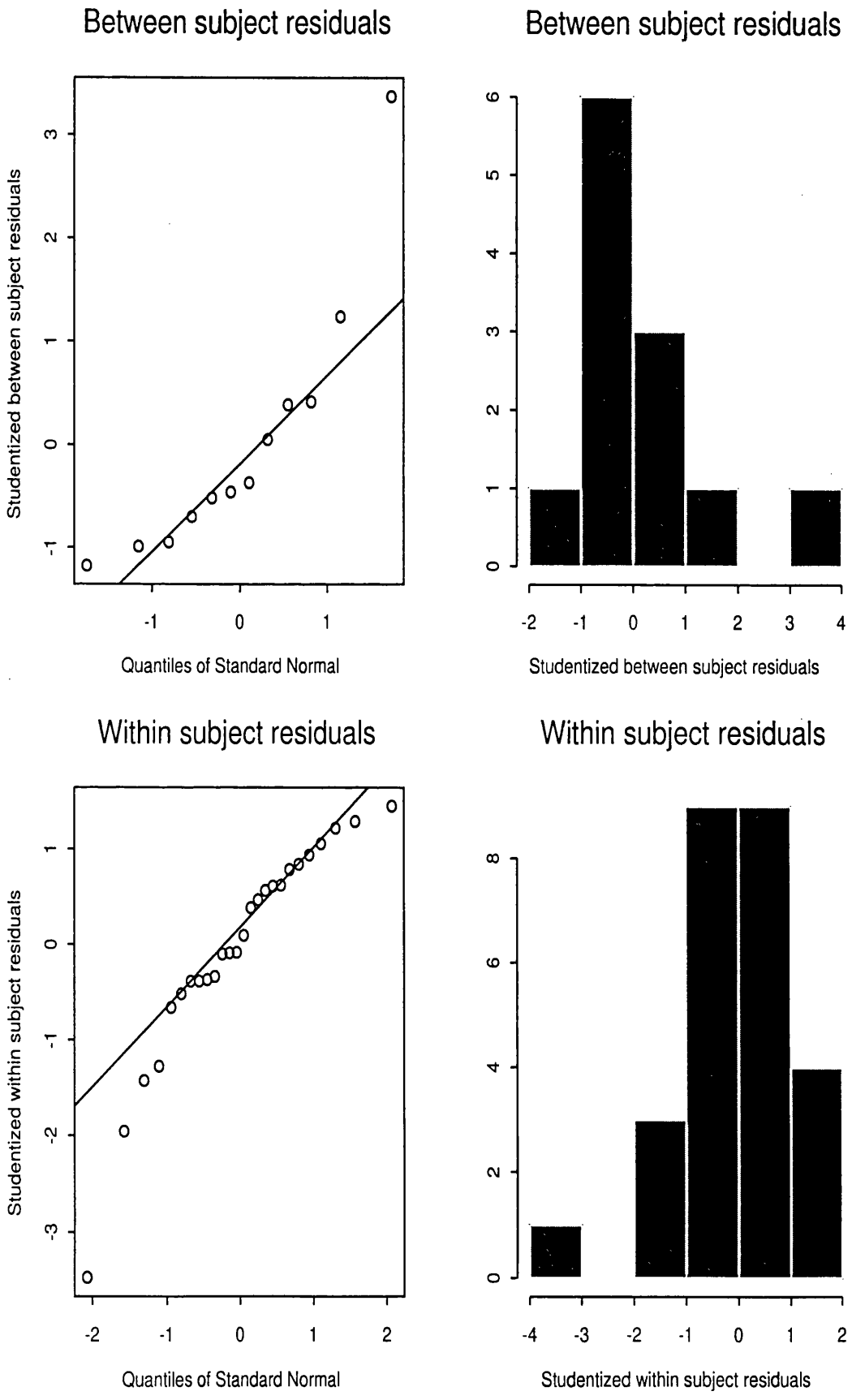


Figure 3.4: Model checking of the asthma trial (without baselines)

Table 3.3: Analysis Of Variance (ANOVA) table

Source	DF	Sum of Squares	Mean Square	F-value	p-value
Between subjects					
Carry-over	1	335.19	335.19	0.03	0.86
BS Residual	11	114878.30	10443.48		
Within Subjects					
Period	1	984.62	984.62	1.31	0.27
Treatment	1	14035.92	14035.92	18.70	0.00
WS Residual	11	8254.46	750.41		

to the estimated within-patient component $\hat{\sigma}_W = 27.39$. Evaluation of the treatment effect shows that, we expect in a future patient an increment in PEF by 46.60 (10.77), when the patient is treated with formoterol compared to being treated with salbutamol.

3.5.1 The Bayesian Approach

The power of this approach lies on the ability of the analyst to report not only an estimate of treatment (or carry-over) difference accompanied with its standard error, but the whole distribution of it, making easier the task to answer further queries of interest about these parameters. The first to present a Bayesian analysis of the cross-over experiment was Grieve (see [26]), who was able to derive explicitly the joint posterior distribution of treatment and carry-over effect, as well as, the marginal posterior distribution of the carry-over effect. Marginal inference for the treatment effect, which is the main purpose for running the clinical trial, was not possible to be evaluated analytically, but Grieve (see [27]) was able to provide a very good approximation to it, based on Patil's approximation to a Behrens-Fisher type distribution. Also the constrain $\sigma_W^2 < \sigma_B^2$ was considered in the analysis, but it turned out to make very little difference to the final conclusions.

Our approach will be based on graphical modeling theory for expressing qualitative relationships between data and unknown parameters, and on Gibbs sampling for performing the necessary computations to derive the posterior quantities of

interest. The presentation of a statistical problem using graphs, where nodes represent random quantities and missing links represent conditional independence assumptions, has the main advantage of breaking a complex model to simpler ones. This implies that the structure of the problem is easier to communicate and furthermore the graph provides the basis for the computation task (Gibbs sampling, see Gilks et al [22]).

In the 2x2 cross-over trial examined here, recall that responses (y_{ij}) on a specific patient are independent conditional on their mean μ_{ij} and the within-subject component of variance σ_W^2. Each patient's mean is a linear function of four parameters : patient, period, treatment and carry-over effect. Each one of these parameters is considered as a random variable and a prior distribution is assigned to it. Note that in the frequentist approach only the "subject effect" is taken as random, with the rest of the parameters regarded as fixed quantities. This model is known in the frequentist literature as the random intercept model.

Schematically the situation is presented in Figure (3.5). In that diagram logical links (dashed arrows) have been used for represented deterministic relationships, while solid arrows represent stochastic dependencies. The Gibbs sampler now generates a Markov chain for each variable. The chain is produced by using the conditional distribution of each unobserved node in the graph given the rest. In the long run the generated draws compose a sample from the posterior distribution of that variable. The diagram indicates the way in which a sample of a random variable is linked with random draws of other variables, so that the statistical restrictions of the model are satisfied (see Spiegelhalter [87]).

The likelihood function can be expressed as the product of the following terms:

$$\begin{aligned} y_{ijk} &\sim N\left(\mu_{ijk}, \sigma_W^2\right) \\ \mu_{ijk} &= \mu + s_{ik} + \pi_j + \tau_{d(i,j)} + \lambda_{d(i,j-1)} \\ s_{ik} &\sim N\left(0, \sigma_B^2\right) \end{aligned}$$

Fully Bayesian analysis requires the specification of prior distributions for all unknown parameters appearing in the above equations. If information regarding those parameters was available from previous cross-over trials this could be incorporated at that stage. In the absence of any prior knowledge the influence of the prior distributions in the final conclusions should be minimal. In our case the

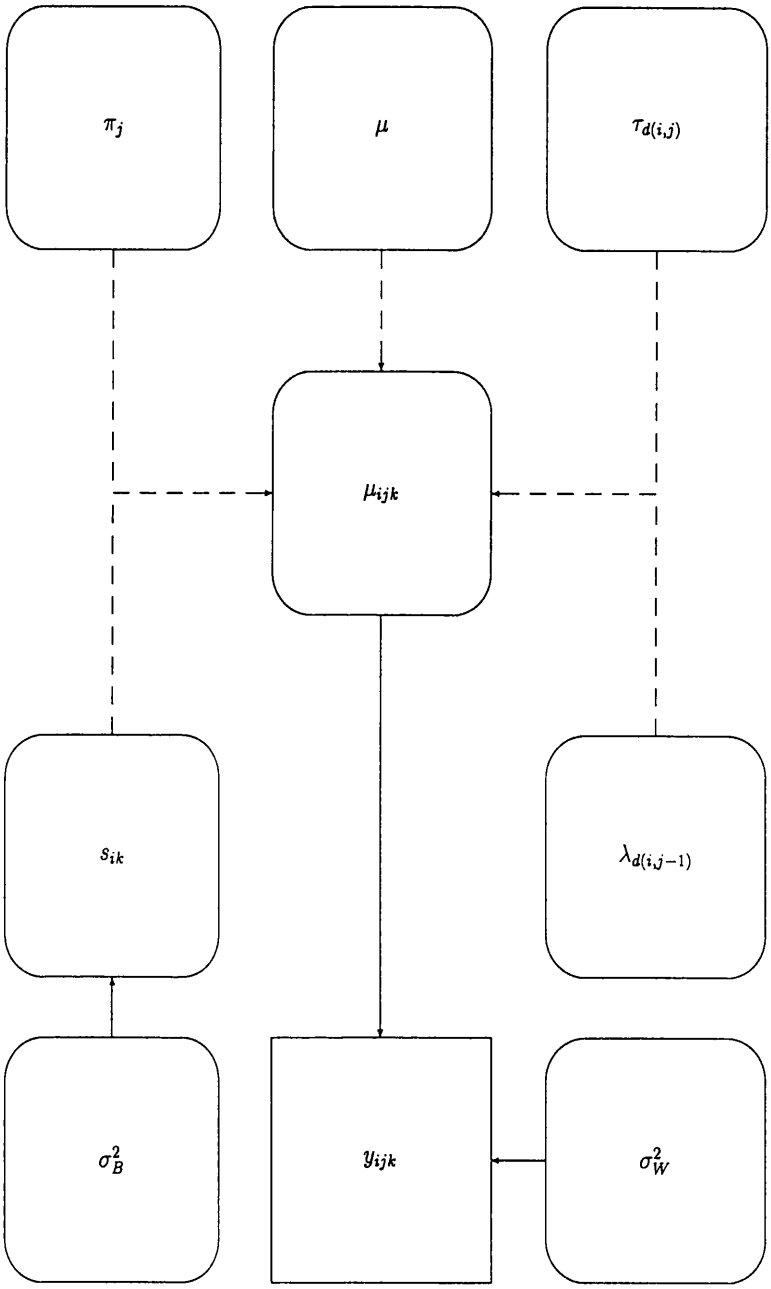


Figure 3.5: Graphical representation of the simple carryover model

following least-informative priors were chosen:

$$\begin{aligned} \mu, \pi_j, \tau_{d(i,j)}, \lambda_{d(i,j-1)} &\sim N\left(0, 10^6\right) \\ \sigma_B^{-2}, \sigma_W^{-2} &\sim \text{Gamma}\left(10^{-6}, 10^{-6}\right) \end{aligned}$$

The starting values chosen to initiate the Gibbs sampler set the location parameters at zero, while the variance components at one. A long chain was run, so that conclusions are insensitive to initial values and most importantly to ensure that the chain has converged to its limited distribution. In this example, convergence monitoring was also performed by generating five simulated sequences with different starting points and using CODA software to evaluate Gelman-Rubin's R-statistic for treatment and carry-over effect. The R-values were almost identically equal to 1, re-assuring that convergence occurred. For each variable 15000 values were generated and only the last 5000 values used for drawing inference. The sampled values used for drawing inference for the various parameters are displayed graphically in Figure (3.6). All calculations were performed using the BUGS software. BUGS code is provided at the end of this chapter.

The posterior distribution of carry-over has mean 13.30 with variance 85.70. We conclude that carry-over must be negligible, although the 95% equal-tailed confidence interval for that effect is (-141.00,201.00) indicating a wide range of possible values for the carry-over difference. This is expected since carry-over is estimated using between-patient information, which implies that no matter if either a Frequentist 95% confidence interval is formed or a 95% Bayesian HPD region is calculated the interval looks always wide. Note here that because of the symmetry of the posterior distributions for all location parameters, 95% equal-tailed intervals or 95% HPD regions lead to similar inferential conclusions.

Other posterior quantities of interest for both models, not only for treatment and carry-over difference, but also for the within and between patient variability are summarized in Table (3.4).

The advantage of the Bayesian approach is that we can form an idea of the most likely values of treatment (carry-over) effect. In Figure (3.7), the posterior distribution of 2λ indicates that the probability of that parameter lying in a symmetric interval around zero is really high. In the same figure the posterior distribution for the treatment difference suggests that under the simple carry-over model it is

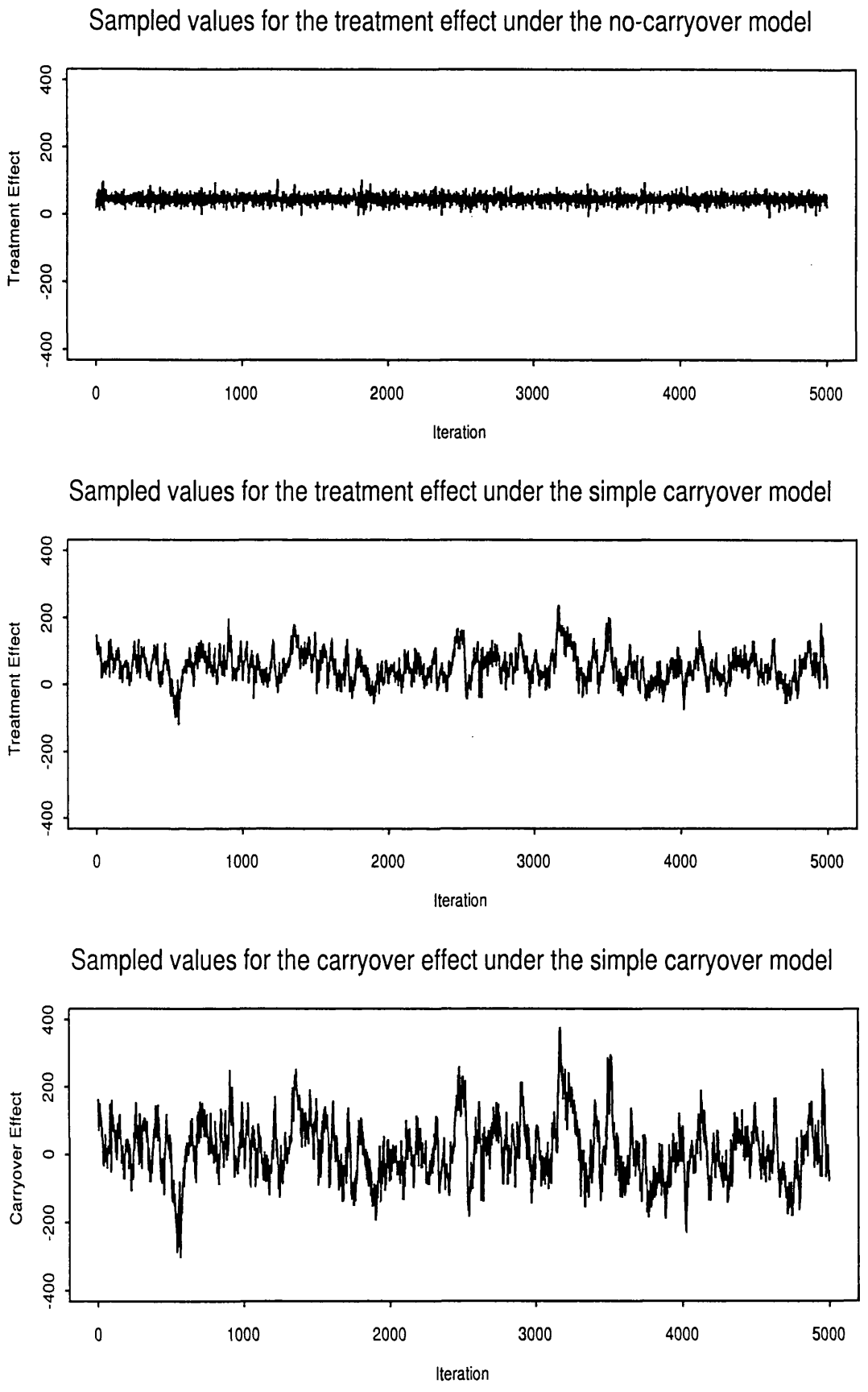


Figure 3.6: Sampled values for treatment and carry-over effect under various assumptions concerning the carry-over term

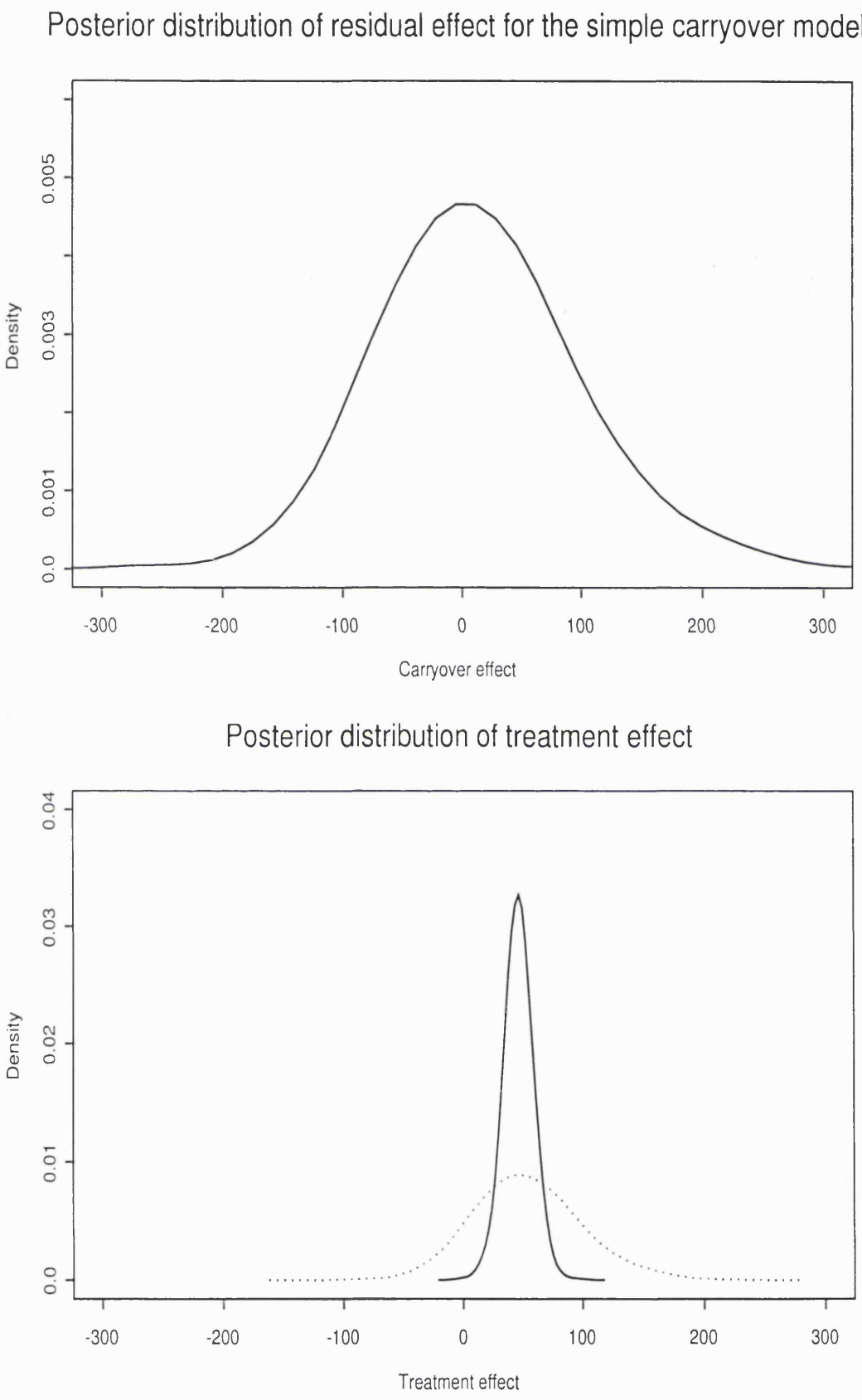


Figure 3.7: Upper half: Posterior for residual effect under simple carry-over model. Lower half: Posterior for treatment effect under model with no carry-over (solid line) and model with simple carry-over (dashed line)

more likely formoterol to give higher PEF measurements than salbutamol, i.e

$$\text{prob}(\tau > 0|y) > \text{prob}(\tau < 0|y).$$

although the possibility of a negligible treatment effect cannot be ruled out, as the posterior distribution of 2τ confirms. However this is not the most likely scenario for that effect. A further model that can be fitted, is the one where no carry-over

Table 3.4: Posterior quantities for parameters of interest

	Model with carry-over term				Model without carryover term			
Parameter	Mean	SD	95% LL	95% UL	Mean	SD	95% LL	95% UL
$2\hat{\tau}$	53.10	44.60	-28.10	150.00	46.50	17.60	21.30	70.00
$2\hat{\lambda}$	13.30	85.70	-141.00	201.00				
$\hat{\sigma}_B$	74.10	19.50	44.60	120.00	69.80	17.60	43.10	112.00
$\hat{\sigma}_W$	30.00	7.52	19.50	48.70	30.00	7.44	19.50	47.90

term is considered. Figure (3.7) shows the posterior density $p(\tau|\lambda = 0, data)$.It is clear now that, although the mean posterior treatment difference is slightly lower compared to the corresponding estimate under the simple carry-over model, the standard error of this difference is substantially lower as well in the simpler model. The 95% HPD region of the simpler model leaves no doubt about the superiority of the new treatment.

The posterior quantities for the within and between patient variability are also affected to some extend by the presence or not of the carry-over term in the model. Both the posterior mean and SD for $\hat{\sigma}_B$ are inflated when carry-over term is included in the model. This is due to the fact that λ and σ_B utilize similar between subject information for inferential purposes and absence of anyone of the two parameters affects our estimate for the other. On the contrary the effect on $\hat{\sigma}_W$ seems to be smaller. Finally the data does not exclude the possibility of the within-patient variance being larger than the between-patient one, although looking at the posterior means of the variance components the posterior probability of that scenario is expected to be small.

In conclusion, the Bayesian analysis of our 2x2 cross-over trial without baselines support the conclusions drawn from the Frequentist approach. Further insight on how treatment and carry-over effect affect each other, is also gained.

3.6 The use of baselines

In the basic 2x2 cross-over experiment, already considered, it is quite common in practice for measurements to be taken on patients just before the start of the first treatment period, and after the completion of the first treatment period and prior to the start of the second treatment period, i.e at the end of the wash-out interval. Let $[y_{i1k}, y_{i2k}, y_{i3k}, y_{i4k}]$ denote the four measurements collected on the k^{th} patient randomized in the i^{th} sequence group. The two baseline measurements provide information about the physical condition of the patient before the start of each treatment period, but they do not help at all in assessing the treatments themselves. Note here that the second baseline measurement might have been influenced by the treatment administered in the first period due to a first order carry-over effect, denoted as θ. Of course carry-over from the first treatment period, might be present when the second treatment measurement is taken and this will be referred to as the second order carry-over effect, and denoted as λ in what follows. An adequate wash-out period would suffice to eliminate both carry-over terms. The linear model adopted here is:

$$y_{ijk} = \mu + \gamma_i + s_{ik} + \pi_j + \tau_{d(i,j)} + \theta_{d(i,j-1)} + \lambda_{d(i,j-2)} + \epsilon_{ijk} \tag{3.36}$$

where,

$$s_{ik} \sim N(0, \sigma_B^2) \quad \text{and} \quad \epsilon_{ijk} \sim N(0, \sigma_W^2)$$

and,

$$\theta_{d(i,0)} = \lambda_{d(i,0)} = \lambda_{d(i,1)} = 0 \quad \text{for} \quad i = 1...n_i, j = 1...4, k = 1, 2.$$

Similar notation to the one used in the 2x2 cross-over experiment without baselines is utilized throughout. For example, carry-over from the first active treatment period to the second active treatment period is denoted by λ, while a new symbol is used to refer to carry-over from first active treatment to wash-out period; namely θ. Once more the second order carry-over term is confounded with the treatment by period interaction. As in the simple carry-over model the conventional uniform covariance structure is implied for observations taken on a

subject, while observations from different subjects are assumed independent. A further term introduced in the model is that of the sequence effect γ_i. Its inclusion ensures that treatment and carry-over terms will be estimated using within subject contrasts.

One way to handle baseline measurements is that proposed by Kenward (see [42]). They use OLS estimators, as they are optimal under uniform covariance structure. These estimators have the form $\hat{c}_1 - \hat{c}_2$, where $\hat{c}_i$ is a contrast of the four cell means $[\bar{y}_{i1.}, \bar{y}_{i2.}, \bar{y}_{i3.}, \bar{y}_{i4.}]$ in sequence i. A treatment or carry-over estimator is completely determined once the weights in those contrasts are explicitly defined. A three stage procedure for drawing inference about the treatment difference allowing at the same time for any adjustments caused by the presence of carry-over terms is now described. Schematically strategy 2 is presented in Figure (3.8).

- **Step 1** : Test the significance of the first order carry-over difference at 10% level, by comparing the two baseline measurement on each subject. Least squares analysis points to the use of the following set of weights:

$$w_{\hat{\theta}} = (\frac{1}{2}, 0, -\frac{1}{2}, 0)$$

- **Step 2** : If the first order carry-over term is found statistically significant from zero then keep that term in the model and check for the significance of the second order carry-over term at 10% level, by comparing the first baseline measurement with the average of the two treatment measurements for each patient. The proposed set of weights in that occasion is:

$$w_{\hat{\lambda}_2} = (1, -\frac{1}{2}, 0, -\frac{1}{2})$$

  On the other hand if the test for the first order carry-over term allows the deletion of that term from the model, then the test for the second order carry-over term (at 10% level again) is based on the comparison of the average of the baseline measurements to the average of the treatment measurements for each patient, suggesting the following scheme of weights:

$$w_{\hat{\lambda}_1} = (\frac{1}{2}, -\frac{1}{2}, \frac{1}{2}, -\frac{1}{2})$$

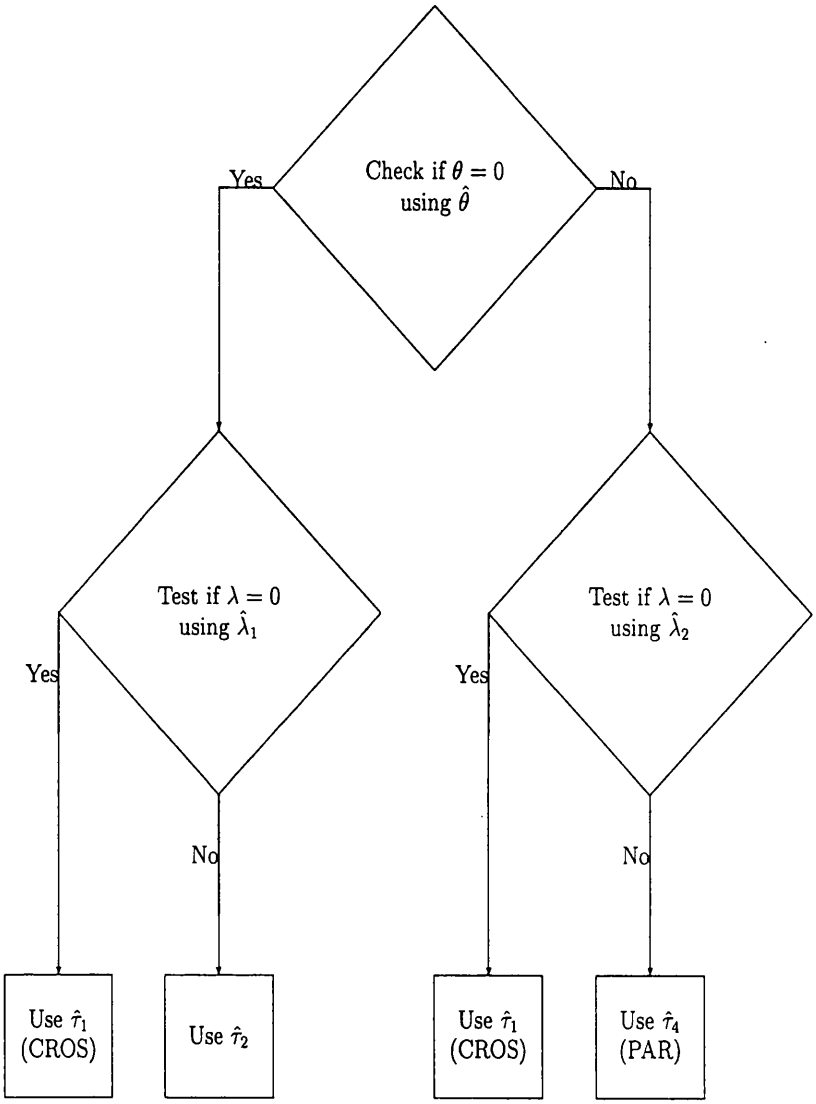


Figure 3.8: Flow diagram of the three stage procedure (staregy2). Strategy 1 is described by a similar diagram by eliminating the third path in the above figure

- **Step 3** : The set of weights for the treatment difference depends upon which carry-over terms have been deleted and which have been kept so far in the model, before entering this final step of analysis. In the most likely case where both carry-over terms have been dropped or only the first order carry-over is still retained in the model, the difference between the treatment measurements for each patient determine the treatment estimator. Note that the baseline measurements are completely ignored in that occasion. This estimator is similar to CROS, used in the analysis of the 2x2 cross-over trial without baselines and without carry-over effect. Each patient's readings are weighted as follows:

$$w_{\hat{\tau}_1} = w_{\hat{\tau}_3} = (0, -\frac{1}{4}, 0, \frac{1}{4})$$

In the least favorable and quite unlikely case of keeping both carry-over terms in the model the treatment estimator is based on the difference between the first baseline and the first treatment measurement for each subject, i.e. the baseline and treatment measurements from the second period are wasted. The weights, presented below, are similar to the ones used on estimating the treatment difference in the classical 2x2 case with carry-over (PAR estimator).

$$w_{\hat{\tau}_4} = (\frac{1}{2}, -\frac{1}{2}, 0, 0)$$

Turning now to the final and most unreasonable possibility of deleting the first order carry-over term but keeping the second one, the difference between the average of baseline measurements and first treatment measurement for each patient forms the treatment estimator. In that case only the second treatment reading is discarded for each patient, pointing to the following scheme of weights:

$$w_{\hat{\tau}_2} = (\frac{1}{4}, -\frac{1}{2}, \frac{1}{4}, 0)$$

All the above tests for the treatment difference are carried out at 5% level.

The above scheme will be called strategy 2. An alternative, simpler scheme, called strategy 1, differs from the previous one in the way that handles the deletion of carry-over terms. More specifically according to strategy 1, if the first

order carry-over term is removed from the model then this automatically implies that the second one is dropped as well (see Kenward [42]). On the contrary in strategy 2 it is feasible to keep the second order carry-over term without including in the model the first one. Because of the more reasonable way of handling carry-over terms, strategy 1 is expected to have a better performance compared to strategy 2 in terms of power or MSE, as would be soon demonstrated.

Before that, a comparison of the treatment estimators proposed in the cross-over experiment with and without baselines is in order. In both cases the estimator used when no carry-over terms retained in the model or when only the first order carry-over is present in the cross-over with baselines, are biased but with lower variance, compared to the estimator used when the full set of carry-over terms is included in either case. A reasonable query at that point is raised. Are the deficiencies of the two stage procedure inherited to the three stage procedure as well? The estimators used in each stage to decide the significance or not of the corresponding terms are highly correlated, and this might force the power of the three stage procedure to be lower than the power of $\hat{\tau}_1$ (CROS), which corresponds to the treatment estimator without pre-testing for carry-over effects at all. A thorough investigation of both strategies requires marginal and conditional probability distributions of various estimators. These are displayed in Table (3.5). A

Table 3.5: Three stage procedure

Marginals of treatment and carryover estimators	
$\hat{\theta} \sim N\left(\theta, \sigma_W^2/n\right)$	$\hat{\tau}_1, \hat{\tau}_3 \sim N\left(\tau - \lambda/2, \sigma_W^2/(4n)\right)$
$\hat{\lambda}_1 \sim N\left(\lambda - \theta, 2\sigma_W^2/n\right)$	$\hat{\tau}_2 \sim N\left(\tau - \theta/2, 3\sigma_W^2/(4n)\right)$
$\hat{\lambda}_2 \sim N\left(\lambda, 3\sigma_W^2/n\right)$	$\hat{\tau}_4 \sim N\left(\tau, \sigma_W^2/n\right)$
Conditionals of 2^{nd} order carryover given 1^{st} order carryover estimators and of treatment estimators given 1^{st} and 2^{nd} order carryover estimators	
$\hat{\lambda}_1 \vert \hat{\theta} \sim N\left(\lambda - \theta, 2\sigma_W^2/n\right)$	$\hat{\lambda}_2 \vert \hat{\theta} \sim N\left(\lambda - \theta + \hat{\theta}, 2\sigma_W^2/n\right)$
$\hat{\tau}_1 \vert \hat{\lambda}_1, \hat{\theta} \sim N\left(\tau - \lambda/2, \sigma_W^2/(4n)\right)$	$\hat{\tau}_2 \vert \hat{\lambda}_2, \hat{\theta} \sim N\left(\tau - \lambda/2 + \hat{\lambda}_1/2, \sigma_W^2/(4n)\right)$
$\hat{\tau}_3 \vert \hat{\lambda}_2, \hat{\theta} \sim N\left(\tau - \lambda/2, \sigma_W^2/(4n)\right)$	$\hat{\tau}_4 \vert \hat{\lambda}_2, \hat{\theta} \sim N\left(\tau - \lambda/2 + \hat{\lambda}_2/2, \sigma_W^2/(4n)\right)$

worth emphasizing property of either strategy is that the treatment effect estimator is independent from the first order carry-over effect estimator, conditionally

upon the second order carry-over difference estimator.

According to that flow diagram three distinct treatment estimators are proposed at the end of each of the four paths. One of them $\hat{\tau}_1$ (similar to CROS) is used twice, while $\hat{\tau}_4$ (similar to PAR) only once. This is an early indication that this procedure might have a good performance in terms of power and MSE, depending upon the proportion of time $\hat{\tau}_4$ is used. But the crucial question is if the three stage procedure is superior than using directly $\hat{\tau}_1$. To answer that question a similar approach to the one used at the investigation of the two stage procedure will be used for the evaluation of power, bias and variance of both strategies as follows:

$$\begin{aligned}
\text{Power (Strategy 2)} &= \sum \int_{\theta_1}^{\theta_2} \int_{\lambda_1}^{\lambda_2} \text{prob}\left(|\hat{\tau}| > c_{\hat{\tau}} | \hat{\theta} = x, \hat{\lambda} = y\right) f_{\hat{\theta},\hat{\lambda}}(x,y)\, dx\, dy \\
\text{E (Strategy 2)} &= \sum \int_{\theta_1}^{\theta_2} \int_{\lambda_1}^{\lambda_2} E\left(\hat{\tau} | \hat{\theta} = x, \hat{\lambda} = y\right) f_{\hat{\theta},\hat{\lambda}}(x,y)\, dx\, dy \\
\text{V (Strategy 2)} &= \sum \left[E\left(V\left(\hat{\tau}|\hat{\lambda},\hat{\theta}\right)\right) + V\left(E\left(\hat{\tau}|\hat{\lambda},\hat{\theta}\right)\right)\right] \qquad (3.37)
\end{aligned}$$

where the summation is over the four paths, while the estimators and the limits of the integrals used for each path are decided according to the plan of the strategy. Assuming that the real treatment difference is 5, the first carry-over difference is a fraction of the treatment difference and finally that the second order carry-over difference is a fraction of the first one, the performance of both strategies is summarized in Table (3.6). In the first third of Table (3.6) the second order carry-over difference has always been kept at zero, while the first one is increased gradually by 10%, reaching finally the treatment effect. In the second and final part of that table, the first order carry-over was taken two and four times respectively higher than the magnitude of the second one.

The most interesting message from that investigation is that when both carry-over terms are negligible (first line of Table (3.6)) the Type I error rate is about 6% for strategy 1, while the corresponding figure for strategy 2 raises to 8.5% similar to the Type I error of the two stage procedure without baselines. Moreover the power of strategy 1 is very close to that of the CROS estimator in the classical 2x2 cross-over, while that of strategy 2 is slightly lower. A simple explanation for the better behaviour of strategy 1 is that more than 88% of the time

Table 3.6: Performance of strategies 1 and 2

		Strategy 2				Strategy 1			
θ	λ	Type 1	Power	Bias	Var	Type 1	Power	Bias	Var
0.00	0.00	0.084	0.872	0.000	4.861	0.059	0.910	0.000	3.136
0.50	0.00	0.085	0.861	-0.062	4.900	0.060	0.910	0.037	3.144
1.00	0.00	0.088	0.847	-0.125	5.014	0.060	0.910	0.072	3.167
1.50	0.00	0.091	0.832	-0.189	5.194	0.060	0.909	0.106	3.202
2.00	0.00	0.096	0.817	-0.252	5.425	0.061	0.908	0.136	3.248
2.50	0.00	0.102	0.802	-0.318	5.690	0.062	0.907	0.162	3.300
3.00	0.00	0.108	0.788	-0.379	5.970	0.063	0.906	0.184	3.356
3.50	0.00	0.114	0.775	-0.434	6.245	0.064	0.905	0.201	3.415
4.00	0.00	0.120	0.766	-0.478	6.497	0.065	0.904	0.214	3.475
4.50	0.00	0.124	0.760	-0.510	6.711	0.065	0.902	0.221	3.537
5.00	0.00	0.128	0.757	-0.526	6.876	0.066	0.901	0.223	3.604
1.00	0.50	0.090	0.839	-0.246	4.976	0.067	0.885	-0.149	3.248
2.00	1.00	0.105	0.800	-0.478	5.298	0.076	0.854	-0.292	3.568
3.00	1.50	0.128	0.760	-0.675	5.768	0.095	0.817	-0.425	4.056
4.00	2.00	0.155	0.722	-0.832	6.307	0.119	0.777	-0.542	4.657
5.00	2.50	0.184	0.686	-0.947	6.846	0.147	0.735	-0.652	5.321
1.00	0.25	0.088	0.844	-0.186	4.983	0.061	0.898	-0.039	3.205
2.00	0.50	0.098	0.813	-0.362	5.314	0.066	0.884	-0.079	3.396
3.00	0.75	0.113	0.783	-0.521	5.764	0.072	0.867	-0.124	3.670
4.00	1.00	0.129	0.758	-0.649	6.218	0.081	0.850	-0.178	3.985
5.00	1.25	0.142	0.742	-0.737	6.578	0.090	0.833	-0.240	4.309

the treatment estimator similar to CROS ($\hat{\tau}_1$) is used while less than 3% of the time the inefficient similar to PAR ($\hat{\tau}_4$) estimator is chosen by the procedure. On the contrary strategy 2 selects $\hat{\tau}_1$ only 81% of the time, $\hat{\tau}_4$ less than 3% while the rest of the time $\hat{\tau}_2$ is chosen.

On those grounds strategy 1 can be considered as an improved version of the two stage procedure at the cost of obtaining two further measurements on each patient. Moreover strategy 1 has Type I error rate and power similar to that of CROS for the whole range of first order carry-over values, provided that the second order carry-over term is kept at zero. This indicates that as long as carry-over terms are handled in a rational way, their inclusion into the model does not affect to a large extend the quality of the estimation procedure concerning treatment effect.

It is also clear from the first third of Table (3.6) that strategy 1 overestimates the real treatment difference, while strategy 2 under-estimates it. However both strategies underestimate the treatment difference in the rest of the cases. Overall in absolute terms, strategy 2 has higher bias and variance compared to strategy 1, and worst performance in terms of Type I error and power, over the whole range of first and second order carry-over combinations considered here. So, strategy 1 should be preferred to strategy 2 for analyzing data from cross-over trials with baselines, although both are inferior compared to using always CROS. In conclusion the incorporation of carry-over terms in the model adversely affects the properties of the final treatment estimator proposed, regardless of the availability of baselines measurements.

3.6.1 A 2x2 cross-over trial with baselines

In the cross-over experiment, already examined, further information was available on each child, i.e. two baseline measurements were taken before the start of each treatment period. A graphical summary of these data is provided in Figure (3.9). The models considered in this subsection include various combinations of carry-over terms. In addition results from analysis without baselines are also reported. This will help to assess the predictive ability of baseline measurements in evaluating treatment and carry-over effects. Following Grieve's notation (see

[27]) four models are fitted:

- **M2** : The saturated model in which carry-over effects of both kinds are included. The trialist, most of the time, allows wash-out and treatment periods to be of the same time length. As a consequence it is quite unlikely for the second order carry-over term to be present at all, unless there is a treatment by period interaction. In our example the second treatment measurement was taken two days after the first one, so that a significant treatment by period interaction is quite unlikely. Overall there is a small chance for this model to have generated our data.

- **M11** : In this model only the second order carry-over term is fitted. This might look unreasonable since if the wash-out has been chosen long enough to eliminate the first order carry-over, why should the second order carry-over be present? In fact carry-over here represents a psychological carry-over. This simply means that some patients suffered discomfort during the first treatment period (probably they were given placebo which does not relieve pain) and they feel unhappy in entering the second treatment period. This feeling might influence the measurement of the second treatment. An alternative motivation (as in model M2) for considering the second order carry-over but excluding the first one, is the presence of a treatment by period interaction.

- **M12** : Only first order carry-over term is now considered, i.e the wash-out period prevent the residual effect of the first active treatment period to be present when the second active treatment measurement was taken, but it was not long enough to eliminate the first treatment's residual effect at the time second baseline measurement was obtained. This model is in accordance with the statistical hierarchy the analyst should follow in a backwards elimination procedure, i.e. second order carry-over is considered for elimination before the first one. Among all the models considered this is the only one that handles carry-over terms in a rational way.

- **M0** : No carry-over terms are now included. The statistician in collaboration with the trialist have already agreed in advance that the proposed

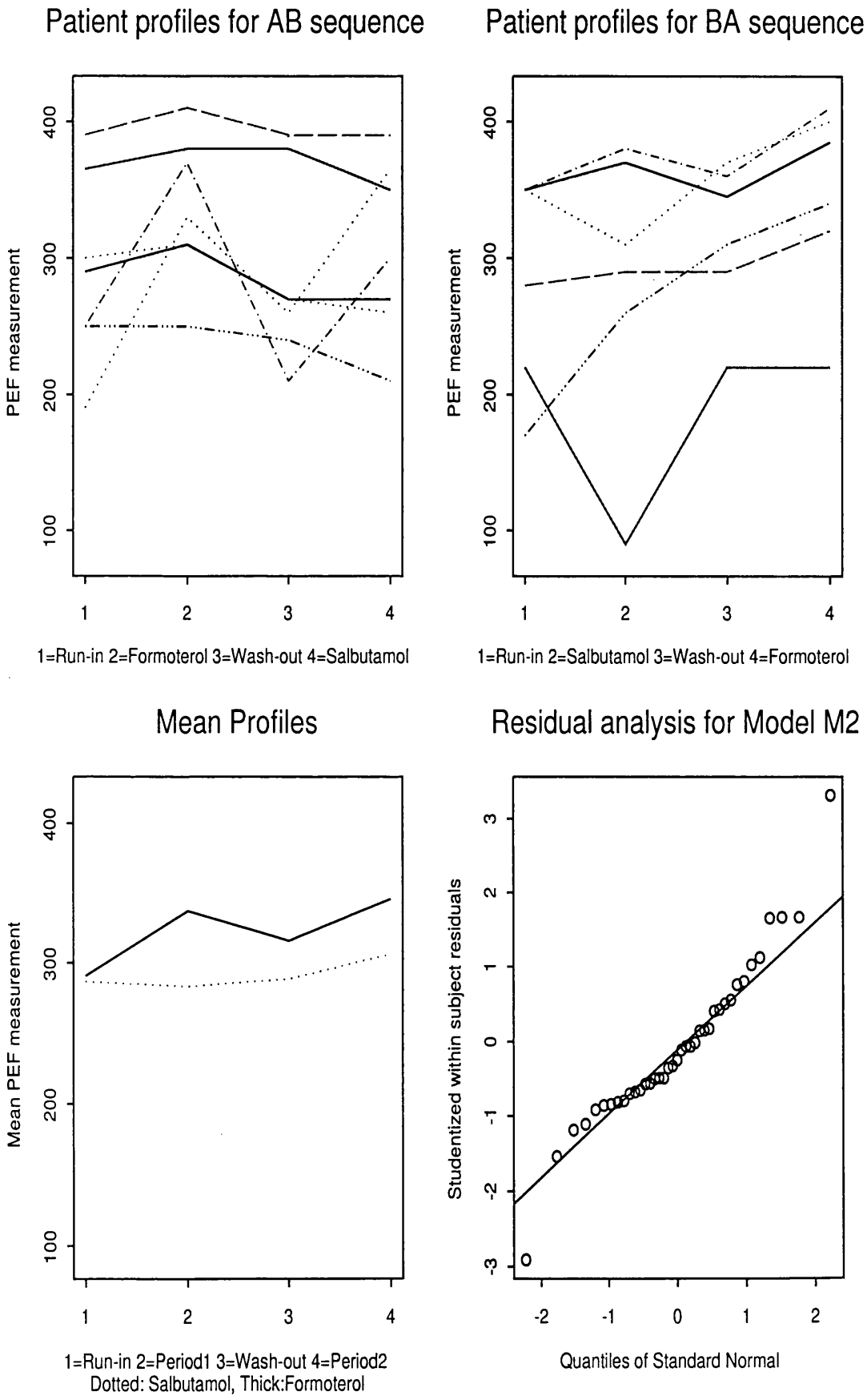


Figure 3.9: Graphical summary of the asthma trial with baselines.

length of the wash-out period is long enough to eliminate any residual effects from the first treatment in all subsequent periods. This model is mostly favored by the practical-oriented data analysts.

If patient effect is taken as random and both carry-over terms are included in the model while REML is the estimation method for the variance components, then formoterol gives higher PEF measurements by an amount of 49.76 (30.26) units when compared to salbutamol (see Table 3.7). Also both carry-over effects are negligible. The between-children variability is twice as high as the within one. When both carry-over terms are removed from the model, treatment effect is estimated at about the same level as before, but with a much smaller standard error, $2\hat{\tau} = 46.60(15.09)$. By keeping both carry-over terms the treatment effect is obscured. However when these terms are removed the new treatment shows its superiority. The inclusion of carry-over terms influences to an appreciable extend our inference about treatment effect, even when baselines are used in the analysis. In some circumstances the collection of baseline measurements might

Table 3.7: Frequentist analysis of a 2x2 trial with and without baselines

Models with baseline measurements								
Model	M0		M12		M11		M2	
$2\hat{\tau}$	46.60	(15.09)	46.60	(14.91)	65.41	(26.23)	49.76	(30.26)
$2\hat{\lambda}$	——	——	——	——	37.61	(42.84)	6.30	(52.41)
$2\hat{\theta}$	——	——	-33.41	(24.34)	——	——	-31.30	(30.26)
$\hat{\sigma}_B$	60.57		60.65		60.55		60.56	
$\hat{\sigma}_W$	38.37		37.90		38.50		38.46	
Models without baseline measurements								
Model	No carryover				Simple carryover			
$2\hat{\tau}$	46.60	(10.77)			53.80	(41.62)		
$2\hat{\lambda}$	——	——			14.40	(80.40)		
$\hat{\sigma}_B$	66.52				69.61			
$\hat{\sigma}_W$	27.39				27.39			

influence the precision with which treatment effect is estimated, although it is

unclear how the estimation of carry-over terms is affected.

In this specific example it seems that the availability of baseline measurements has some effect, not only on the magnitude of the treatment or carry-over difference, but most importantly in their estimated standard errors. In absolute terms, estimated carry-over effect from first active to second active treatment period for the model with baselines is double when compared to the corresponding figure of the model without baselines, while its standard error is about half. As far as the treatment effect is concerned, the estimates along with their standard errors from both models are comparable.

Note that the sequence effect is always included in the above models. In addition the precision with which the treatment effect is estimated gets higher as carry-over terms are eliminated from the full model. For purpose of completeness only, it is worth noting that the correlation between any two measurements on a child (intra-class correlation coefficient) is estimated at 0.71, regardless if any carry-over terms are included or not in the model and if baselines are used or not in the analysis.

3.6.2 The Bayesian Solution

The set of models studied here are identical to the ones considered in the previous section, but perceived from a Bayesian perspective. Least-informative priors, similar to the ones used for the analysis of the same dataset without baselines, were assigned to each unknown quantity and Gibbs sampling was used for the derivation of the relevant posterior distributions. Kernel estimates for the posterior density of the treatment and the carry-over terms are displayed in Figure (3.10). Posterior summaries for the parameters of interest, are presented in Table (3.8).

Once more it is confirmed that if both carry-over terms are included in the model then there is a non negligible posterior probability for the treatment difference to lie in a symmetric interval around zero. On the contrary the elimination of any residual term (first or second) in the model seems to produce stronger evidence that formoterol gives on average higher PEF measurements than salbutamol. More specifically the irrational model M11 indicates that formoterol is superior

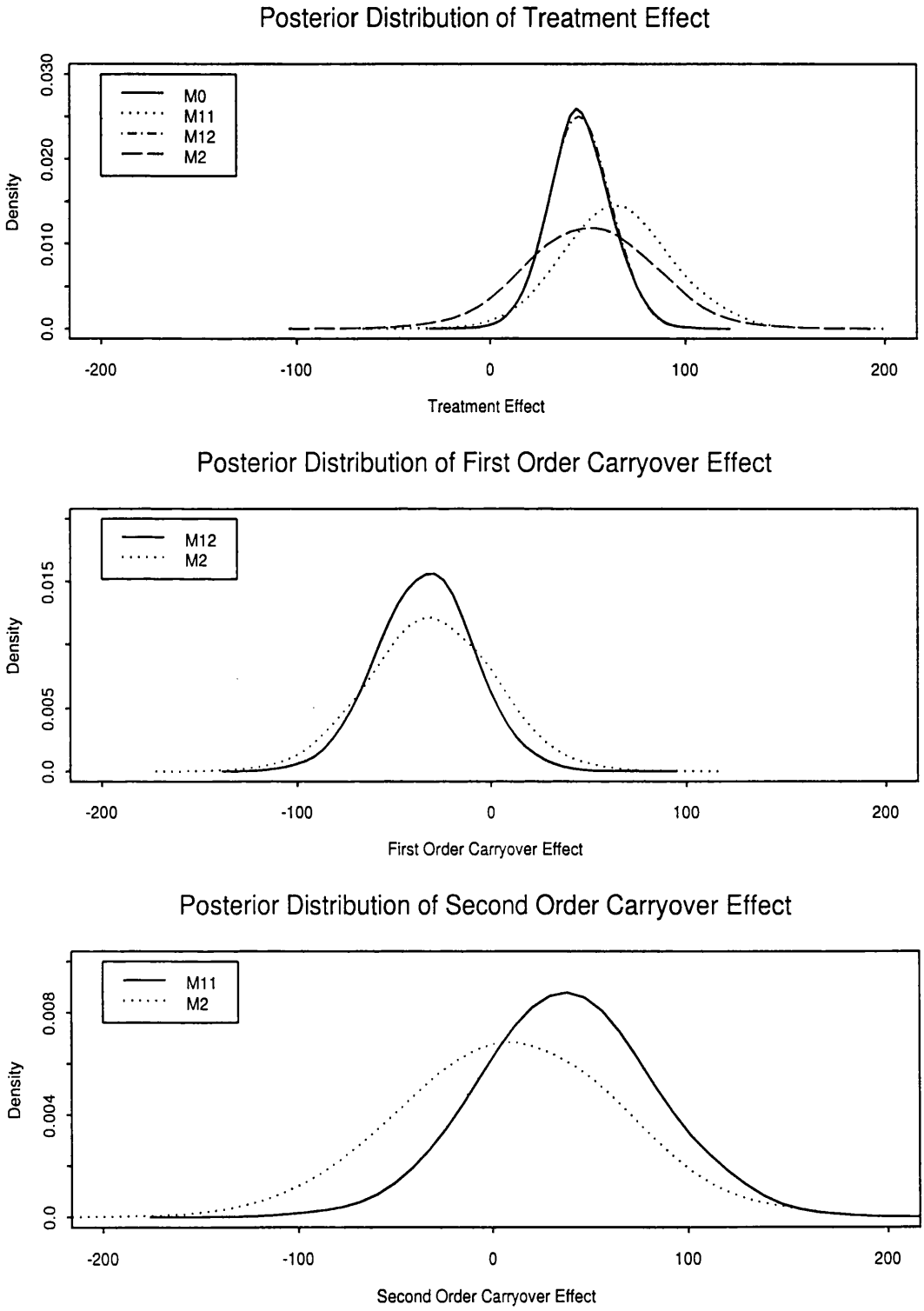


Figure 3.10: Bayesian analysis with baselines.

Table 3.8: Bayesian analysis allowing for baseline effect

	Model M2		Model M11		Model M12		Model M0	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
$2\hat{\tau}$	51.00	32.10	65.50	27.00	46.80	15.40	46.70	15.50
$2\hat{\lambda}$	8.17	56.00	37.40	44.20	——	——	——	——
$2\hat{\theta}$	-30.70	32.20	——	——	-33.70	24.80	——	——
$\hat{\sigma}_B$	64.90	17.40	64.90	17.40	65.00	17.40	64.40	16.60
$\hat{\sigma}_W$	39.70	5.16	39.40	4.95	38.80	4.89	39.40	5.09

to salbutamol by 65.50 (27.00) units, while for models M12 and M0 the corresponding figure is about 20 units lower (46.80 or 46.70) with half standard error (15.40 or 15.50). Posterior inference for the first order carry-over is not affected considerably from the presence of the second order carry-over term in the model and vice-versa (see Figure 3.10).

Finally, in this case, although the posterior distribution of the second order carry-over difference is centered around zero, the same is not true for the first order carry-over effect (see Figure (3.10)). Baseline measurements taken after administration of salbutamol give lower PEF values than baselines taken after administration of formoterol. On subjective grounds it seems that model M12 is the most coherent with the observed data, though model M0 is an equally good alternative. Which one of the two is the best choice, will be formally investigated by using appropriate model selection techniques.

3.6.3 Another use of baselines

If we take a more careful look at the data it is clear that higher baseline measurements tend to be followed by higher treatment outcomes no matter which treatment has been administered to the patient. This implies a patient trend effect affecting both baselines and outcome measurements. If this hypothesis is true, then baseline measurements convey useful information not accounted for by patient or period effects. If we denote by y_{ijk}, x_{ijk} the treatment outcome and baseline measurement respectively at the i^{th} sequence group, in the j^{th} period,

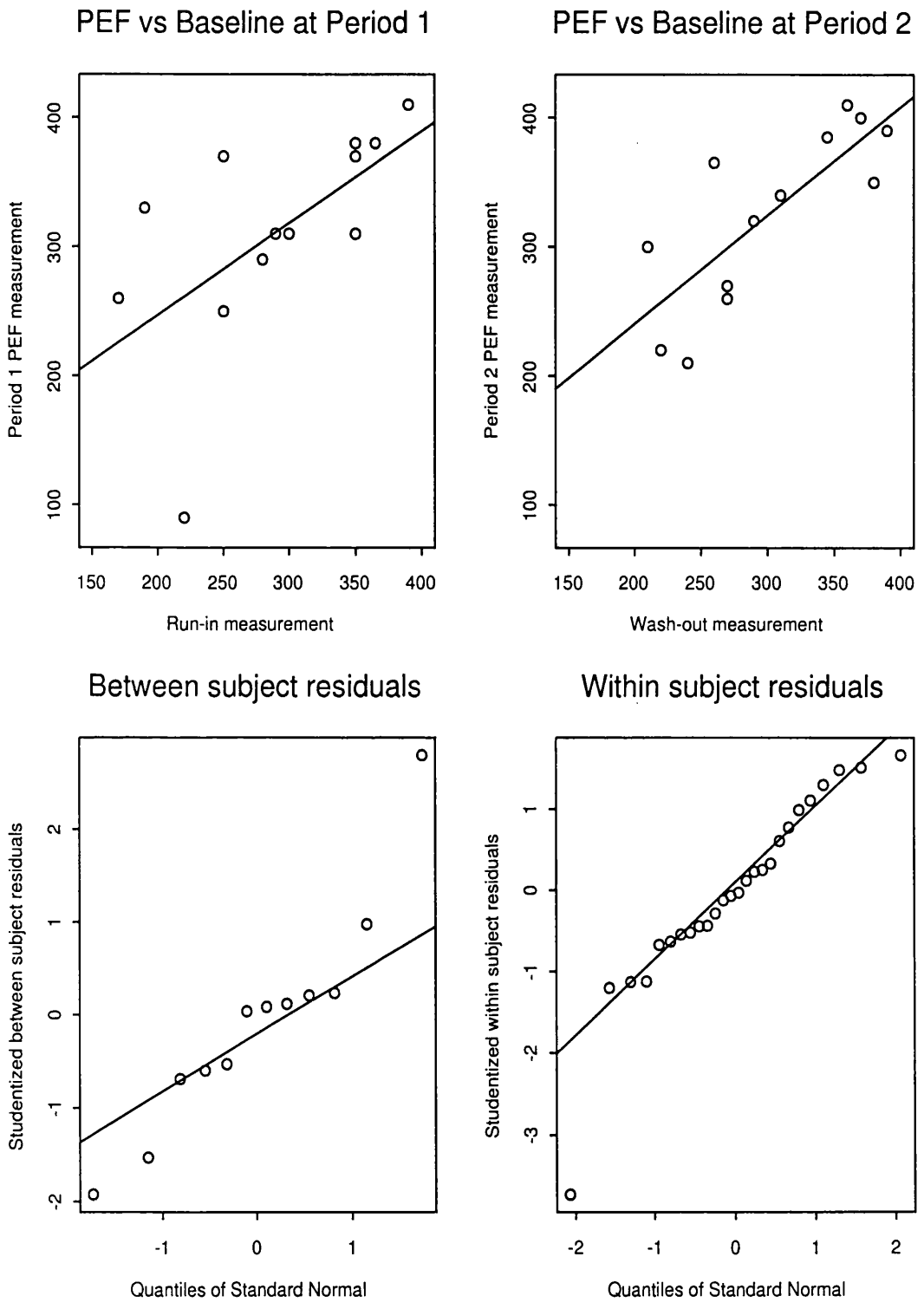


Figure 3.11: Graphical summary and model checking of the asthma trial when baselines are used as covariates

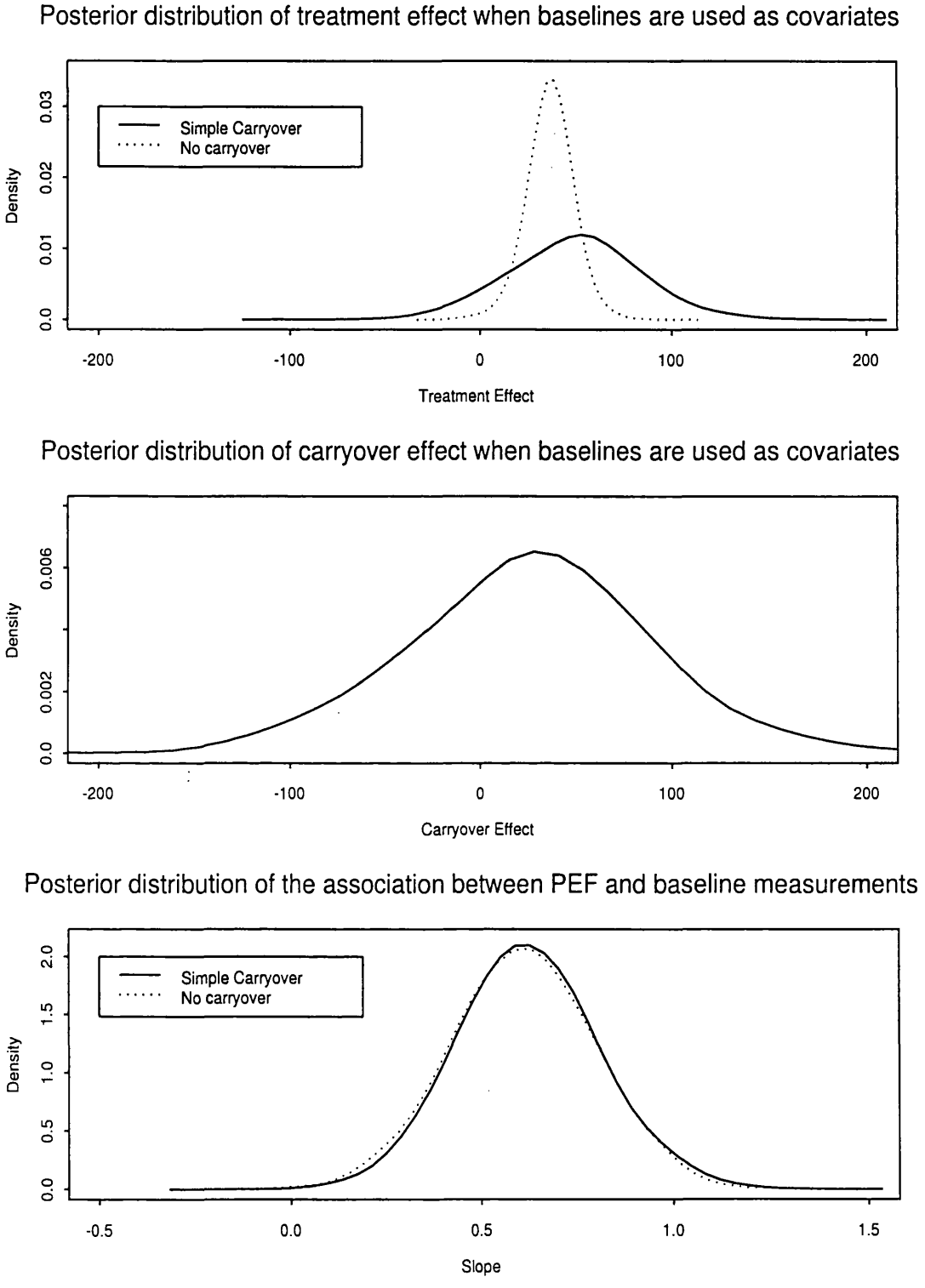


Figure 3.12: Posterior distribution of various parameters of interest of the asthma trial when baselines are used as covariates

on the k^{th} patient, then the model adopted here is:

$$y_{ijk} = \mu + s_{ik} + \pi_j + \tau_{d(i,j)} + \lambda_{d(i,j-1)} + \beta(x_{ijk} - \bar{x}_{...}) + \epsilon_{ijk} \qquad (3.38)$$

where, as usual, $s_{ik} \sim N(0, \sigma_B^2)$ and $\epsilon_{ijk} \sim N(0, \sigma_W^2)$. A graphical summary of the association between active treatment measurements and baseline readings, is displayed in Figure (3.11). Both Bayesian and Frequentist approaches, with and without carry-over terms have been considered and the results are presented in Table (3.9). Model checking graphical summaries for the Frequentist approach is provided in Figure (3.11). Running the Gibbs sampler, 15000 values were generated for each variable, but only the last 5000 ones were used for drawing inference. Posterior distributions for parameters of primary interest are displayed in Figure (3.12).

Note that in equation (3.38) baseline measurements have been standardized by subtracting their mean. This strategy is typical in regression problems since it achieves orthogonality between the standardized variable and the constant term. The ideal situation is when the estimated parameters are orthogonal to each other. This is hardly achieved when an unbalanced design is used, but in regression problems the standardization of covariates stabilizes considerably the estimation process. Parameterization issues are common in MCMC methodology as well, and usually tackled in a similar fashion. Convergence of the Markov chain to the posterior distribution is highly accelerated by using a balanced design or appropriately transforming the original parameters in the unbalanced case.

Here we have another close agreement, as far as the usefulness of the baseline measurements is concerned. In both Bayesian and Frequentist analysis baseline measurements have a strong predictive value for the response, i.e. higher baseline measurements tend to be followed by higher response outcomes (see upper half of Figure (3.11) and lower part of Figure (3.12)). Our main question is always the clinical effectiveness of the new treatment against the old one. As in all previous models considered, when carry-over is included, both the Frequentist and the Bayesian statistician will agree that no treatment difference is evident from the data, although the Bayesian will stresses that the superiority of formoterol is the more likely scenario (see upper part of Figure (3.12)). Both will agree that in the absence of carry-over from the model there is strong evidence for suggesting

formoterol in a future patient as the best treatment regime.

Table 3.9: Results of model fitting with baselines as covariates

	Frequentist viewpoint				Bayesian viewpoint			
	Model with λ		Model without λ		Model with λ		Model without λ	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
$2\hat{\tau}$	51.37	30.99	37.16	10.25	49.50	34.10	37.10	12.10
$2\hat{\lambda}$	28.39	58.83	——	——	25.60	63.30	——	——
$\hat{\beta}$	0.60	0.16	0.60	0.16	0.61	0.18	0.60	0.18
$\hat{\sigma}_B$	49.64		47.78		51.30		48.30	
$\hat{\sigma}_W$	25.23		25.21		28.00		28.80	

The practical implication from the discussion above is that the collection of baseline measurements during the course of a cross-over trial hardly alters the conclusions about treatment effect, already drawn from previous analysis where baselines were completely ignored. This statement is true no matter if baselines are considered as part of the response or fitted as a covariate. It might be the case that if baselines perform poor in explaining variability of the treatment outcome, then their inclusion in the model might increase the variance with which the treatment effect is estimated. In our example this seems to be true when carry-over term is included in the model, but not when an adequate wash-out period prohibits the consideration of such term in the model.

3.7 Covariates

In most clinical trials, either cross-over or parallel, demographic information is usually available for the patients participating in the study, such as age, sex, weight etc. Two kinds of covariates commonly met are: continuous or categorical. An example of a continuous covariate is the baseline measurement already studied in the previous section. In the formoterol/salbutamol example patients 2, 3, 6 and 11 are female, the rest being male. Patient profiles for male and female patients are displayed in Figure (3.13).

In that case the main concern is if drug acts differently on various patient sub-

groups. If it does then treatment effect should be studied separately for each level of the covariate. On the other hand if treatment effect is not related to the levels of the covariate then the inclusion of it might explain a substantial proportion of the between-subject variability, implying a reduction in the between-subject residual sum of squares. In that case more precise statements could be derived for the carry-over difference, since the investigation of its statistical significance relies upon between subject information. On the contrary treatment or other effects which are usually estimated using within subject contrasts will not be affected by the inclusion or not of the covariate in the model. The introduction of a further factor (gender) into our model generalizes the ANOVA table as follows: (Table (3.10)).

Table 3.10: Extended ANOVA table after incorporation of covariates

Source of variation	DF	Sum of Squares	Mean Square	F-value	p-value
Between Subjects					
carryover	1	335.19	335.19	0.03	0.86
gender	1	18482.80	18482.80	1.76	0.21
carryover:gender	1	1991.72	1991.72	0.18	0.67
Residuals	9	94403.75	10489.31		
Within Subjects					
period	1	984.62	984.62	1.16	0.30
treatment	1	14035.92	14035.92	16.65	0.00
period:gender	1	621.06	621.06	0.73	0.41
treatment:gender	1	49.66	49.66	0.05	0.81
Residuals	9	7583.75	824.64		

According to the analysis above there is no strong evidence that the effect of drugs on PEF measurements depends on the gender of the child. On the other hand the covariate has accounted for more than 20% of the between subject variability, but carry-over effect is still far from statistical significance. An idea about the variability explained by treatment, carry-over effect and their interactions with other terms is also summarized in Table (3.10). The model fitted here treats patient effect as the only random parameter, while the fixed parameters allowed for

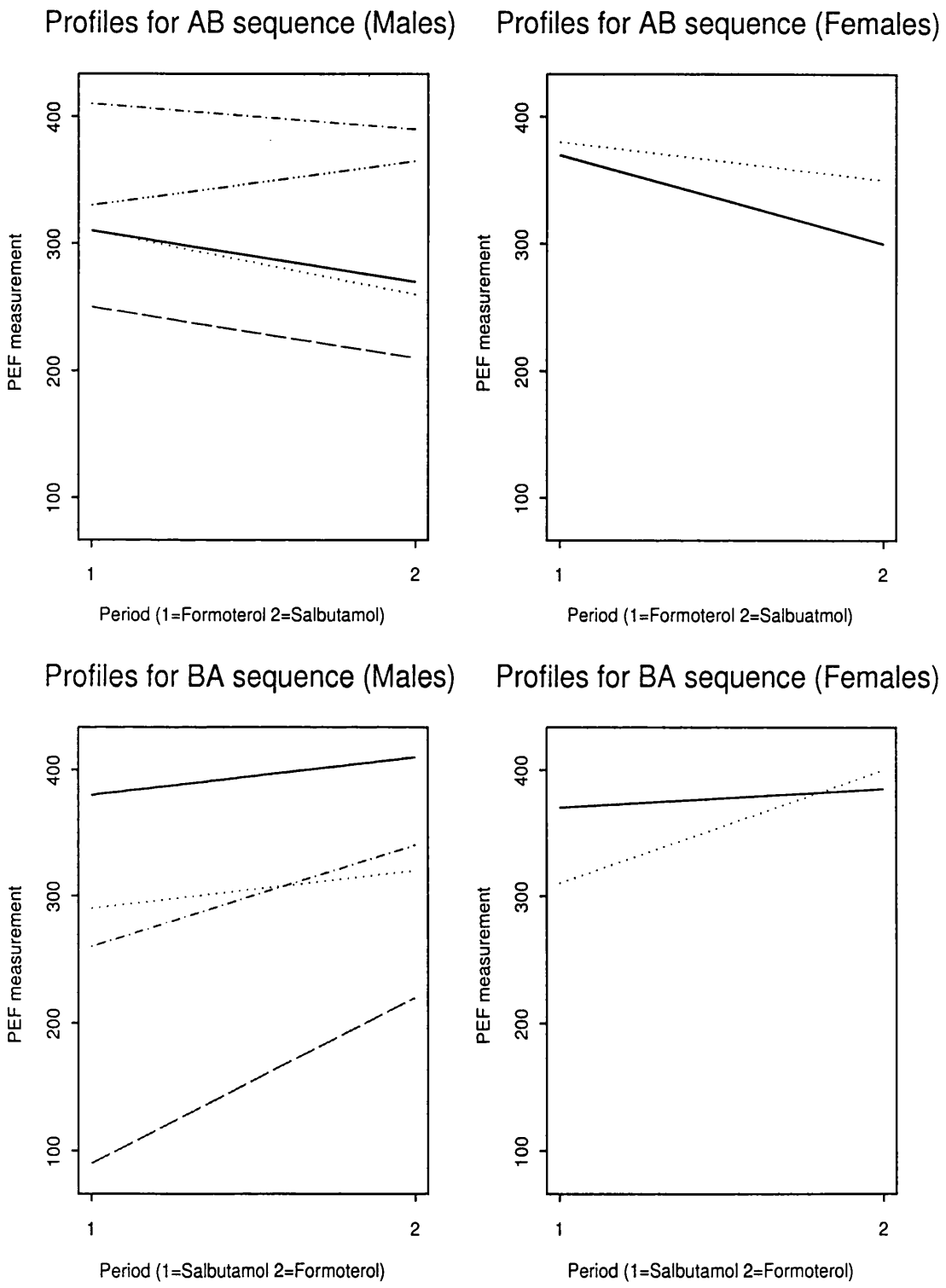


Figure 3.13: Graphical summary of the asthma trial without baselines, but "gender" included as covariate

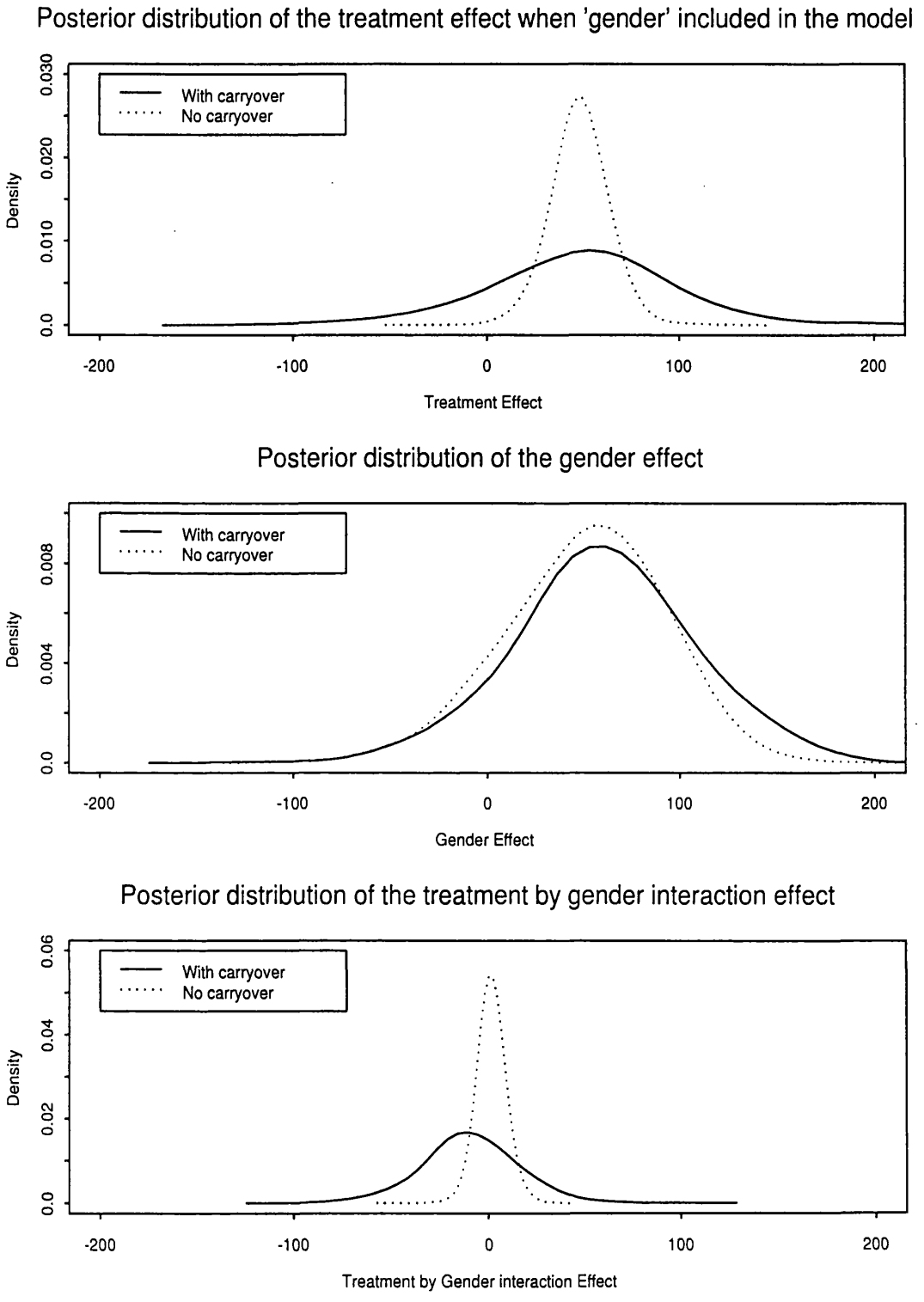


Figure 3.14: Posterior distribution of various parameters of interest of the asthma trial without baselines, but "gender" included as covariate

are: period, gender, treatment, and carry-over. Two-way interactions between "gender" and each one of the fixed effects are also considered.

One of the less discussed issues in the cross-over literature, but of high practical importance, is the treatment by patient interaction term. Modern statistical thinking (implemented in widely used commercial software packages like SAS or S+) allow the inclusion of that term in either the fixed and/or the random part of our model. This term implies not only that the mean treatment effect but also the volatility of the response are patient dependent. Irrespective of the way this interaction term is treated, an enormous number of new parameters will be introduced to describe it. A typical way to overcome estimation-related problems is via modern Bayesian techniques. More specifically both the treatment effect and the variability of the observations for each patient could be modeled as a random sample from a population distribution, characterized by a set of hyper-parameters. Finally the implication of this hierarchical modeling structure on the choice of the optimal design for running cross-over experiment have not been fully investigated.

In our example we have only 13 participants, not enough information to estimate accurately the hyper-parameters. This is true not only for the patient by treatment interaction but also for any interaction between the random component ("patient") and any fixed term. As a result considering a two-way interaction between a fixed and a random effect was ruled out for that analysis. On the contrary interactions between "gender" (the covariate of interest in this section) and any fixed parameter require only one degree of freedom to model it. It was felt that adequate information was available to include these terms. Results are summarized in Table (3.11)

There is no statistical evidence that average PEF measurements on male patients differ significantly from the female ones. Both approaches confirm that treatment effect is the same irrespective of the gender of the child (i.e negligible treatment by gender interaction, see lower part of Figure (3.14)). The estimates along with the standard errors of all the parameters presented above decrease (in absolute terms) when carry-over term is removed from the model. There seems to be a fairly close agreement concerning parameter estimation under either Frequentist

Table 3.11: Summary statistics when a covariate is included in the model

	Frequentist viewpoint		Bayesian viewpoint	
	With λ	No λ	With λ	No λ
Parameters	Estimate (SD)			
Gender	58.50 (43.60)	57.29 (39.90)	60.10 (47.00)	51.00 (41.20)
Treatment	51.00 (45.32)	48.25 (12.35)	49.00 (48.60)	48.30 (14.60)
Carryover	5.50 (87.20)		2.33 (92.90)	
Treatment:Gender	-8.00 (22.66)	1.50 (6.17)	-9.23 (25.00)	1.41 (7.29)
Carryover:Gender	-19.00 (43.60)		-21.10 (47.70)	

or Bayesian point of view.

Overall, inclusion of carry-over terms affect to an appreciable expend our inferences regarding treatment differences (see Figure (3.14)). Results are not altered by the incorporation of a covariate in the model; a similar conclusion was drawn for baselines as well.

3.8 A Non-Linear approach to the carry-over

Our modeling approach till now, is based on the assumption that carry-over and treatment effects are mathematically unrelated. The majority of the medical investigators, involved in a cross-over study, would implicitly assume that the residual effect, if it exists, is a small proportion of the treatment effect and should be modeled as such. A typical medical statistician would object to the idea of modeling carry-over effect by incorporating a non-linear term into his model, simply because on one hand it adds unnecessary complexity to the problem and on the other hand computationally can be quite difficult to be tackled by widely used statistical software. In my view, any statistician keen on modeling residual effects, should consider this approach as the only pragmatic one, which in addition provides reasonable results for the treatment effect. There is no doubt that there is limited information in estimating the unknown proportion of treatment that carries over to the next period. This causes problems in the estimation process for that non-linear term, but as soon as more patients are recruited per sequence,

this limitation is largely removed.

The model considered in this section is a slight modified version of the simple carry-over model and can be written as follows:

$$E\left(y_{ijk}\right) = \mu + \pi_j + \tau_{d(i,j)} + \tau_{d(i,j-1)}\rho \tag{3.39}$$

where notation is similar to the one used in previous sections. When baseline measurements are considered as part of the response, the mean function can be specified as follows:

$$E\left(y_{ijk}\right) = \mu + \gamma_i + \pi_j + \tau_{d(i,j)} + \tau_{d(i,j-1)}\rho + \tau_{d(i,j-2)}\rho\kappa \tag{3.40}$$

Terms introduced in this non-linear form of the simple carry-over model with baselines, are closely linked with the ones considered in the linear case. More specifically the following relations hold:

$$\theta_{d(i,j)} = \tau_{d(i,j-1)}\rho \tag{3.41}$$

$$\begin{aligned} \lambda_{d(i,j)} &= \tau_{d(i,j-2)}\rho\kappa \\ &= \theta_{d(i,j-1)}\kappa \end{aligned} \tag{3.42}$$

where $\rho,\kappa \in (0,1)$. With this parameterization residual effect dies out as time progresses, i.e. $\mid \lambda_{d(i,j)} \mid \leq \mid \theta_{d(i,j)} \mid \leq \mid \tau_{d(i,j)} \mid$. According to the above model, the proportion of treatment that carries over from period i to period $i+1$ is ρ, while that from i to $i+2$ is $\rho\kappa$. The two terms added, although it may reflect drug activity more realistically, it can lead to problems during the estimation process. Following our modeling philosophy so far, the within subject variance-covariance matrix will be of the form $\sigma^2 R(\alpha)$. Simple correlation structures (compound symmetry) will be considered in the sequel, since a limited number of repeated measurements are available per subject. Generalized least squares principle will be used to accommodate simultaneous estimation of mean and covariance parameters. Details of the estimation scheme are as follows:

- Step 1: Estimate α using a preliminary fit to our data, like the typical Ordinary Least Squares fit.

- Step 2: Using the value of α from Step 1, an estimate for the mean parameters can be derived by minimizing the following quadratic form:

$$\left(y - E(y)\right)^T R^{-1}(\alpha)\left(y - E(y)\right) \tag{3.43}$$

- Step 3: Using the estimates from Step 2, re-estimation of the covariance parameters takes place, by minimizing with respect to σ^2 and α the pseudo-likelihood function below:

$$log\left|\sigma^2 R(\alpha)\right| + (y - E(y))^T R^{-1}(\alpha)\,(y - E(y))\,/\sigma^2 \tag{3.44}$$

The final estimate of σ^2 is:

$$\hat{\sigma}^2 = \left(y - \hat{E}(y)\right)^T R^{-1}(\hat{\alpha})\left(y - \hat{E}(y)\right)/(N - p) \tag{3.45}$$

where N is the number of patients recruited in the trial, while p the number of estimated mean parameters. An estimate of the approximate covariance matrix for the mean parameters is:

$$\hat{\sigma}^2\left(X^T R^{-1}(\hat{\alpha})X\right)^{-1} \tag{3.46}$$

where X is the Nxp matrix of partial derivatives of the mean function with respect to the mean parameters, evaluated at the final estimates of these parameters.

3.8.1 Frequentist approach without baselines

In the non-linear case the treatment estimate is 53.54(39.08) in favor of the new treatment (formoterol). The carry-over effect from the first to the second active treatment period is estimated at 15.95(55.47). The proportion of treatment that carries-over to the next period is 29.79%. These results are in agreement with the linear approach, in which the carry-over effect is modeled independently of the treatment difference, since in either case both treatment and carry-over effect are statistically non-significant. It is worth noting that both estimates and their standard errors for treatment and carry-over difference are similar in the non-linear case compared to the linear approach. The mean function is described in equation (3.39). A compound symmetry covariance structure has been assumed for the repeated measurements obtained in each subject, with an estimated intra-subject correlation coefficient of 0.07. In order to obtain the above estimates for the mean parameters and variance components, the following function was optimized by following the steps 1-3 described before:

$$2nln(\sigma^2) + nln(1 - \rho_*^2) + \frac{1}{\sigma^2(1 - \rho_*^2)}(y - E(y))^T\left[I_n \otimes V^{-1}\right](y - E(y)) \tag{3.47}$$

This is simply minus twice the log-likelihood function and it was minimized using S+ routines. Relevant S+ code is given at the end of the chapter. Regarding notation, n is the number of patients recruited for the study (13 in our case), $\rho_* = \sigma_B^2/(\sigma_B^2 + \sigma_W^2)$ being the intra-subject correlation coefficient, $\sigma^2 = \sigma_B^2 + \sigma_W^2$, while

$$V^{-1} = \begin{pmatrix} 1 & -\rho_* \\ -\rho_* & 1 \end{pmatrix} \tag{3.48}$$

3.8.2 Frequentist approach with baselines

Turning now to the case where baselines are incorporated into the analysis and both first and second order carry-over terms enter into the model (M2) as described in equation (3.40), the estimated proportion of treatment effect that carries over from first treatment period to first wash-out period is identical to zero. This implies that the proportion of treatment that carries over from first treatment to second treatment period must be zero as well. Since, $\hat{\rho} = 0$ and $\hat{\kappa} = 0$, the matrix $\left(X^T R^{-1}(\hat{\alpha}) X\right)$ is non-invertible and an estimate of the standard error of the treatment effect cannot be derived. The treatment difference itself is estimated at 46.61. These findings are in close agreement with the linear approach, where both Bayesian and Frequentist approaches indicate that carry-over of any order is unlikely to be present. Due to the linear nature of the latter approach, standard errors for the parameters of interest are available in this case.

If we omit anyone of the carry-over terms but retain the other one, then treatment effect is still statistically insignificant, but the estimate of the retained residual effect is zero in either case. More specifically when the second order carry-over term is eliminated from the model (M12), then the treatment estimate is 46.61(26.12) in favor of formoterol, while in the case where the first order carry-over is omitted (M11), then the corresponding treatment estimates raises at 65.42(45.13). In conclusion, in both linear and non-linear analysis the only model that is highly supported by the data is the one with no carry-over terms. All the above conclusions are drawn under the assumption that repeated measurements within a subject are related via a compound symmetry error structure.

The function minimized for inferential purposes looks as follows:

$$3nln(\sigma_W^2) + nln(\sigma_W^2 + 4\sigma_B^2) + (y - E(y))^T \left[I_n \otimes V^{-1}\right] (y - E(y)) \qquad (3.49)$$

where $E(y)$ is described by equation (3.40) or any of its variants depending on which carry-over terms included in the model, while

$$V^{-1} = \frac{1}{\sigma_W^2}\left(I_{4x4} - \frac{\sigma_B^2}{\sigma_W^2 + 4\sigma_B^2} J_{4x4}\right) \qquad (3.50)$$

Once more n is the number of subjects recruited in our trial, I is the identity matrix and J is a square matrix having every element equal to unity. For illustrative purposes, S+ code used to optimize the function described in equation (3.49) when only the first order carry-over is included in the model, is provided at the end of the chapter.

3.8.3 Bayesian approach without baselines

In the Bayesian analysis without baselines, the only added complication is the specification of a prior distribution for ρ. Since this parameter is constrained to the interval $(0, 1)$, a natural family of distributions from which this prior could be chosen from is the $Beta(a, b)$ one. The parameters a, b can be modified to reflect opinion of medical experts, or experience gained from similar studies in the past. In practice, hardly such information exists, and a useful starting point is the Beta(1,1) distribution (or equivalently the Uniform distribution in the $(0, 1)$ interval). After a 10000 iteration burn-in, a further 5000 iterations confirm the superiority of the new compound compared to the old one. According to this analysis a typical user of formoterol will have his PEF measurement raised by 51.30(14.40) units compared to salbutamol. The 95% HPD for treatment effect is (21.70, 79.50). The point estimate of the treatment effect from the non-linear analysis is in close agreement to the one derived under the linear case, where carry-over and treatment terms were unrelated. Posterior kernel density estimates of the treatment difference under both linear and non-linear approaches are presented in Figure (3.15). The estimated standard error of the treatment effect is about three times smaller in the non-linear analysis when compared to the corresponding figure in the linear case. This implies that the posterior

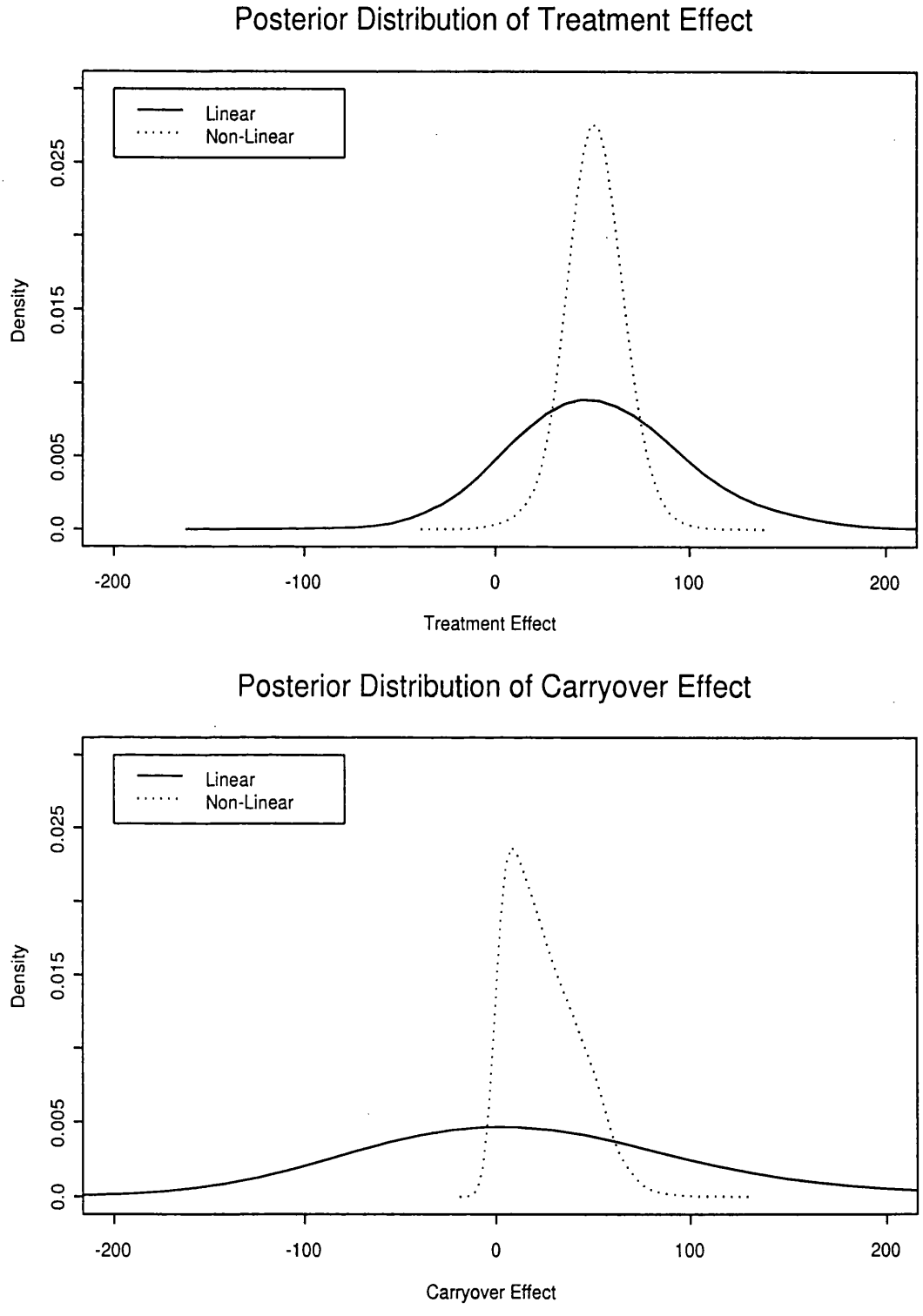


Figure 3.15: Posterior distribution of various parameters of interest of the asthma trial without baselines under the simple carry-over model

probability of the treatment effect lying in a symmetric interval around zero is far less in the non-linear case compared to the linear one. Both analysis though, indicate clearly that formoterol is the appropriate therapy for asthma.
It is worth noting that the estimated carry-over effect in the non-linear case is 24.2(17.7). The posterior density of the carry-over difference is slightly skewed to the left in the non-linear case, contrary to the density resulted from the linear analysis which looks symmetric. This is because carry-over effect is calculated as the product of the treatment effect with the proportion of drug remaining in the body from the previous treatment assignment. The point estimates of the carry-over effect closely agree in the linear and non-linear analysis, though the standard error in the linear analysis is about four times higher when compared to the non-linear case (see Figure (3.15). Both analysis agree that presence of carry-over is highly unlikely.

3.8.4 Bayesian approach with baselines

Similar conclusions were drawn when baseline measurements were incorporated as part of the response into the analysis. As before, $Beta(1,1)$, has been chosen as the prior distribution for these extra parameters. Analysts might be tempted to consider informative prior inputs, although clinical justification for these choices should be provided.
The model that includes first and second order carry-over terms (M2), gives a treatment estimate of 44.30(17.30) in favour of formoterol. Only 28.3% of the treatment effect persists from first active to first wash-out period, while the corresponding figure from first active to second active treatment period is 14.7%. Similar results are derived when the model under which only one carry-over term, that from first active to first baseline period is allowed for (M12). The estimated treatment effect slightly lowers to 41.30(16.80), but a similar proportion of active treatment persists to the next period, 28.4%. Finally, the irrational model (M11) where second order carry-over is fitted in the absence of the first one, gives an inflated but significant treatment estimate: 63.80(22.70). This estimate may reflect not only real treatment difference but also first or higher order residual effects, which have been eliminated from our model. Posterior distribution of

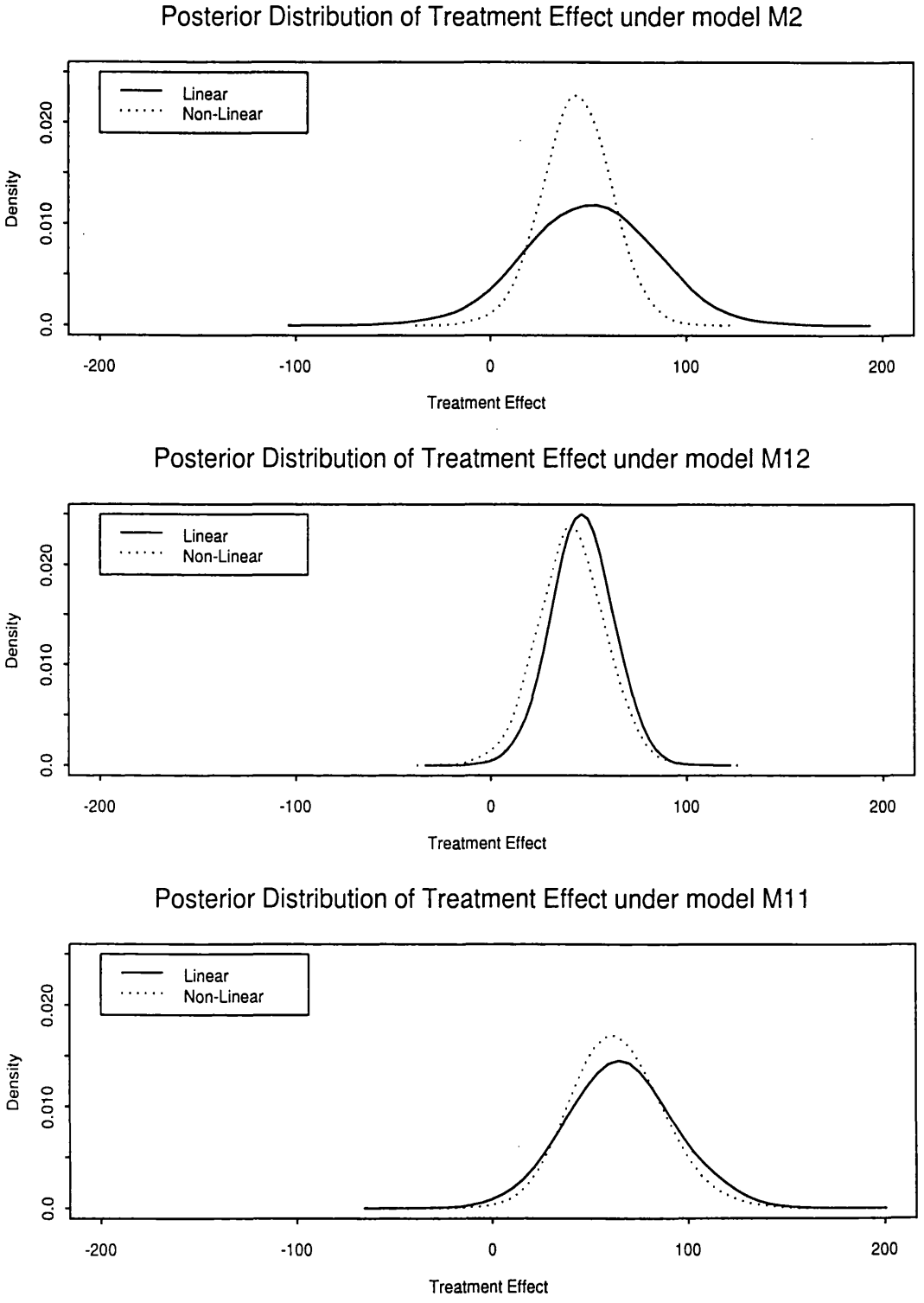


Figure 3.16: Posterior distribution of treatment effect of the asthma trial with baselines under models M2, M12 and M11

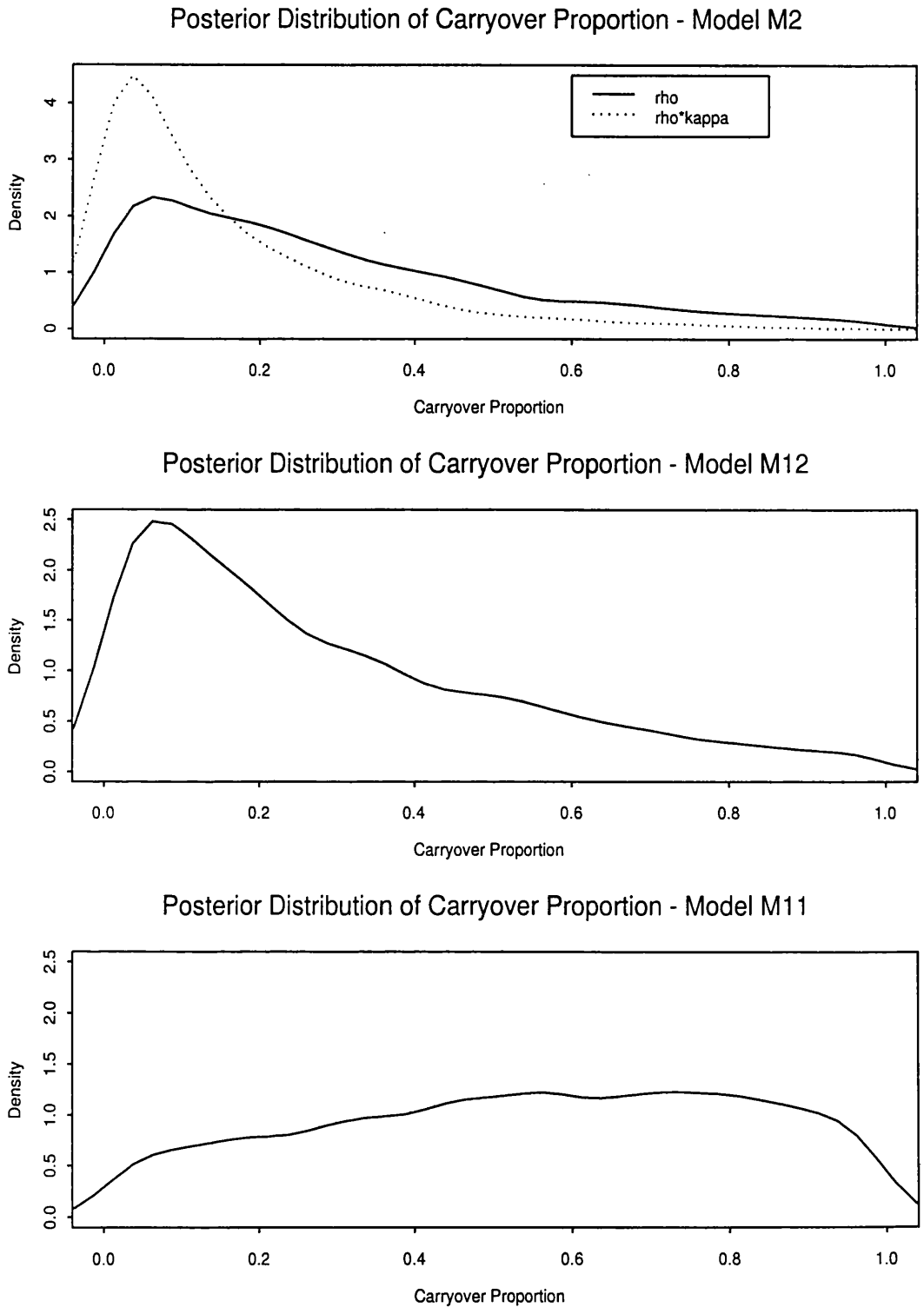


Figure 3.17: Posterior distribution of carry-over proportion of the asthma trial with baselines under models M2, M12 and M11

carry-over proportion under all models considered is displayed in Figure (3.17). Note that the linear and non-linear Bayesian analysis provide comparable results concerning estimation of the treatment difference (see Figure (3.16)), as long as the same assumptions are made for the carry-over effect. The effectiveness of the new treatment is unquestionable. The posterior distribution of carry-over effects of any order, show clearly that inclusion of such terms in the model is unnecessary. This is true in either linear or non-linear approach. It is worth noting that posterior densities of carry-over terms are slightly skewed to the left in the non-linear case, contrary to the linear approach where a rather symmetric shape is observed.

In conclusion the Frequentist analysis (linear or non-linear) of this cross-over trial strongly rejects the inclusion of any residual terms, while treatment difference is masked when carry-over terms are included into the model. The Bayesian approach (linear or non-linear) supports the superiority of the new treatment regime under different carry-over schemes. Note that for the Frequentist approach, a constrained non-linear optimization problem was solved using S+ routines.

3.8.5 Model checking

A more formal way is normally required for the selection of the best among competing models. Obviously in the class of models fitted, some are nested within others, which implies that a likelihood ratio approach for choosing between them is valid. But if a selection is required between a linear an a non-linear model then the final decision should be based on a criterion that rewards a good fit but punishes for model complexity. One such criterion is the Akaike's Information Criterion (AIC) defined as follows:

$$AIC = -2ln(\text{likelihood}) + 2(\text{number of estimated parameters})$$

Although AIC has received some criticism as a model selection tool (especially in the time series literature), it is still the most popular criterion used by practitioners. In a Bayesian analysis, the posterior distribution of AIC is evaluated and model choice is based on a summary statistic of that distribution. In what follows the posterior mean of AIC calculated from the last 1000 MCMC runs is

compared among competing models and summarized in Table (3.12).

Table 3.12: Posterior Mean of AIC

	M2	M11	M12	M0
Linear	-238.705	-240.433	-239.853	-242.096
Non-linear	-238.693	-240.115	-240.876	

The model which has the lowest value of AIC is selected as the best one. Following that principle the model with no carry-over terms (M0) is the preferable one. For model M12 the non-linear fit gives slightly better results when compared to its linear counterpart. The reverse argument is true for models M11 and M2. In the linear case AIC gives the following model ordering: $M0 \prec M11 \prec M12 \prec M2$, while in the non-linear context we get the more sound result $M0 \prec M12 \prec M11 \prec M2$. The operator $\prec$ means that the model on the left hand side provides a better fit compared to the right hand side one. In conclusion the non-linear approach gives sensible results for treatment effect irrespective of the type of residual term (if any) fitted in the model. Moreover it tends to provide accurate outcome during the model selection process.

3.9 Conclusions

In the 2x2 cross-over trial, the performance of various treatment estimators (CROS, PAR, TS) has been studied in some detail. In summary CROS should be the preferable treatment estimator, no matter if carry-over is included or not in the model. The alternative (TS procedure), where CROS is selected with probability p and PAR with probability $1-p$, should be avoided, because it has lower power and higher MSE when compared to CROS. If the analyst insists in using the two stage procedure then one can replace the original scheme with a new one in which, the sizes of the tests for carry-over and treatment difference are set so that the overall size of the procedure is 5%. Unfortunately the improved plan does not perform better when compared to CROS in terms of power, or MSE. This investigation leads to the conclusion that TS procedure should be gradually abandoned by the analyst of the cross-over experiment.

Based on a representative example of a cross-over trial in asthma, both Bayesian and Frequentist analysis suggest that carry-over is very unlikely to be present in a well-planned trial. The use of baselines or covariates hardly affect our conclusions about treatment difference, although their incorporation might increase precision for inferences about carry-over effect. In conclusion magnitude and standard error of treatment difference are affected by the presence of carry-over terms. Treatment effect tends to be statistically unimportant when carry-over is incorporated in the final model, while in the absence of it treatment difference is highly significant. The trialist should carefully investigate the potential for pharmacological carry-over and choose the appropriate length of the wash-out period for eliminating such an effect. Once this precaution has been taken the analysis model should not include carry-over terms of any kind.

3.10 BUGS and S+ code used for the derivation of the results in this chapter

3.10.1 BUGS code for the linear Bayesian analysis without baselines - subsection 3.5.1

Bayesian analysis of simple carry-over model without baselines

model pef1;

const

N=13, *number of patients*

P=2; *number of periods*

var

pef[N,P], *response matrix*

carryover, *parameter for carryover*

carryover.effect, *real carryover effect*

carry[N,P], *carry-over matrix*

treatment, *parameter for treatment*

treatment.effect, *real treatment effect*

treat[N,P], *treatment matrix*

pi, *period effect*

period[N,P], *period matrix*

intercept, *intercept of the model*

mu[N,P], *mean of the response*

subject[N], *random subject effect*

precision.within, *within patient precision*

precision.between, *between patient precision*

sigma.within *within patient standard deviation*

sigma.between; *between patient standard deviation*

Next we simply read data and set initial values for the parameters in our model

```
data period, treat, carry, pef in "agsc.dat";
inits in "agsc.in";
```

Priors for the parameters in our model

```
{
intercept ~ dnorm(0,1.0E-06); pi ~ dnorm(0,1.0E-06);
treatment ~ dnorm(0,1.0E-06); carryover ~ dnorm(0,1.0E-06);
precision.within ~ dgamma(1.0E-06,1.0E-06);
precision.between ~ dgamma(1.0E-06,1.0E-06);
treatment.effect <- 2*treatment; carryover.effect <- 2*carryover;
sigma.within <- sqrt(1/precision.within);
sigma.between <- sqrt(1/precision.between);
```

Next we simply define our model

```
for (i in 1:N) {
    subject[i]~dnorm(0,precision.between);
    for (j in 1:P) {
    pef[i,j]~dnorm(mu[i,j],precision.within);
    mu[i,j]<-intercept+pi*period[i,j]+treatment*treat[i,j]+
    carryover*carry[i,j]+subject[i]; } }
}
```

3.10.2 BUGS code for the linear Bayesian analysis with Baselines as part of the response (model M2) - subsection 3.6.2

```
model pef1;
const
   N=13,  number of patients
   P=4;  number of periods
var
   pef[N,P],
   theta, first.carry, carry1[N,P],
   lambda, second.carry, carry2[N,P],
   tau, treatment, treat[N,P],
   pi1,pi2,pi3,pi4,
   period1[N,P], period2[N,P], period3[N,P], period4[N,P],
   sequence, seq[N,P], intercept, mu[N,P], subject[N],
   precision.within, precision.between, sigma.within, sigma.between;

Reading data and initial values for the Gibbs sampler
data period1, period2, period3, period4,
treat, carry1, carry2, seq, pef in "nagm2.dat";
inits in "nagm2.in";

Defining priors
{
intercept ~ dnorm(0,1.0E-06);
pi1 ~ dnorm(0,1.0E-06); pi2 ~ dnorm(0,1.0E-06); pi3~dnorm(0,1.0E-06);
treatment ~ dnorm(0,1.0E-06); first.carry~dnorm(0,1.0E-06);
second.carry ~ dnorm(0,1.0E-06); sequence~dnorm(0,1.0E-06);
precision.within ~ dgamma(1.0E-06,1.0E-06);
precision.between ~ dgamma(1.0E-06,1.0E-06);
theta<-2*first.carry; lambda<-2*second.carry; tau<-2*treatment;
```

```
sigma.within<-1/precision.within;
sigma.between<-1/precision.between;

Model definition
for (i in 1:N) {
    subject[i]~dnorm(0,precision.between);
    for (j in 1:P) {
    pef[i,j] ~ dnorm(mu[i,j],precision.within);
    mu[i,j]<-intercept+subject[i]+sequence*seq[i,j]+
    pi1*period1[i,j]+pi2*period2[i,j]+pi3*period3[i,j]+
    (pi1+pi2+pi3)*period4[i,j]+
    treatment*treat[i,j]+first.carry*carry1[i,j]+second.carry*carry2[i,j]; }}
}
```

Similar code has been applied for fitting models M11 and M12.

3.10.3 S+ code for the Non-linear Frequentist analysis without baselines -subsection 3.8.1

```
Reading the data-set
data1<-read.table("agsc.dat"); data1<-data.frame(data1)
data<-as.matrix(data1)
period.data<-data[, 1 : 2]; period.col<-matrix(period.data,2*nrow(data),1)
treatment.data<-data[, 3 : 4]; treatment.col<-matrix(treatment.data,2*nrow(data),1)
carryover.data<-data[, 5 : 6]; carryover.col<-matrix(carryover.data,2*nrow(data),1)
response.data<-data[, 7 : 8]; response.col<-matrix(response.data,2*nrow(data),1)
no.param<-4; no.times<-2; no.subj<-nrow(data)
ones<-matrix(1,26,1); epsilon<-1.0E-06
var.parameters.old<-c(10,0)
mean.parameters.old<-c(1,1,1,0.5)

Mean Function
mean.estimation<-function(mean.vector)
{
```

```
    x<-mean.vector[1]; y<-mean.vector[2]
    z<-mean.vector[3]; w<-mean.vector[4]
    inv.v1<-(1/(1-ρ^2))*matrix(c(1,-ρ,-ρ,1),2,2)
    inv.sigma1<-kronecker(diag(nrow(data)),inv.v1)
    mean1<-x*ones+y*period.col+z*treatment.col+z*w*carryover.col
    likelihood1.value<-t(response.col-mean1)%*%inv.sigma1%*%(response.col-mean1)
    return(likelihood1.value)
}

Variance function
var.estimation<-function(var.vector)
{
    a<-var.vector[1]
    b<-var.vector[2]
    inv.v2<-(1/(1-b^2))*matrix(c(1,-b,-b,1),2,2)
    inv.sigma2<-kronecker(diag(nrow(data)),inv.v2)
    mean2<-mu*ones+period*period.col+tau*treatment.col+tau*theta*carryover.col
    likelihood2.value<-2*nrow(data)*log(a)+
    nrow(data)*log(1-b^2)+
    (1/a)*(t(response.col-mean2)%*%inv.sigma2%*%(response.col-mean2))
    return(likelihood2.value)
}

Here is where the estimation process starts
ρ <-var.parameters.old[2]
meanlikelihood.old<-mean.estimation(mean.parameters.old)
mean.nonlinear.list<-nlminb(start=mean.parameters.old, objective=mean.estimation,
lower=c(-Inf,-Inf,-Inf,0), upper=c( Inf, Inf, Inf,1))
mean.parameters.new<-mean.nonlinear.list$parameters
meanlikelihood.new<-mean.nonlinear.list$objective

while (abs(meanlikelihood.new-meanlikelihood.old)>epsilon)
```

```
{
    mu<-mean.parameters.new[1]; period<-mean.parameters.new[2]
    tau<-mean.parameters.new[3]; theta<-mean.parameters.new[4]
    var.nonlinear.list<-nlminb(start=var.parameters.old,
    objective=var.estimation, lower=c(0,-1), upper=c(Inf,1))
    var.parameters.new<-var.nonlinear.list$parameters
    mean.parameters.old<-mean.parameters.new
    var.parameters.old<-var.parameters.new
    meanlikelihood.old<-meanlikelihood.new
    ρ <-var.parameters.old[2]
    mean.nonlinear.list<-nlminb(start=mean.parameters.old,
    objective=mean.estimation, lower=c(-Inf,-Inf,-Inf,0), upper=c( Inf, Inf, Inf,1))
    mean.parameters.new<-mean.nonlinear.list$parameters
    meanlikelihood.new<-mean.nonlinear.list$objective
}
```

Final estimation steps

```
μ.final<-mean.parameters.new[1]
period.final<-mean.parameters.new[2]
τ.final<-mean.parameters.new[3]
θ.final<-mean.parameters.new[4]
λ.final<-τ.final*θ.final
sigtot.final<-var.parameters.new[1]
ρ.final<-var.parameters.new[2]
mean.final<-μ.final*ones+period.final*period.col+
τ.final*treatment.col+λ.final*carryover.col
inv.varcov.ind<-(1/(1-ρ.final^2))*(1/sigtot.final)*matrix(c(1,-ρ.final,-ρ.final,1),2,2)
inv.sigma.all<-kronecker(diag(nrow(data)),inv.varcov.ind)
derivative.matrix<-cbind(ones,period.col,treatment.col+
θ.final*carryover.col,τ.final*carryover.col)
corr.fixed.effects<-solve(t(derivative.matrix)%*%inv.sigma.all%*%derivative.matrix)
τ.se<-sqroot(t(c(0,0,1,0))%*%corr.fixed.effects%*%c(0,0,1,0))
λ.se<-sqroot(t(c(0,0,θ.final,τ.final))%*% corr.fixed.effects%*% c(0,0,θ.final,τ.final))
```

ll.τ <-τ.final-qt(0.975,(no.times*no.subj-no.param))*τ.se

ul.τ <-τ.final+qt(0.975,(no.times*no.subj-no.param))*τ.se

ll.λ <-λ.final-qt(0.975,(no.times*no.subj-no.param))*λ.se

ul.λ <-λ.final+qt(0.975,(no.times*no.subj-no.param))*λ.se

3.10.4 S+ code for the Non-linear Frequentist analysis with baselines (model M12) -subsection 3.8.2

Reading the data-set

```
data1<-read.table("nagm2.dat"); data1<-data.frame(data1)
data<-as.matrix(data1)
period1.data<-data[, 1 : 4]; period2.data<-data[, 5 : 8]
period3.data<-data[, 9 : 12]; period4.data<-data[, 13 : 16]
treatment.data<-data[, 17 : 20]
carry1.data<-data[, 21 : 24]; carry2.data<-data[, 25 : 28]
sequence.data<-data[, 29 : 32]; response.data<-data[, 33 : 36]
period1.col<-matrix(period1.data,4*nrow(data),1)
period2.col<-matrix(period2.data,4*nrow(data),1)
period3.col<-matrix(period3.data,4*nrow(data),1)
period4.col<-matrix(period4.data,4*nrow(data),1)
treatment.col<-matrix(treatment.data,4*nrow(data),1)
carry1.col<-matrix(carry1.data,4*nrow(data),1)
carry2.col<-matrix(carry2.data,4*nrow(data),1)
sequence.col<-matrix(sequence.data,4*nrow(data),1)
response.col<-matrix(response.data,4*nrow(data),1)
no.param<-7; no.times<-4; no.subj<-nrow(data)
ones<-matrix(1,4*nrow(data),1); epsilon<-1.0E-06
mean.parameters.old<-c(1,1,1,1,1,1,0.5)
var.parameters.old<-c(10,10)
```

Mean function

```
mean.estimation<-function(mean.vector)
{
```

```
    x1<-mean.vector[1]; x2<-mean.vector[2]
    x3<-mean.vector[3]; x4<-mean.vector[4]
    x5<-mean.vector[5]; x6<-mean.vector[6]
    x7<-mean.vector[7]
    v1<-ρ1*diag(no.times)+ρ2*matrix(1,no.times,no.times)
    inv.v1<-(1/ρ1)*(diag(no.times)-(ρ2/(ρ1+4*ρ2))*matrix(1,no.times,no.times))
    inv.sigma1<-kronecker(diag(nrow(data)),inv.v1)
    mean1<-x1*ones+x2*sequence.col+
    x3*period1.col+x4*period2.col+x5*period3.col+(x3+x4+x5)*period4.col+
    x6*treatment.col+x6*x7*carry1.col
    likelihood1.value<-t(response.col-mean1)%*%inv.sigma1%*%(response.col-mean1)
    return(likelihood1.value)
}
```

Variance function

```
var.estimation<-function(var.vector)
{
    a<-var.vector[1]; b<-var.vector[2]
    v2<-a*diag(no.times)+b*matrix(1,no.times,no.times)
    inv.v2<-(1/a)*(diag(no.times)-(b/(a+4*b))*matrix(1,no.times,no.times))
    inv.sigma2<-kronecker(diag(nrow(data)),inv.v2)
    mean2<-μ*ones+sequence*sequence.col+
    period1*period1.col+period2*period2.col+period3*period3.col+
    (period1+period2+period3)*period4.col+
    τ*treatment.col+τ*carry1*carry1.col
    likelihood2.value<-3*nrow(data)*log(a)+nrow(data)*log(a+4*b)+
    (t(response.col-mean2)%*%inv.sigma2%*%(response.col-mean2))
    return(likelihood2.value)
}
```

Estiamtion process

```
ρ1 <-var.parameters.old[1]
```

```
ρ2 <-var.parameters.old[2]
meanlikelihood.old<-mean.estimation(mean.parameters.old)
mean.nonlinear.list<-nlminb(start=mean.parameters.old, objective=mean.estimation,
lower=c(-Inf,-Inf,-Inf,-Inf,-Inf,-Inf,0), upper=c( Inf, Inf, Inf, Inf, Inf,Inf,1))
mean.parameters.new<-mean.nonlinear.list$parameters
meanlikelihood.new<-mean.nonlinear.list$objective

while (abs(meanlikelihood.new-meanlikelihood.old)>epsilon)
{
    μ <-mean.parameters.new[1]; sequence <-mean.parameters.new[2]
    period1<-mean.parameters.new[3]; period2<-mean.parameters.new[4]
    period3 <-mean.parameters.new[5]
    τ <-mean.parameters.new[6]; carry1<-mean.parameters.new[7]
    var.nonlinear.list<-nlminb(start=var.parameters.old, objective=var.estimation,
    lower=c(0,0), upper=c(Inf,Inf))
    var.parameters.new<-var.nonlinear.list$parameters
    mean.parameters.old<-mean.parameters.new
    var.parameters.old<-var.parameters.new
    meanlikelihood.old<-meanlikelihood.new
    ρ1 <-var.parameters.old[1]; ρ2 <-var.parameters.old[2]
    mean.nonlinear.list<-nlminb(start=mean.parameters.old, objective=mean.estimation,
    lower=c(-Inf,-Inf,-Inf,-Inf,-Inf,-Inf,0), upper=c( Inf, Inf, Inf, Inf, Inf, Inf,1))
    mean.parameters.new<-mean.nonlinear.list$parameters
    meanlikelihood.new<-mean.nonlinear.list$objective
}

μ.final<-mean.parameters.new[1]
sequence.final<-mean.parameters.new[2]
period1.final<-mean.parameters.new[3]
period2.final<-mean.parameters.new[4]
period3.final<-mean.parameters.new[5]
τ.final<-mean.parameters.new[6]
```

carry1.final<-mean.parameters.new[7]

θ.final<-τ.final*carry1.final

ρ_1.final<-var.parameters.new[1]

ρ_2.final<-var.parameters.new[2]

mean.final<-μ.final*ones+sequence.final*sequence.col+
period1.final*period1.col+period2.final*period2.col+period3.final*period3.col+
(period1.final+period2.final+period3.final)*period4.col+
τ.final*treatment.col+τ.final*carry1.final*carry1.col

inv.varcov.ind<-(1/ρ_1.final)*(diag(no.times)-
(ρ_2.final/(ρ_1.final+4*ρ_2.final))*matrix(1,no.times,no.times))

inv.sigma.all<-kronecker(diag(nrow(data)),inv.varcov.ind)
derivative.matrix<-cbind(ones,sequence.col,
period1.col+period4.col,period2.col+period4.col,period3.col+period4.col,
treatment.col+carry1.final*carry1.col,τ.final*carry1.col)

corr.fixed.effects<-solve(t(derivative.matrix)%*% inv.sigma.all%*% derivative.matrix)
τ.se<-sqroot(t(c(0,0,0,0,0,1,0))%*%corr.fixed.effects%*%c(0,0,0,0,0,1,0))
θ.se<-sqroot(t(c(0,0,0,0,0,carry1.final,τ.final))%*%corr.fixed.effects%*%
c(0,0,0,0,0,carry1.final,τ.final))
ll.τ <-τ.final-qt(0.975,(no.times*no.subj-no.param))*τ.se
ul.τ <-τ.final+qt(0.975,(no.times*no.subj-no.param))*τ.se
ll.θ <-θ.final-qt(0.975,(no.times*no.subj-no.param))*θ.se
ul.θ <-θ.final+qt(0.975,(no.times*no.subj-no.param))*θ.se
Similar code has been written for fitting models M2 and M11.

3.10.5 BUGS code for the non-linear Bayesian analysis without baselines - subsection 3.8.3

model pef1;
const

```
    N=13, P=2;

Defining parameters
var
    pef[N,P],
    rho, carryover.effect, carry[N,P],
    treatment, treatment.effect, treat[N,P],
    pi, period[N,P],
    intercept, mu[N,P], subject[N],
    precision.within, precision.between, sigma.within, sigma.between;

Reading data-set
data period, treat, carry, pef in "agsc.dat";
inits in "coragsc.in";

Defining priors
{
intercept~dnorm(0,1.0E-06); pi~dnorm(0,1.0E-06);
treatment~dnorm(0,1.0E-06); rho~dbeta(1,1);
precision.within~dgamma(1.0E-06,1.0E-06);
precision.between~dgamma(1.0E-06,1.0E-06);
treatment.effect<-2*treatment; carryover.effect<-2*treatment*rho;
sigma.within<-1/precision.within;
sigma.between<-1/precision.between;

The model
for (i in 1:N) {
    subject[i]~dnorm(0,precision.between);
    for (j in 1:P)
    pef[i,j]~dnorm(mu[i,j],precision.within);
    mu[i,j]<-intercept+pi*period[i,j]+treatment*treat[i,j]+
    treatment*rho*carry[i,j]+subject[i]; } }
```

}

3.10.6 BUGS code for the non-linear Bayesian analysis with baselines (model M2) - subsection 3.8.4

```
model pef1;
const
    N=13, P=4;

Defining model parameters
var
    pef[N,P],
    rho1, first.carry, carry1[N,P],
    rho2, kappa, second.carry, carry2[N,P],
    treatment.effect, treatment, treat[N,P],
    pi1,pi2,pi3, period1[N,P], period2[N,P], period3[N,P], period4[N,P],
    sequence, seq[N,P],
    intercept, mu[N,P], subject[N],
    precision.within, precision.between,
    sigma.within, sigma.between;

Reading data-set
data period1, period2, period3, period4,
treat, carry1, carry2, seq, pef in "nagm2.dat";
inits in "coragm2.in";

Defining priors
{
intercept~dnorm(0,1.0E-06);
pi1~dnorm(0,1.0E-06); pi2~dnorm(0,1.0E-06); pi3~dnorm(0,1.0E-06);
rho1~dbeta(1,1); kappa~dbeta(1,1); rho2<-rho1*kappa;
treatment~dnorm(0,1.0E-06); sequence~dnorm(0,1.0E-06);
precision.within~dgamma(1.0E-06,1.0E-06);
```

```
precision.between~dgamma(1.0E-06,1.0E-06);
treatment.effect<-2*treatment;
first.carry<-2*treatment*rho1; second.carry<-2*treatment*rho2;
sigma.within<-1/precision.within;
sigma.between<-1/precision.between;

The model
for (i in 1:N) {

    subject[i]~dnorm(0,precision.between);
    for (j in 1:P) {
    pef[i,j]~dnorm(mu[i,j],precision.within);
    mu[i,j]<-intercept+sequence*seq[i,j]+
    pi1*period1[i,j]+pi2*period2[i,j]+pi3*period3[i,j]+
    (pi1+pi2+pi3)*period4[i,j]+
    treatment*treat[i,j]+
    treatment*rho1*carry1[i,j]+
    treatment*rho1*kappa*carry2[i,j]+
    subject[i]; } }
}
```

Similar code has been applied for fitting models M12 and M11

Chapter 4

Multi-period, multi-sequence designs for two treatments

4.1 General considerations

When a clinical trial is conducted, the number of periods used is usually chosen to be equal the number of treatments the trialist is prepared to compare. This need stems from the limited time horizon within which the trial must be completed, but also from limited financial resources. A direct implication of that restriction is that when only two treatments are compared, no more than two periods will be used.

There are a number of advantages when higher-order designs are used for comparing two treatments. By higher-order we mean that in the clinical trial plan either multi-period designs are allowed for, or more than two sequence groups are used, or both. To begin with, better insight for the treatment difference can be gained for each patient, if a multi-period trial is preferred to a conventional 2x2 solution. Imagine for the moment that the treatment sequence ABAB... is administered to a patient. In that case the treatment difference A-B can be evaluated more than once, and assuming negligible time trends, a more accurate patient-based estimate for the treatment effect is possible. By combining these individual-based treatment estimates a better overall picture for the superiority or not of the newly proposed treatment (A) compared to the old one (B) can be obtained. Furthermore, if subject effect is modeled as a fixed parameter, multi-period designs offer

the opportunity to estimate carry-over differences using within subject contrasts. On the contrary in the 2x2 case carry-over effect is estimated by utilizing only between subject information. Finally in multi-sequence, multi-period cross-over clinical trials treatment by period interaction and carry-over difference are separately estimable. In the 2x2 case these effects are intrinsically aliased.

Another important issue, which can be investigated using the multi-period designs, is the statistical significance or not of the treatment-by-carryover interaction. Till now it has been assumed that residual effects depends on the previous treatment and not at all on the current one. A well-known alternative to that scenario (see Fleiss [17]) proposes a scheme under which carry-over from A to A might be negligible, but the one from A to B might be present. On pharmacological grounds this might be the case if the two compounds have similar but not identical therapeutic activity and react to each other. In clinical trials (cross-over or parallel group ones) when a prespecified dose is administered to the patient the clinician allows the drug to reach its pharmacological peak effect (known from PK/PD studies) before the measurement is taken. Therefore, when the wash-out period is not long enough between successive measurements of the same compound, the residual effect from the previous period will force the current measurement to reach its asymptote value earlier, implying that the final current measurement will be similar to the one obtained as if no residual effect is present at all. On the contrary between successive measurements of different compounds, which might react chemically, the effect of the previous compound to the current measurement might be influential in determining the final current treatment outcome. This type of residual effect is called Fleiss (or steady-state) carry-over (see Fleiss [17]).

Overall three models can be considered by the analyst during the design phase for choosing the optimal plan: the simple carryover, the Fleiss-carryover type model, and finally the model with no carryover terms at all. A question worth consideration, is what are the losses in estimating the treatment difference using anyone of the models mentioned before, but in reality any of the other two has produced the observed data.

Although more statistical issues are possible to be tackled in a multi-period setup,

there are some practical disadvantages in conducting such a trial worth considering in more detail. To begin with, the trialist should always keep in mind that sources and time for drug development are limited. As a consequence if more than two treatment periods are used, the time a trial lasts will be extended substantially and the possibility of drop-outs from the study becomes significantly higher. If the time length of wash-out periods is added to that of treatment periods then use of multi-period designs might be strictly cost-prohibited. Furthermore, in cases where a complicated multi-period design is used, clinicians will find difficult to administer treatments according to the protocol of the trial, while statisticians will face difficulties in communicating results from the final analysis. Overall we have to weight carefully the advantages for conducting such a trial, which should be in agreement with the objectives of the study, before we prefer the more complicated design from the conventional 2x2 solution.

4.2 The approach considered here

Suppose that a multi-period cross-over design consisting of s sequence groups and p periods is to be used. The statistician responsible for choosing the optimal within a family of designs, exploits the fact that any treatment difference estimator can be expressed as a weighted average of the ps sequence by period means. This assumption will be adopted in what follows. Note here that it is quite common in practice repeated measurements are collected on each subject within a treatment period. It is debatable if the mean of these measurements is the appropriate statistic for summarizing patient's response at that specific period. An interesting query of how results altered when a different summary measure is chosen, or alternatively when the repeated measurements are used without any attempt to summarize them, is raised.

In a typical ANOVA table, where the results of the analysis of a cross-over trial will be summarized, the total number of available degrees of freedom (df) is $sp-1$. Now, $s-1$ of them will be used for estimating sequence effects (if fitted as fixed effects), $p-1$ for period effects, while the rest, $(s-1)(p-1)$, will be partitioned for assessing treatment and carry-over difference as well as any other estimable

interactions of interest. This implies that the number of periods, number of sequences and number of patients allocated in each sequence, should provide the analyst with adequate information to estimate these parameters. In some trials the treatment effect(s) on sub-group of patients with specific demographic characteristics (e.g. males, aged 50-60) can be of interest. This information can be used by the GP to individualize the treatment regime. For such a trial not only is the choice of the treatment sequences under question, but also the proportion of males/females allocated to different sequences is controversial. On the other hand, in other types of clinical trials (e.g PK/PD studies) the long-term effect of the compound on a target subpopulation might be the focus of attention; if that is the case, the way the total study-completion time is divided into subperiods is debatable. The statistician, in close collaboration with the clinician, should choose the appropriate sub-period length so that the possibility of carry-over presence diminishes, while the proportion of patients who drop-out from the study is kept to a minimum level.

In most studies, both the number of periods and sequences used, are usually fixed in advance by relying mostly not on statistical methodologies but on practical needs. We follow the same policy here by restricting the number of periods and sequences at low levels, so that the whole set of designs for that family can be easily listed. The statistically optimal designs for the simple carry-over model have been derived under the assumption of fixed subject effects (or equivalently fixed sequence effects) and independent within-subject errors. For a full review of these results, see Kershner and Federer [43].

An attempt to relax both assumptions has already been reported in the literature, for example determining optimum design plans assuming random subject effects (see, Laska and Meisner [51]). If the subject effects are considered as random quantities, then the sequence effect which is the average effect of the subjects allocated to that sequence should be random as well. Results under different kinds of carry-over (e.g. Fleiss) will be presented. More specifically, the model considered here is:

$$\bar{y}_{ij} = \mu + s_i + \pi_j + \tau_{d(i,j)} + \lambda_{d(i,j-1,j)} + \bar{\epsilon}_{ij} \tag{4.1}$$

where

$$s_i \sim N(0, \sigma_B^2/n) \text{ and } \bar{\epsilon}_{ij} \sim N(0, \sigma_W^2/n) \tag{4.2}$$

Note that the residual treatment effect ($\lambda_{d(i,j-1,j)}$) depends on the current and previous period treatment in an unspecified way. This equation includes simple, Fleiss and no-carryover model as special cases. If patients are followed up for a long time, within-subject error structure can be safely assumed to follow a stationary first-order auto-regressive process. The (j,k) element of the variance-covariance matrix for the i^{th} subject's error vector is:

$$\text{Cov}(\bar{\epsilon}_{ij}, \bar{\epsilon}_{ik}) = \frac{\sigma_W^2/n}{1-\rho^2}\rho^{|j-k|} \tag{4.3}$$

where n is the number of patients per sequence. In equation (4.1) a serially correlated error structure and a random subject effect are considered simultaneously. This implies that correlation decays with increasing time difference and approaches a limit that is greater than zero for large time differences. This limit is the between subject variance. If that model is adopted for designing a study, knowledge in advanced of this variance component is necessary. Because that information is not readily available, the random sequence effect is removed from the model, leaving the AR(1) structure to describe stochastic dependence between measurements made on a subject.

The d_s optimality criterion is used for determining the optimum plan, regardless of the assumptions made for the fixed or the random part of the model. More specifically the design plan under which the treatment effect is estimated with minimum MSE, will be declared as optimum. Note that the optimum design for estimating treatment difference might be sub-optimum for estimating carry-over difference or any other interaction terms and vice versa. A further restriction on the class of designs studied is that of dual balance i.e. if a treatment sequence (e.g. AABB) is present, then its dual (BBAA) should be present as well. In addition, equal proportion of patients are allocated to that pair of sequences. The concept of duality for treatment sequences is meaningful only when two treatments are compared in the study (e.g placebo-control). In the case where more than two treatments are examined, the duality concept applies to the design as a whole, but not to individual treatment sequences. The main reason for considering this

special family of designs stems from a result proved by Laska and Meisner (see [51]), under which optimal designs are not necessarily dual balanced, but a dual balance design always exists in the family of best plans.

4.2.1 Three period-two sequence designs

The only possible design plans in this case are listed in Table (4.1). In any design presentation each sequence is accompanied by its dual. Following standard notation introduced by Jones and Kenward (see [39]) the above plans will be referred to as 3.2.1, 3.2.2 and 3.2.3, where the first number reflects the number of periods, the second stands for the number of sequences, while the final one is an index to distinguish between different designs. The efficiency of a design for a prespecified effect (e.g treatment or carry-over difference) is defined as the ratio of the variance of the treatment or carry-over estimator at the optimum over the corresponding figure for the design in question. The efficiency of three-period two-sequence designs for the treatment as well as the carry-over difference have been evaluated over a range of possibilities concerning the modeling of the carry-over effect and the within-subject covariance structure. It is well-known that

Table 4.1: Two, four sequence three-period designs

3.2.1	3.2.2	3.2.3	3.4.12	3.4.13	3.4.23
A B B	A B A	A A B	A B B	A B B	A B A
B A A	B A B	B B A	B A A	B A A	B A B
			A B A	A A B	A A B
			B A B	B B A	B B A

under fixed subject effects and independent within-subject errors (see, Jones and Kenward [39] or Kershner and Federer [43]), design 3.2.1 gives minimum variance unbiased estimator both for the treatment and carry-over effect under the simple carry-over model. If no-carryover is assumed and AR(1) error structure is used, design 3.2.2 estimates treatment difference most efficiently over the positive range of the correlation coefficient (ρ). Under Simple (Fleiss) type of carryover design 3.2.3 (3.2.1) is the optimum for estimating treatment effect this time, while all designs are equally efficient for that effect when a second order carry-over term

is added to the first one (second order carry-over model). Similar results are given by Matthews (see [61]). The majority of the results presented above, are valid when the assumption of a uniform covariance structure is made, with the exception that all designs are equally efficient in the case of no-carryover while 3.2.1 instead of 3.2.3 is the optimum under simple carry-over model.

In the cross-over literature, designs which estimate efficiently not only the treatment but also the carry-over difference are preferred. Design 3.2.1 is optimum for estimating carry-over difference for both Simple and Fleiss type of carry-over. This is true irrespective of the covariance structure assumed, although for the Fleiss type of carry-over and under uniform structure 3.2.2 is an equally efficient alternative.

Overall design 3.2.1 seems to have a good performance for estimating both the treatment and residual effect no matter the within-subject error structure assumed, while for those with a special interest on the treatment effect, design 3.2.3 is an excellent alternative. Under no circumstances design 3.2.2 should be used (unless no carry-over is assumed), while for all designs variance of the treatment and carry-over effect decreases as intra-class correlation increases. Sensitivity of best plans under different model assumptions, show that it is easier to propose robust solutions for the residual effect rather than for the treatment effect.

4.2.2 Three period-four sequence designs

The possible design plans in that occasion are listed in the right half of Table (4.1). The available degrees of freedom for the group by period interaction in the three period two sequence designs are $(p-1)(s-1) = 2$ allowing only the estimation of treatment and first-order (or Fleiss) residual effects. In all these designs first-order carry-over is aliased with treatment by period interaction, while the assumption of carry-over being dependent not only upon previous treatment but also upon the current one (treatment by first-order carry-over interaction) cannot be tested.

A way to overcome this problem is to allow for more sequences and/or more periods in the design. Here the first possibility is only considered and the two-sequence three-period designs presented in the previous section, are combined

in pairs giving four-sequence three-period designs. For example, by combining 3.2.i with 3.2.j the three-period four-sequence design 3.4.ij is generated. The three possibilities, labeled 3.4.12, 3.4.13 and 3.4.23, are given in Table (4.1). In this way we increase the available degrees of freedom for the group by period interaction from two to six. If we include treatment, first-order and second-order carry-over terms then two d.f remain for estimating uninteresting sequence by period interaction terms. Note that the 2 d.f of the treatment by period interaction are aliased with first and second order carry-over effects, as for the 2x2 case. Under uniform or AR(1) within-error structure four models will be studied:

- **M1** : Inclusion of residual terms of any kind is not considered based on knowledge about the pharmacological effect of the drug on humans. Only treatment effect (τ) is fitted.
- **M2** : First order carry-over effect added to model M1 (simple carry-over model).
- **M3** : A second order carry-over effect further added to the simple carry-over model (second order carry-over model).
- **M4** : A special type of treatment by first-order carry-over interaction is fitted in addition to the treatment term (Fleiss model).

It was concluded that under model M1 and AR(1) error structure designs 3.4.12 and 3.4.23 are equally efficient for estimating treatment difference, while under the simple carry-over model the ideal choice is 3.4.13. In the Fleiss model, the optimum decision depends upon the correlation coefficient. More specifically design 3.4.13 is preferred for small values of ρ, while for the larger values 3.4.12 becomes the favored one. Finally under the completely unrealistic model M3, design 3.4.23 is the best choice. Similar results for the simple carry-over model are proven by Matthews (see [61]).

If now uniform covariance structure is assumed under model M1 all designs are equally good for assessing treatment effect, whereas for the Fleiss and second order carry-over model, optimum decision depends again upon ρ. For the simple carry-over model, optimum design does not depend on the assumed covariance

Table 4.2: Optimum three period four sequence designs for τ and λ

Upper half : AR(1) structure - Bottom half : Uniform structure						
Optimum for τ (3.4.index)				Optimum for λ (3.4.index)		
M1	M2	M3	M4	M2	M3	M4
.12 or .23 $\forall\rho$	.13 $\forall\rho$	.23 $\forall\rho$	.13 if $\rho \in (0, 0.6]$ .12 if $\rho \in (0.6, 1)$	.13 $\forall\rho$	.13 $\forall\rho$	.12 $\forall\rho$
.12 or .13 or .23 $\forall\rho$	.13 $\forall\rho$	.23 if $\rho \in (0, 0.8]$ .12 if $\rho \in (0.8, 1)$	.13 if $\rho \in (0, 0.5]$ .12 if $\rho \in (0.5, 1)$	.13 if $\rho \in (0, 0.2]$ .12 if $\rho \in (0.2, 1)$	.13 $\forall\rho$	.12 $\forall\rho$

structure.

Turning now to the issue of estimating efficiently the residual effect designs 3.4.12 and 3.4.13 are preferred in all the cases (AR(1) error structure), but in other occasions the unknown value of ρ plays a key role in the final choice (Uniform error structure). Results are shown in Table (4.2).

Obviously if the number of sequences used in the trial is increased, then the precision with which we estimate treatment or residual terms will be increased as well. This implies that four sequence designs should be preferred for running a cross-over study than two sequence ones. But a four sequence design is normally more expensive to conduct and requires the management of four groups of patients. In conclusion, if the experimenter decides to run a four sequence design then 3.4.13 is a good choice as it has good performance for estimating treatment difference when carry-over terms are included in the model, irrespective of the covariance structure assumed.

4.3 Using more periods

In that section designs made of four treatment periods in either two, four or six treatment sequence groups are considered. Only designs made up of dual balanced treatment sequences are investigated. The logistics of running such a study are far more complicated from the study-designs considered so far. If we

assume that the experimenter keeps the completion time of the trial fixed, then sub-dividing this time into four equal time-intervals (instead of three or two), may cause difficulties in collecting the amount of information required for regulatory or other authorities. In addition the cost for conducting such study might not be negligible. From the statistical point of view, by using more periods it is expected that all the effects of interest will be estimated more precisely, but also non-estimable effects in two or three period plans become estimable in the four period family. As in the previous section treatment, first-order and second-order residual effects will be included in the model, but also the best design plan when different carry-over types (e.g. Fleiss) assumed, will be presented.

The optimum design will be the one which estimates treatment (or carry-over) difference with minimum variance. It is obvious that other functions could be considered to optimize, but these choices depend upon the interests of the experimenter. For example minimizing the variance of the overall treatment effect (i.e. treatment plus residual component) or the total study cost are two such functions. As before, both uniform and AR(1) within error structure will be assumed throughout. Finally note that when Fleiss type of carry-over is incorporated into the analysis, second order carry-over of the same type is not included, because it is quite unlikely in practice to occur. The same argument can be put forward for the simple carry-over model, but the reason for considering such a term here, is simply to study the sensitivity of optimum plan when higher order carry-over terms are considered. Third, fourth or higher order residual terms will not bother us in what follows. There are seven different four period dual-sequence designs, listed in Table (4.3).

By allowing more periods, the set of estimable interactions increases, but some of them like the treatment by carry-over one ($\tau\lambda$) are still not estimable. Under the AR(1) within-error structure when treatment and all carry-over terms are included (M3), designs 4.2.6 and 4.2.7 are the optimum ones for estimating treatment effect but the decision depends on intra-class correlation, while in the case of the simple carry-over model (M2) design 4.2.3 is the preferable one. If the Fleiss type of carry-over holds then 4.2.1 and 4.2.6 are equally efficient for small values of ρ, but 4.2.3 is the optimum for the large ones. Finally if the

trialist is confident enough that no residual terms should be present because an adequate wash-out period has been allowed for, then the advisable design is the 4.2.2. These results are also comfirmed by Matthews (see [61]). From the above

Table 4.3: Two-sequence, four-period designs

4.2.1	4.2.2	4.2.3	4.2.4
A A B B	A B A B	A B B A	A B A A
B B A A	B A B A	B A A B	B A B B
4.2.5	**4.2.6**	**4.2.7**	
A A B A	A B B B	A A A B	
B B A B	B A A A	B B B A	

discussion it can be concluded that if a model with elaborated carry-over terms is used, then optimum designs are made of sequences with non-equal replication of A's and B's (designs 4.2.6 or 4.2.7), while as residual terms are removed gradually from the model then equal number of A's and B's appear in each sequence for the optimum plan. This is the price we have to pay for including carry-over terms. The dangers from administered the same drug in a number of adjacent periods is to bias the clinician's assessment of the subject's response, as the randomization code could be easily broken. More importantly if one of the treatments is placebo and design 4.2.6 (or 4.2.7) is used, then a group of patients will suffer discomfort for a long period and be willing to abandon the trial. All the above shows that a lot of conflicting objectives have to be reconciled, one of which is the statistical efficiency, before a specific design is chosen. Under uniform covariance structure similar conclusions derived when compared to the AR(1) case.

Turning now to the optimum estimation of carry-over effect design 4.2.6 seems to have good performance over the range of the models studied and irrespective of the covariance structure assumed. Generally speaking it is easier to find a robust plan for estimating carry-over rather than treatment effect.

By combining two-sequence four-period generic designs in pairs we form 21 distinct four-period four-sequence designs, each one referred to as 4.4.ab if designs 4.2.a and 4.2.b are joined together. Contrast to the two sequence plans, studied before, the treatment by first order carry-over interaction is now estimable.

Under uniform covariance structure and when the full set of carry-over terms is present (model M3), designs 4.4.12 and 4.4.14 estimate τ optimaly, but the decision which one to use depends on ρ. For the simple carry-over model design 4.4.13 is our best choice, while 4.4.16 is the favorite one for the Fleiss type of carry-over, irrespective of the value of ρ. In the absence of any residual terms any combination of 4.2.1, 4.2.2 and 4.2.3 in pairs can be used to estimate optimaly the treatment difference. From the above discussion the major two-sequence design for constructing the optimum four-sequence plan is 4.2.1. If the AR(1) structure is assumed and model M1 (no-carryover) is used for analysis purposes then 4.2.2 is the major building block for the optimum four-sequence design, while design 4.2.3 plays that role for the simple and Fleiss type of carryover; for more details see Table (4.4). This family of plans has also been studied by Matthews (see [61], [62]) and similar conclusions were derived.

Table 4.4: Optimum four-period designs for treatment effect

Upper:Two sequences-Middle:Four sequences-Lower:Six sequences AR(1) within-subject error structure assumed			
M1	M2	M3	M4
.2 $\forall\rho$	.3 $\forall\rho$	.6 if $\rho \in (0.0, 0.4]$ .7 if $\rho \in (0.4, 1.0)$	.1 or .6 if $\rho \in (0.0, 0.5]$ .3 if $\rho \in (0.5, 1.0)$
.23 or .24 or .25 $\forall\rho$	.13 if $\rho \in (0.0, 0.7]$ .35 if $\rho \in (0.7, 1.0)$	.56 or .47 if $\rho \in (0.0, 0.1]$ .47 if $\rho \in (0.1, 0.4]$ .12 if $\rho \in (0.4, 1.0)$	.16 if $\rho \in (0.0, 0.5]$ .13 or .36 if $\rho \in (0.5, 0.7]$ .34 if $\rho \in (0.7, 1.0)$
.235 or .245 .234 $\forall\rho$	.134 if $\rho \in (0.0, 0.2]$ .135 if $\rho \in (0.2, 1.0)$	.126 if $\rho \in (0.0, 0.5]$ .127 if $\rho \in (0.5, 1.0)$	.167 if $\rho \in (0.0, 0.2]$ .136 if $\rho \in (0.2, 0.8]$ .134 or .346 if $\rho \in (0.8, 1.0)$

Following the same principle, combining any three two-sequence four-period designs, a six-sequence four-period design is produced. There are 35 distinct designs in that family and the optimum combination for estimating τ is presented again in Table (4.4). Note here that the 35 distinct designs produced in that way in two treatments constitute all the members for that family. The strategy of producing more complicated designs by combining generic ones can be extended to

the situation where more than two treatments are compared. Worth noting that the plans generated in that way constitute a new design family the size of which grows too fast. Identification of subsets with high probability of containing plans with optimum properties is highly desirable. Formal proof that the optimum plan for the subset is the optimum for the family as well, or at least that the efficiency of the former is quite high, could be difficult to derive. Efficient subset construction could considerably simplify the design search for the original family. It should be noted that the optimality conclusions drawn so far do depend upon the intra-class correlation coefficient ρ.

A related work by J.N.S Matthews (see [61]) in which the simple carry-over model with fixed subject effects and AR(1) within-error structure is assumed as the model generated the data at hand, manages to determine mathematically the optimum design. This work restricts attention on three and four-period design families. Under these assumptions, negative correlation between successive measurements on a patient is possible. Matthews concludes that the final decision concerning the design to use is highly affected not only by the value of ρ, but also the proportion of patients allocated in each sequence group. This is an uninteresting result since ρ is unknown in practice while equal number of patients are usually allocated to the sequence groups. Being aware of these facts, Matthews goes even further and examines the robustness of various designs considered before. He deduces that over the full range of ρ and under the simple carry-over model a design with good performance for estimating treatment and residual effect is 4.4.13 in our notation. Unfortunately our results do not suggest a specific design with good properties over the range of models studied. This is an indication that this line of research will be difficult (if at all possible) to be taken any further.

4.4 Model mis-specification

A wide range of different criteria have been proposed in the literature for choosing the optimum design. In the cross-over set-up the assumed carry-over effect is crucial in deciding the best design for the analysis. Types of carry-over, already

discussed, make this term depending upon current and previous treatment regime (Fleiss carryover). These ideas can be further extended in various directions, producing more elaborated carry-over schemes, although the validity of those plans in real life problems has been questioned a lot in the past (see Matthews [62] and Fleiss [18]). One such direction allows the current patient's response to depend on the whole treatment history of that patient, i.e the residual effect at time t is a function of all treatment effects up to and including time t. Although such a scenario assumes that carry-over from current treatment is present in all (or some) subsequent treatment periods, it is extremely unlikely to be encountered in practical applications.

A more general scheme, can be described as follows: if carry-over from treatment A to treatment B (or from B to A) is denoted by λ, then carry-over from A to A (or from B to B) will be $\phi\lambda$ for some $0 \leq \phi \leq 1$. This will be referred to as the mixed carry-over model in the sequel. Essentially it is an intermediate scenario between simple and Fleiss type of carry-over, since when $\phi = 0$ ($\phi = 1$) then the Fleiss model (simple carry-over model) is recovered. Note that even under that new scheme, residual effect from previous treatment lasts for one period only.

An analyst might feel insecure in using either the Simple or the Fleiss type of carry-over as his analysis model, so he may prefer to let the data decide upon the value of ϕ. In other words, his analysis model should allow for all possibilities. On the other hand if "nature" decides that the observed data are generated by the Simple (or Fleiss) carryover model, and that information is not captured by the analysis model, then the analyst would be interested to know which design gives the minimum mean square error (MSE) for estimating treatment effect under model mis-specification. In other words, if Model 1 is used for the analysis, but Model 2 is the correct one and should have been used instead, which design plan recovers the real treatment difference. In all the above, only two treatments are compared, the within-error structure is assumed known and only dual balanced designs in two, four or six sequence groups and four periods will be considered. The above question can be extended in the case where not only the systematic part, but also the within-patient error structure of the model has been mis-specified; for example although the analyst is using the OLS treatment estimator

which is optimal if one assumes uniform within-subject covariance structure, the same estimator might be sub-optimal and should be modified, if the structure adequately describing the correlation between successive observations in a patient is the AR(1).

Some minimal notation will now be introduced. The design matrices for the "analysis" and the "true" model will be denoted by X_1 and X_2 respectively. Following Laird-Ware's notation for models with fixed and random effects, both the true and analysis model can be described as follows:

$$\mathbf{y} = X_i\beta + Js + \epsilon \qquad i = 1, 2 \tag{4.4}$$

where β includes overall mean, period, treatment and carryover effect (if present), all fixed, while $s \sim N(0, \sigma_B^2)$ being the random sequence effect and J is a vector of ones. The equation above, implies a uniform covariance matrix Σ for the vector of repeated measurements on a particular subject. The AR(1) structure cannot be expressed using a random effects model, but usually is presented in the following form:

$$\mathbf{y} = X_i\beta + \epsilon \qquad i = 1, 2 \tag{4.5}$$

where $\epsilon \sim N(0, \Sigma)$ and $\Sigma_{ij} = \rho^{|i-j|}$. The correlation coefficient is assumed positive throughout, in agreement with the findings in most practical applications. The generalized least squares estimator for the fixed effects under the "analysis" model will be:

$$\hat{\beta}_{analysis} = \left(X_1^T\Sigma^{-1}X_1\right)^{-1} X_1^T\Sigma^{-1}\mathbf{y} \tag{4.6}$$

Note that the model used for "analysis" differ from the "true" only in their systematic part, and more specifically on the type of carryover assumed. Under the "true" model we have:

$$E(\mathbf{y}) = X_2\beta$$

so that:

$$E\left(\hat{\beta}_{analysis}\right) = \left(X_1^T\Sigma^{-1}X_1\right)^{-1} X_1^T\Sigma^{-1}X_2\beta \tag{4.7}$$

It is easy now to evaluate the bias and the variance matrix of the fixed effects as follows:

$$Bias(\hat{\beta}_{analysis}) = \left(I - \left(X_1^T\Sigma^{-1}X_1\right)^{-1}X_1^T\Sigma^{-1}X_2\right)\beta \quad (4.8)$$

$$V(\hat{\beta}_{analysis}) = \left(X_1^T\Sigma^{-1}X_1\right)^{-1} \quad (4.9)$$

"True" model candidates includes the model with no carry-over terms, the Simple and the Fleiss-type carry-over model. Similarly "analysis" model candidates includes all the "true" model nominees plus the mixed carry-over type with ϕ ranging from 0 to 1. In what follows the real treatment difference is assumed to be 5 units, while the carry-over effect is taken as a proportion of the treatment effect. Both treatment and carry-over terms enter linearly into the model and are mathematically unrelated. To facilitate results presentation the ratio λ/τ and the correlation coefficient ρ are classified as follows : low (0.1-0.3), medium (0.4-0.6) and high (0.7-0.9) range of values. In practice 20% or less (low range) of the real treatment effect carries over to the next period, while the correlation among successive measurements on the same subject is usually estimated at about 0.7 (high range). Optimum plans under these restrictions will mainly be discussed in the sequel.

As expected the design used for running the trial depends upon the statistician's choice for the "analysis" model. For example, in the quite likely case, where the model with no residual terms is the "true" model and statistician 1 uses the Simple carry-over, while statistician 2 uses the Fleiss type of carry-over as his "analysis" model respectively, when it comes to design selection they will choose differently. If we restrict attention to four-sequence four-period designs, then statistician 2 should be running his trial using the design (AABB, ABBB, duals) while statistician 1 can choose any one of the about equally efficient plans (AABB, ABBA, duals) and (AABB, AAAB, duals). The two statisticians will disagree on which design should be used to run the trial, even when the Simple or Fleiss carryover schemes are the models responsible for generating our data ("true" model), as long as they pick up different "analysis" models.

General recommendations on the choice of the best design cannot be made; a notable exception to that rule might be when the family of two-sequence four-period designs is decided to be used from the outset. If that is the case a design with

good performance on estimating treatment effect with minimum MSE, irrespective of the "true" or "analysis" model assumed, is ABBA/BAAB. Unfortunately this argument does not hold when designs with more than two sequences are considered. The set of optimum designs, under various scenarios, are summarized in Tables (4.9), (4.10) and (4.11). Each design is identified by a number pre-assigned to it. The key to that index is presented in Table (4.12).
From the statistician's point of view, the robustness of the chosen design to "nature's" choices is the ultimate goal. Suppose for the moment that a statistician after reviewing the cross-over literature feels comfortable in using the simple carry-over model as his analysis model. If he had to choose a six-sequence plan, normally he would choose the design as if the "true" and "analysis" model coincide. Suppose now that "nature" disappoints his expectations and chooses the model with no carry-over terms as the one responsible for generating the observed data. If that is the case, design 22 (see index) should be used for running the study, while in the alternative case where "nature" chooses as "true" model the one with the Fleiss type of carry-over, design 15 becomes his best choice. The wise statistician would prefer a design with high efficiency over a wide range of "nature's" choices. This design usually is not the optimum under anyone of the "true" models, but it has good performance (efficiency more than 90%) over the range of "nature's" choices. Some of these designs will now be presented when six sequences are used. We still assume that $\lambda = 0.2\tau$ and $\rho = 0.7$.

- A down-to-earth statistician decides to use for his analysis model the one with no residual terms. In that case, design 3 has 97% efficiency under the Simple carry-over model, while it is nearly optimum under the Fleiss type of carry-over (efficiency more than 99%).

- A conservative choice could be to use in the analysis phase the simple carry-over model. If that is the case, designs that are highly efficient when the "true" model is the Fleiss one, tend to be of low efficiency under the no-carryover scheme. A design with reasonably good performance under either of these "true" model candidates, is design 33 (ABBA, ABBB, AAAB, duals), which has efficiency of 58% under no-carryover model and only 50% under the Fleiss type of carryover.

- In the unlikely case where the statistician chooses the Fleiss type of carry-over for his analysis model, then design 12 has 94% efficiency if "true" model is the one without residual terms, while it has only 80% efficiency if "nature" selects the Fleiss type of carry-over as the "true" model.

If we are given "true", "analysis" models, correlation coefficient ρ, and it is assumed that residual effect is a small proportion of the treatment effect, then the set of designs the analyst can choose from to run his trial is wider than the corresponding set when λ is a substantial proportion of τ. On the other hand, given "true", "analysis" models and proportion of treatment effect carrying over to the next period, the set of designs the analyst can choose from is similar across possible values of the correlation coefficient. To illustrate the point suppose that the practitioner decides to use a four-sequence design. On pharmacological grounds, he decides to use for his analysis model the Mixed one with $\phi = 0.2$. "Nature" on the other hand produces the observed data using the Fleiss model. Which is the best design to use, so that the MSE of estimating treatment effect is minimized, if correlation among successive measurements on a subject is high? The answer, of-course, depends on the true value of λ. If λ is in the low range then the statistician is free to choose any one of the four designs (5,6,13,14). He has two alternatives if λ lies in the middle range (designs 5,14) and only one choice (design 5) when λ is in the high range. In other words the analyst should worry more to capture correctly residual difference, rather than the way observations are related within subjects.

In our discussion so far it has been assumed that if the Mixed carry-over model is used the proportionality coefficient ϕ is known. This is a strong assumption to be made and one would expect this quantity to be estimated from the data. If that is the case then we are dealing with a non-linear model with respect to ϕ and λ. Usually not sufficient data are available in practice to estimate both ϕ and λ. To prove the point, suppose that the two-sequence three-period design AAB/BBA is used. If n patients are allocated in each sequence, then only the second period data will be used to estimate ϕ. Given that the number of patients recruited in a typical cross-over trial is usually low to moderate, the estimation of ϕ and its standard error will be unstable. As a consequence, ϕ will be assumed

known throughout, implying a linear model in the unknown parameters.
Furthermore the situation in which not only the systematic part, but also the random part of the model is mis-specified by the analyst will be discussed in some detail. The majority of the analysts cannot express any prior opinion about second order behaviour of their data without actually analyzing these data; usually they are much more confident in modeling first order properties. Let's assume for the moment that the analyst uses uniform correlation structure to model the within-patient dispersion matrix, while the appropriate one for the data actually observed is the AR(1). Recall here that the family of designs studied so far has at most four periods. In such occasion, the uniform structure sounds a sensible choice, unless the observations on a subject are quite distant apart in time, in which case the AR(1) is a viable alternative. It is further assumed that each subject offers a complete set of measurements, obtained under an agreed therapy time-table common to all subjects (e.g each patient has his second measurement taken one week after the first one). On the other hand patients can visit the clinic at different dates. Finally intervals between successive measurements are assumed to be similar across subjects. Those assumptions are crucial when second order structure is modeled.
Under those circumstances, the bias vector for the fixed effect parameters is provided by a similar relation to the one used when only the systematic part of the model is mis-specified. The variance-covariance matrix for the same set of parameters is as follows:

$$V\left(\beta_{analysis}\right) = B\Sigma_2 B^T \tag{4.10}$$

where:

$$B = \left(X_1^T \Sigma_1^{-1} X_1\right)^{-1} X_1^T \Sigma_1^{-1} \tag{4.11}$$

and Σ_1 is the dispersion matrix used in the analysis model (uniform in our case), while Σ_2 is the true error structure (AR(1)) which the analyst fails to correctly identify. It seems that the error-structure chosen for the "analysis" model influences to a greater extend the chosen design, compared to the error-structure chosen for the "true" model. This can be seen from the bias and variance equations above, which depend more heavily on Σ_1 rather than on Σ_2.

Optimum designs derived under mis-specification of the systematic part, tend to be highly efficient even in the case where both systematic and random part are wrongly modeled. For the six-sequence family of designs, in the case of practical importance where no carry-over is present, but the analyst insists on incorporating residual term(s) (of some kind) into his model, designs 12 and 31 (see index) are still the best choices. The same set of designs, along with design 9, have high efficiency under the Fleiss carry-over model. When the simple carry-over model become the "true" model, then design (ABAB, ABBA, ABBB, duals) has excellent performance over the whole range of ϕ, ρ and λ.

The logic behind the strategy for choosing the design which gives the minimum variance for the treatment estimator when systematic and/or random part of the model is mis-specified is quite artificial. The statistician has to select his analysis model based on his intuition, experience and background information about the nature and objectives of the study (see Senn and Lambrou [82]). In practice though this intuition is built after experiments with similar set-ups have been analyzed and sometimes after combining results from various studies (meta-analysis). In this way information is obtained not only about the nature and magnitude of residual terms, but also about secondary parameters, such as intra-class correlation. The effective planning of future trials highly depends on the quality of this background information. It is exactly this experience used throughout in this section concerning plausible values of ρ and λ/τ. If the analyst feels that not enough information is available to justify his choice on key parameters, then a Bayesian approach could be adopted to incorporate this uncertainty. Results of optimum designs under the Bayesian perspective are very limited in the cross-over literature, mainly due to computational difficulties that frequently arise during the implementation phase.

4.5 What makes a good plan

To understand why some designs estimate treatment effect more accurately than others under specific model assumptions, it is wise to take a closer look at the treatment estimators proposed under these plans. Recall that treatment estima-

tor is a weighted average of the *ps* sequence by period cell means. The expectation of this estimator is free of period, sequence and residual terms.

The procedure followed for the specification of the weights has already been considered in the 2x2 case, and will be further illustrated here using the four period, two sequence design AABB/BBAA. Following Senn (see [77]), we first eliminate sequence and period effects. This implies that for a given column (row) the weights must add up to zero. This results in the scheme displayed in Table (4.5). Note that only those two constrains reduce the number of unknown weights from eight to three. An estimator of the difference between the effect of treatment A and the effect of treatment B would require that if weights summed over the symbol A the result should be 1, while summed over the symbol B the result ought to be -1. This imposes the further constraint: $w_1 + w_2 = 1/2$. Suppose

Table 4.5: Weights after eliminating sequence and period effects

A	A	B	B
w_1	w_2	w_3	$-(w_1 + w_2 + w_3)$
B	B	A	A
$-w_1$	$-w_2$	$-w_3$	$w_1 + w_2 + w_3$

now that three experts express three different opinions about the type of carry-over occurred during the study. The first expert after considering the half-lives of the drugs involved and the length of the wash-out period used, he strongly supports the opinion that the presence of carry-over is high unlikely (no carry-over). The second expert believes that no matter the serious attempt made to eliminate residual effects, there will still be a carry-over of meaningful size (simple carry-over). Finally, the third one who do not feel so confident that the wash-out period used excludes in all cases the possibility of residual term being present, he combines the opinions of the other two and suggests that residual effect is not possible when a treatment is followed by itself, but it might be possible in other scenarios (Fleiss carry-over).

The next step is to eliminating residual effects form the treatment estimator. The weights produced, shown in the three lay-outs below, correspond to the the three carry-over types; no carry-over, simple carry-over and Fleiss carry-over.

Simple carryover scheme				Fleiss carryover scheme			
A	Aa	Ba	Bb	A	A	Bα	B
$1+2w$	$-1/2-2w$	w	$-1/2-w$	$1/2-w$	w	0	$-1/2$
B	Bb	Ab	Aa	B	B	Aβ	A
$-1-2w$	$1/2+2w$	$-w$	$1/2+w$	$-1/2+w$	$-w$	0	$1/2$

As far as the notation is concerned, English lowercase letters correspond to simple carry-over effect, while Greek ones to the Fleiss-type carry-over. Note that the associate weights for either letters add up to zero, a fact which ensures the elimination of the residual term, when the expectation of the treatment estimator is evaluated. In order to obtain a unique set of weights in each case further

No carryover scheme

A	A	B	B
$1/2-w_2$	w_2	w_3	$-1/2-w_3$
B	B	A	A
$-1/2+w_2$	$-w_2$	$-w_3$	$1/2+w_3$

constrains are needed. For those who include carry-over term of any kind in their model only one weight is still unknown, but for the model without residual terms two weights need specification. To keep things simple, we assume that observations are independent. In that case, the variance of the treatment estimator is proportional to the sum of squares of the weights for each scheme. One way of selecting the unknown weight(s) is by minimizing that variance.
Applying that rule, the following set of weights is derived, under the three models considered:

$$\hat{\tau}_1 = \left(\frac{5}{20}\bar{y}_1 + \frac{5}{20}\bar{y}_2 - \frac{5}{20}\bar{y}_3 - \frac{5}{20}\bar{y}_4\right) + \left(-\frac{5}{20}\bar{y}_5 - \frac{5}{20}\bar{y}_6 + \frac{5}{20}\bar{y}_7 + \frac{5}{20}\bar{y}_8\right)$$

$$\hat{\tau}_2 = \left(\frac{6}{20}\bar{y}_1 + \frac{4}{20}\bar{y}_2 - \frac{7}{20}\bar{y}_3 - \frac{3}{20}\bar{y}_4\right) + \left(-\frac{6}{20}\bar{y}_5 - \frac{4}{20}\bar{y}_6 + \frac{7}{20}\bar{y}_7 + \frac{3}{20}\bar{y}_8\right)$$

$$\hat{\tau}_3 = \left(\frac{5}{20}\bar{y}_1 + \frac{5}{20}\bar{y}_2 - \frac{0}{20}\bar{y}_3 - \frac{10}{20}\bar{y}_4\right) + \left(-\frac{5}{20}\bar{y}_5 - \frac{5}{20}\bar{y}_6 + \frac{0}{20}\bar{y}_7 + \frac{10}{20}\bar{y}_8\right)$$

where $\hat{\tau}_1$,$\hat{\tau}_2$ and $\hat{\tau}_3$ estimate the treatment effect under the model with no carry-over terms, Simple and Fleiss type of carry-over respectively. Note that $\hat{\tau}_1$ (the estimator with all weights equal in absolute terms) has the smallest variance of the three. Another scheme of weights could have resulted if the statistician tried to eliminate both Simple and Fleiss-type carryover, or Simple and second order carryover simultaneously.

The above approach is applied to decide the optimum six sequence, four period dual balanced design, assuming AR(1) within error structure with $\rho = 0.7$. The new element added here is that a different proportion of the available patients is allocated in each dual sequence group. Three dual sequences are involved in the family of designs considered, which implies that a proportion p of them is allocated to the first one, a proportion q to the second one, while the rest (1-p-q) to the third one. In that case the precision with which the treatment effect is estimated, will be affected not only by the auto-correlation coefficient (known here), but also by the allocation scheme. The weights are chosen so that Simple, Fleiss and Mixed type of carry-over are eliminated from the treatment estimator in the corresponding model. Varying ρ will not qualitatively alter our conclusions. Similar work has been done by Matthews (see [65]).

In the majority of reported clinical trials there is an equal allocation of patients to the sequence groups, usually controlled by a central randomization system. Obviously this is the optimal allocation of patients in a longitudinal study, if the serial observations on each subject are assumed independent or equally correlated. Under more elaborated error structures this might not be the case. There are also other practical needs to study the efficiency of designs with non-equally replicated sequences. For example, in a multi-center study small centers will be running part of the planned sequences with a moderate number of subjects randomized in each one, while larger medical units have the infrastructure to recruit an adequate number of subjects equally allocated in each sequence. Putting these facts together, inevitably one concludes that equal proportions per sequence is the exception rather than the rule in real life applications.

Results concerning the optimal plan along with the optimal allocation of subjects to sequences for that plan, over the six-sequence, four-period family of designs

are presented in Table (4.6). As it is expected the highest proportion of the

Table 4.6: Optimum 6-sequence, 4-period designs

AR(1) within-subject error structure with $\rho = 0.7$. Unequal proportion of patients is allowed in each dual sequence. No model mis-specification is allowed for						
Model	Design			Proportion		
No carryover	ABAB	ABBA	ABAA / duals	0.8	0.1	0.1
	ABAB	ABBA	AABA / duals	0.8	0.1	0.1
	ABAB	ABAA	AABA / duals	0.8	0.1	0.1
Simple carryover	AABB	ABBA	AABA / duals	0.1	0.8	0.1
Fleiss carryover	AABB	ABBA	ABAA / duals	0.1	0.8	0.1
	ABBA	ABAA	ABBB / duals	0.8	0.1	0.1
Mixed with $\phi = 0.2$	AABB	ABBA	ABBB / duals	0.1	0.8	0.1
Mixed with $\phi = 0.5$	AABB	ABBA	ABBB / duals	0.1	0.8	0.1
Mixed with $\phi = 0.8$	AABB	ABBA	AABA / duals	0.1	0.8	0.1

recruited patients is assigned to the optimal two-sequence plan in the family of two-sequence, four-period designs. For example, for the model free of carry-over terms 80% is allocated to the dual sequence ABAB/BABA (40% in each single sequence), while under any model which includes residual term of any kind, the same percentage of subjects is now allocated to the dual sequence ABBA/BAAB. In design theory both the selection of sequences and the proportion of the available resources assigned to each one of them are treated as unknown quantities. The optimization problem need to be solved is considerably simplified when the assumption of equal allocation of subjects to sequences is made. The implication of that assumption is hard to be assessed in practice. Note that the optimality criterion has to be slightly modified, in the case where non-equal number of patients is allowed for in each sequence. More specifically if X_s denotes the design matrix of the s^{th} sequence, then the information matrix, i.e. the inverse of the covariance matrix for the fixed effects, for any design plan is:

$$\sum_s p_s X_s^T \Sigma^{-1} X_s \tag{4.12}$$

where p_s is the proportion of subjects allocated to the s^{th} sequence and Σ stands for the within-subject error structure (AR(1) here). In the case of a balance allocation of subjects to sequences, the information matrix is similar to the unbalanced case one presented above, with the p_s terms removed.

4.6 A cross-over clinical trial in 7 treatments

A cross-over trial was carried out for comparing two different formulations of a compound, called formoterol, used to treat patients suffering from asthma. The old formulation was a dry powder of formoterol delivered from a single dose device, called ISF, while the new one was a multi-dose inhaller named MT&A and developed by a pharmaceutical company. This was a multi-center clinical trial carried out in four different countries under the close supervision and assistance of people at company's headquarters. The data were kindly provided by Senn et al (see [83]).

An important query needed to be tackled at the planning stage of the trial was the number and level of different doses for each formulation. Three doses of MT&A (6, 12 and 24mg per puff) and three of ISF (6, 12 and 24mg per puff) are to be compared. One of the study-objectives was to determine the time-response curve at each dose for each formulation. For ethical reasons placebo was also given during the course of the study. This implies seven treatments altogether. The response variable was force expiratory volume in one second (FEV). The recommended treatment regime for each patient was one or two puffs daily. Each patient was followed for a time period of five days and according to well-known PK/PD properties of formoterol, a wash-out period of at least two days would eliminate any residual effect. A four-day wash-out period between successive active treatment periods was agreed. This ensures that carryover should not be a consideration for the statisticians involved, neither in the design choice nor in the analysis. The design plan was produced by cycling the sequences:

(MT&A6, Placebo, ISF24, MT&A24, ISF6, MT&A12, ISF12)

(MT&A6, ISF24, ISF6, ISF12, Placebo, MT&A24, MT&A12)

(MT&A6, MT&A24, ISF12, ISF24, MT&A12, Placebo, ISF6)

until a 7x7 latin square is produced from each one of them. In that way a cross-over design of 21 sequences in 7 periods is generated. If we delete periods 6 and 7, a 21-sequence, 5-period cross-over plan is finally produced.

The above design may not be the optimal one for minimizing the variance of treatment contrasts on which the trial investigators were interested, but it is a reasonable choice given the time constraints faced by the statisticians involved. Although the precaution of an adequate wash-out period to eliminate the possibilities of presence of carry-over effect of any kind were taken, an interesting query of how this optimal plan may change if we assume the existence of certain forms of carry-over, like the Simple or the Fleiss one, is now raised. Note that a baseline measurement was taken before each treatment measurement. It turned out that baseline measurements had a tremendous explanatory power for the response variable, which in our case is the logarithm of FEV measurements.

This is an incomplete block design and estimation of treatment contrasts can be done in various ways depending on how the patient effect is treated. If a fixed patient effect is assumed then all treatment comparisons are made using within patient differences. On the contrary by modeling patient effect as a random component some inter-block information can be recovered, i.e. a weighted combination of between and within patient differences forms now the treatment estimator. Both cases will be covered.

With 7 treatments a set of at most 6 treatment contrasts is estimable. The orthogonal set of treatment contrasts chosen by the authors (see [83]) were the following: the first one compares the average treatment effect to the placebo one. The second one addresses the question for which the trial was set up, i.e. does the new treatment formulation of formoterol (MT&A) gives on average higher FEV measurements compared to the old one (ISF) or not? The third contrast, ("slope"), examines the linear effect of dose level on the response aggregated over formulations, while the "curvature" contrast tests for a similar quadratic effect. The next contrast checks whether or not the linear effect of MT&A is parallel to the linear effect of ISF, while the final one examines if the vertical distance of the average reading at 12mg from the line joining the average readings at 6mg and 24mg is the same for the two formulations. For simplicity the last two contrasts

will be referred as "Parallelism" and "Opposite Curvature" in what follows.
Only the first three contrasts were found statistically significant at 5% level, after fitting a variety of models. The first model excludes residual terms of any kind, in agreement with the investigator's beliefs. The second model is a simplification of the Simple carry-over scheme, in which only active agents effect can persist to the next period, while placebo's residual effect is negligible. In our final model the amount of dose that carries-over to the next period depends not only on the current dose level and type of formulation, but also on the type of formulation of the next period. This new type of carry-over, a special case of the Fleiss type, will be referred to as the "proportional" type of carry-over in the sequel and is presented in Table (4.7). Note that from the 161 patients participating in the trial, only 148 provided complete sequence of 5 measurements. Overall 158 patients were available for analysis. Those who discontinued for any reason gave 31 measurements. The incomplete sequences were taken into account at the analysis stage. Frequentist analysis in which subject effect is considered as

Table 4.7: Fleiss type of carry-over for the 7 treatment trial

ISF6	$\rightarrow$	ISF12, ISF24	λ	MT&A6	$\rightarrow$	MT&A12, MT&A24	μ
ISF12	$\rightarrow$	ISF6, ISF24	2λ	MT&A12	$\rightarrow$	MT&A6, MT&A24	2μ
ISF24	$\rightarrow$	ISF6, ISF12	3λ	MT&A24	$\rightarrow$	MT&A6, MT&A12	3μ

either fixed or random will be discussed for all three models. Bayesian analysis with random subject effect is also covered. Results are presented in Table (4.8). For sake of presentation M1 refers to the model with no-carryover terms, M2 to the simple carry-over model, while M3 to the proportional type of carry-over. The important message though, is that the dose-log(response) curve for ISF can be derived from the corresponding curve for MT&A by a vertical shift upwards of about 0.01. Unfortunately the new formulation proved unsuccessful and this conclusion stays valid irrespective of the inclusion or not of any residual effect in the model, or the way the patient effect is treated (fixed or random).
The analysis of that multi-period, multi-treatment trial points out that carry-over effect of any kind, if fitted, does not alter the conclusions concerning the treatment effect in a substantial way, if the precaution of eliminating residual

Table 4.8: Analysis of the 7 treatment cross-over trial (x 10^{-2})

Model	Drug vs Placebo	ISF vs MT&A	Slope	Baseline effect	σ_B^2	σ_W^2
Frequentist with fixed subject effect						
M1	12.15 (0.71)	9.98 (0.93)	5.40 (0.76)	46.01 (3.87)		7.67
M2	12.27 (0.82)	10.04 (0.99)	4.62 (0.91)	45.71 (3.89)		7.68
M3	12.09 (0.73)	10.06 (1.04)	5.45 (0.77)	46.00 (3.88)		7.68
Frequentist with random subject effect						
M1	12.43 (0.74)	10.10 (0.97)	6.01 (0.79)	78.40 (2.19)	1.04 (0.15)	0.68 (0.03)
M2	12.20 (0.88)	10.24 (1.04)	5.52 (0.94)	78.54 (2.19)	1.03 (0.15)	0.65 (0.03)
M3	12.41 (0.76)	10.31 (1.10)	6.03 (0.80)	78.46 (2.19)	1.04 (0.15)	0.65 (0.03)
Bayesian with random subject effect						
M1	12.40 (0.75)	10.10 (0.97)	6.00 (0.79)	77.90 (3.26)	1.09 (0.18)	0.65 (0.04)
M2	12.20 (0.88)	10.20 (1.05)	5.50 (0.94)	78.20 (2.93)	1.07 (0.16)	0.66 (0.03)
M3	12.40 (0.76)	10.30 (1.12)	6.02 (0.81)	78.30 (2.91)	1.07 (0.16)	0.65 (0.04)

effect at the design phase has been taken.

4.7 Discussion and other related results

As has already been explained there are many reasons to use more than two periods in a cross-over trial. For the clinicians who use those designs in practice a reasonable question arises: what is the optimal design to use if the number of periods, number of sequences and number of treatments one decides to compare are provided. Furthermore what is the optimal allocation of the available number of patients to each sequence group? For notation's sake a cross-over designs in t treatments, n subjects (units), and p periods, will be denoted as co(t,n,p). In the majority of optimality results drawn so far in the literature, the simple carry-over model is assumed with independent within-subject error structure. The optimality criterion used is universal optimality, which implies A-,D- and E-optimality criteria. A design is universal optimal if its information matrix, C, satisfies the following conditions, as described by Kiefer (see [44]):

- The rows of C sum to zero.
- The diagonal elements of C are equal as well as the off diagonal elements.
- C has maximal trace over the family of designs which universal optimality is claimed.

Most of the optimum designs possess certain characteristics; they are uniform balance designs. A design is uniform if each treatment appears the same number of times in each period and is administered the same number of times in each subject. This implies that both the number of periods and the number of subjects must be a multiple of the number of treatments the trialist is prepared to compare. The balance property requires that each treatment is proceeded equally often by any other treatment and never by itself. The "strongly balanced" property ensures that each treatment follows any other treatment the same number of times including itself.

The first important optimality result for cross-over designs was derived by Hedayat and Afsarinejad (see [33]), who proved that uniform balance designs are optimal for estimating treatment and carry-over effects over the class of uniform designs, when the number of periods used is the same as the number of treatments. Cheng and Wu (see [5]) were able to relax the uniformity assumption and prove that uniform balance cross-over designs are optimal for estimating only the carry-over effect over all co(t,λt,t) designs. To be able to extent this result for the treatment effect as well, a uniformity assumption need to be imposed over the family of designs in which optimality is claimed. More specifically, if the class of designs which are uniform on units and uniform on the last period only is considered, then uniform balance designs are optimum for the estimation of the treatment and carry-over effect over this class.

From results in previous sections it is evident that the three-period, two-sequence design ABB/BAA is optimal (or near optimal) under various model assumptions. This design results, if the last period of the classical 2x2 design (AB/BA) is repeated. More generally Cheng and Wu proved that if in a balance uniform co(t,λt,t) design, the last period is repeated, then a universally optimal design for the treatment and carryover effect is obtained over the designs in co(t,λt,t+1).

An even more general result, proved again by Cheng and Wu (see [5]) and raises further restrictions is the following: A **strongly** balanced uniform design is universal optimal for the estimation of treatment and carry-over effect over all designs in co(t,n,p). The practical difficulty in implementing that result stems from the fact that strongly balance uniform designs are more restricted than balance uniform designs. As a consequence it is more difficult for the statistician to suggest a strongly balance uniform design than a balance uniform one.

In the case where only two treatments are compared, optimality results were derived by Laska and Meisner (see [51]), who were the first including random subject effects in their model, so that the covariance matrix of each subject's response is the uniform one. They proved that when the number of periods p is even, then a strongly balance uniform design always exists and it is optimal for estimating treatment and carry-over effect over all co(2,λp,p) designs. If the number of periods used is odd, then the optimal design for the estimation of treatment and carry-over effect consists of a strongly balance uniform design in the first p-1 periods, while the last period is a repetition of the $(p-1)^{th}$ period. This result is valid even in the case where baseline measurements are available. Unfortunately Laska and Meisner were not able to provide an analytical result if an AR(1) within-error structure with positive correlation assumed for the covariance matrix of each subject's response, but a computer program was used to search all possible designs and find the optimum for p=3 and p=4. Conclusions depend upon the correlation coefficient. If more than two treatments are compared and an AR(1) within-error structure is assumed then Gill and Shukla (see [23]) suggest that if $\rho < 0$ then the optimal design should change over the treatments as little as possible, whereas if $\rho > 0$ then in the optimum design each treatment should be preceded and followed by other treatments.

4.8 Suggestions-Conclusions-Future Directions

The results presented in this chapter, provide a better insight why the two stage procedure (presented in the previous chapter) performs worse than the CROS estimator under any performance criterion we consider. It is clear that lack of

information concerning residual effect results in highly inefficient two stage procedure. Furthermore, treatment effect is estimated better when designs made of sequences where treatments appear equally often in each sequence are used, whereas designs made of sequences with un-equal repetitions of A and B are performing better when estimation of residual terms is under consideration. It was pointed out that the inclusion of more periods and/or sequences may improve the power with which residual effects of any kind are detected, but the best design plans used for testing carry-over terms may prove bad choices for detecting treatment differences.

In the Bayesian analysis of the cross-over example in the previous chapter, we modeled the residual effect as a proportion of the treatment difference. It was assumed that if $a\%$ of treatment A carries over to treatment B, then the same proportion carries over from B to A. At first glance this argument may sound un-reasonable, but bearing in-mind that the two asthma drugs we try to compare have similar pharmacokinetic profiles, then this assumption may be justified. In practice, any analyst would like to hypothesize that the fraction of treatment that carries over from A to B is different from the fraction that carries over from B to A. Then, it would make sense to explore how our design choice changes, if that non-linear model, or any modification of it, is considered as the "true" expected to generate the data. A first attempt to tackle this question is made in the next chapter.

Another important issue, not well-explored in the cross-over literature, is the use of N-of-1 trials. In such a trial individual patients are given repeated administration of at least two treatments with the objective to learn something about the effect of the drug in a given patient rather than for patients in general. This requires that patients must be willing to be treated at least three times, although the possibility of using 6 to 8 periods is reasonable. A further study objective would be to investigate the variation of individual response to treatment. Optimum plans for this scenario would be quite useful to be derived.

Table 4.9: Optimum two, four and six sequence designs

Only the systematic part of the model is mis-specified. True model:No carryover. Within error structure:AR(1).									
Correlation Coefficient (ρ)									
	Low			Medium			High		
λ	Low	Med	High	Low	Med	High	Low	Med	High
Analysis model:Simple Carryover									
Two	3,6	6	6	3,7	7	7	3,7	7	7
Four	2	2,21	2,21	2,11,16,21	11,16,21	11,16,21	2,16	16	16
Six	12,22,34	12,22,34	12,22,34	22,31	22,31	31	22,31	22,31	22,31
Analysis model:Fleiss Carryover									
Two	1,6	1,6	1,6	1,3,6	1,6	1,6	1,3,6	1,6	1,6
Four	11	11	11	11	11	11	11	11	11
Six	31	31	31	31	31	31	12,31	31	31
Analysis model:Mixed with $\phi = 0.2$									
Two	1	1	1	1,3	1	1	1,3	1	1
Four	11	11	11	11	11	11	2, 16	16	16
Six	31	31	31	31	31	31	22, 31	31	31
Analysis model:Mixed with $\phi = 0.5$									
Two	1	1	1	1, 3	1	1	1, 3	1	1
Four	11	11	11	16	16	16	16	16	16
Six	31	31	31	31	31	31	22,31	22,31	22,31
Analysis model:Mixed with $\phi = 0.8$									
Two	3,6,7	6,7	6,7	1,3,7	1,7	1,7	1,3	1	1
Four	2,21	2,21	21	2,16	16	16	2,16	16	16
Six	12,22,31	12,31	12,31	22,31	31	31	22,31	22,31	22,31

Table 4.10: Optimum two, four and six sequence designs

Only the systematic part of the model is mis-specified. True model:Simple carryover. Within error structure:AR(1).									
Correlation Coefficient (ρ)									
	Low			Medium			High		
λ	Low	Med	High	Low	Med	High	Low	Med	High
Analysis model:No carryover									
Two	3	3	3	3	3	3	3	3	3
Four	2	2	2	2	2	2	9	9	9
Six	3,6	3,6	3,6	6	6	6	6	6	6
Analysis model:Fleiss carryover									
Two	3,6	2	2	3,6	2	2	2,3,6	2	2
Four	5,13	5	5	5,13,14	5	5	12,13,14	5,12	5,12
Six	4,15,16	4	4	15,16	15	15	15,16	15	15
Analysis model:Mixed with $\phi = 0.2$									
Two	3,6	2,6	2	3,6	2,6	2	2,3,6	2,6	2
Four	5,13	5	5	5,13,14	5	5	12,13,14	12	5,12
Six	4,16	4	4	15,16	15	15	15,16	15	15
Analysis model:Mixed with $\phi = 0.5$									
Two	3	6	2,6	3	6	2,6	3,6	2,6	2
Four	13	5,13	5	6,13,14	5,14	5	6,13,14	12,14	12
Six	4,16	4,15	4	15,16	15	15	15,16	15	15
Analysis model:Mixed with $\phi = 0.8$									
Two	3	3	3,6	3	3	3,6	3	3,6	6
Four	2,13	13	13	2,6,13	13	13,14	6,13	13,14	12,14
Six	3,16	16	4,15,16	3,16	16	15,16	3,16	15,16	15

Table 4.11: Optimum two, four and six sequence designs

Only the systematic part of the model is mis-specified. True model:Fleiss carryover. Within error structure:AR(1).									
Correlation Coefficient (ρ)									
	Low			Medium			High		
λ	Low	Med	High	Low	Med	High	Low	Med	High
Analysis model:No carryover									
Two	1,6	1,6	1,6	1,3,6	1,3,6	1,3,6	3	3	3
Four	11	11	11	2,11,13	2,11,13	2,11,13	2,6,13	2,6,13	2,6,13
Six	12,31	12,31	12,31	12	12	12	3,12,16	3,12,16	3,12,16
Analysis model:Simple carryover									
Two	2,6	2	2	2,6	2	2	2,4	2	2
Four	5,6,8,12	8	8	5,12	5,8	8	5	5	5
Six	4,9	9	9	4,9,15	9,15	9,15	15	15	15
Analysis model:Mixed with $\phi = 0.2$									
Two	6	6	4	3,6	4,6	2,4	3,6	2,4,6	2,4
Four	11,14	5,6,12,14	5	11,13,14	5,12,14	5	5,6,13,14	5,14	5
Six	12,16,31	4,15	4	12,16	4,15	4,15	15,16	15	15
Analysis model:Mixed with $\phi = 0.5$									
Two	4,6	2	2	4,6	2	2	2,3,4,6	2	2
Four	5,6,11,12	5,8	8	5,12,13,14	5	5,8	5,14	5	5
Six	4,14,16,31	4,9	9	4,15,16	9,15	9,15	15,16	15	15
Analysis model:Mixed with $\phi = 0.8$									
Two	2,4,6	2	2	2,4,6	2	2	2,4,6	2	2
Four	5,14	8	8	5,14	5,8	5,8	5,14	5	5
Six	4,9,15	9	9	4,9,15	9,15	9,15	15	15	15

Table 4.12: Key to optimal designs under different model assumptions

Key to the four sequence designs			
2	5	6	8
AABB/BBAA	ABAB/BABA	ABBA/BAAB	ABAB/BABA
ABBA/BAAB	ABAA/BABB	ABAA/BABB	AABA/BBAB
9	11	12	13
ABBA/BAAB	AABB/BBAA	ABAB/BABA	ABBA/BAAB
AABA/BBAB	ABBB/BAAA	ABBB/BAAA	ABBB/BAAA
14	16	21	
ABAA/BABB	AABB/BBAA	ABBB/BAAA	
ABBB/BAAA	AAAB/BBBA	AAAB/BBBA	
Key to the six sequence designs			
3	4	6	9
AABB/BBAA	ABAB/BABA	AABB/BBAA	ABAB/BABA
ABBA/BAAB	ABBA/BAAB	ABBA/BAAB	ABAA/BABB
ABAA/BABB	ABAA/BABB	AABA/BBAB	AABA/BBAB
12	14	15	16
AABB/BBAA	AABB/BBAA	ABAB/BABA	ABBA/BAAB
ABBA/BAAB	ABAA/BABB	ABAA/BABB	ABAA/BABB
ABBB/BAAA	ABBB/BAAA	ABBB/BAAA	ABBB/BAAA
22	31	34	
AABB/BBAA	AABB/BBAA	ABAA/BABB	
ABBA/BAAB	ABBB/BAAA	ABBB/BAAA	
AAAB/BBBA	AAAB/BBBA	AAAB/BBBA	

Chapter 5

Multi-period, multi-sequence designs in general

5.1 Designing for a purpose

The aim of any clinical trial, cross-over or parallel one, is to compare two or more treatments on the basis of experimental data. The confidence with which the effects of the various treatments will be assessed, depends to a large extent on the plan chosen to conduct the trial. A typical 2x2 cross-over experiment can be seen as a randomized block design with 2 blocks (AB, BA), in which n experimental units randomly allocated in each block. Similarly any cross-over experiment in s sequences and p periods, can be seen as a randomized design with s blocks. The allocation of treatments to these blocks should be carried out in such a way that treatment contrasts of direct interest are estimated with the highest precision.
A good plan usually depends on the number of times a treatment appears in each sequence. In a cross-over trial, each patient should try all available treatments at least once. This implies that if t treatments are compared, then the number of repeated measurements collected from each patient should be at least t. Allowance of adequate wash-out intervals is essential to ensure high data quality in these circumstances. If $p > t$ then replication of treatments in each sequence is inevitable, leading to more efficient treatment estimates, especially for two treatment comparison as illustrated in the previous chapter.
On other occasions, due to situation restrictions, it is not possible to have a single

replication of every treatment into each sequence, in other words the number of periods used is less than the number of compounds compared. This type of design known as "incomplete block design", is a popular choice in the industry, though more sequences might be required to achieve similar treatment effect accuracy as in a design where each treatment appears at least once in each sequence. The clinical management of such a study might be difficult as well. In the "incomplete block designs" the number of times treatments appear in each sequence is crucial in the precision with which pair-wise treatment comparisons are made. An excellent review of these type of designs can be found in Fisher and Yates (see [16]).

In a typical analysis of a cross-over trial, treatment comparisons are based on weighted averages of within-sequence treatment estimates. Optimal plans are selected on the basis that both the chosen sequences and the weights attached to the treatment estimates derived from them, provide the best overall picture of drug activity. In any cross-over design though, treatment comparisons can also be evaluated using between-sequence information, although such estimates are given less credibility because of the high between-sequence variability. Combination of these two pieces of information, known widely as "recovery of inter-block information" in the statistical literature, is being routinely implemented in everyday statistical analysis.

Recovering the inter-block information will make sense only if the "sequence" effect fails to achieve a marked reduction in the error mean square. If that is the case, then the amount of information regarding treatment activity recovered from the inter-block analysis will alter conclusions to an appreciable extent. On the other hand if "sequence" effects are large, then inter-sequence treatment information could be safely ignored.

The families studied in this chapter, are usually composed of multi-period designs (number of periods $>= 5$). For long-period families it makes sense to assume that the within-sequence error structure is described by an AR(1) process. It will be shown that the AR(1) correlation coefficient plays a key role in determining the optimal plan under specific carry-over schemes. Recovering inter-block information, implying compound symmetry covariance structure, is appropriate for

cross-over plans with short sequences (see previous chapter).

5.2 Setting the scene

In the previous chapter cross-over plans for the comparison of two treatments only were considered. In the current chapter up to six treatments are compared under different assumptions made for the carry-over term. A typical user of cross-over plans would require each patient to act as his own control, trying all available treatments. This implies, that all treatments appear at least once in each sequence (i.e. $p > t$).

The full model used throughout for comparing two treatments contains terms for the general mean, period, treatment and first-order carry-over effects. For the comparison of three or more treatments a fixed sequence effect is added to the previous model. The reason a fixed sequence effect is not included for the comparison of two treatments is because it leads to non-estimable treatment effect under specific carry-over schemes (e.g. simple carry-over and design AAAAB/BBBBA). In addition, it is assumed that there is a standard treatment (labeled A) and the contrasts we are interested in estimating with the highest precision are the ones that compare each of the newly proposed treatments with the standard one. This implies that comparisons between the new treatments may have lower precision. Furthermore, designs for efficient comparison of carry-over differences will not bother us in the sequel, since carry-over terms are of less importance.

5.2.1 Comparing two-treatments

Best plans were derived under different carry-over assumptions. More specifically for two treatment comparison the following type of residual effects have been studied:

- No carry-over terms included.
- Simple carry-over model (described in detail elsewhere).
- Fleiss carry-over, i.e a treatment can carry-over to any other treatment but not to itself.

- Mixed carry-over, i.e if A carries over to B λ units, then B carries over to A $\phi\lambda$ units where $0 \leq \phi \leq 1$. Three values of ϕ are considered: $0.2, 0.5$ and 0.8.

The within-sequence covariance structure is the stationary auto-regressive of order one. Negative values of ρ, the correlation between adjacent measurements, are not considered. The sensitivity of the optimal plan to positive values of ρ is studied. Three values for the correlation coefficient ($\rho = 0.2, 0.5$ and 0.8) are considered, reflecting the different spacing (long, medium and short, respectively) of the repeated measurements collected within sequences that may occur during the course of the cross-over experiment.

5.2.2 Comparing more than two treatments

More possibilities arise in that occasion, due to the larger number of treatment contrasts likely to be tested. It is also easier to consider more elaborate carry-over schemes, which in the two-treatment case are not applicable. Simple and Fleiss carry-over schemes are extended in a natural way to the multi-treatment case. Mixed carry-over scheme is impractical here, since different values of ϕ could be assumed for different pairs of treatments. For ethical reasons, placebo is typically one of the treatments in this type of trials. If that is the case, it is assumed that carry-over from placebo to any other treatment is nonexistent. Furthermore, there is no placebo treatment effect. For sake of argument this type of model refers to as "Simple2" in what follows.

Finally in most pharmacological studies, treatments are administered in increasing doses. The aim of these experiments is to discover the dose with the highest response. In a typical study three doses of each compound are considered. The doses are chosen in the low, medium and high part of the dose-response curve. It is assumed here, that increasing the dose by a factor k would increase the pharmacological response by the same factor. A similar argument applies to the carry-over effect as well. The typical objection against this approach is that doubling dose will not necessarily double the response, since if both doses are close to the asymptote of the dose-response curve then similar responses will be generated. However, it may be the case that the middle part of the dose-response

curve is studied and our three doses have been selected from that range. Under these circumstances, the proportionality argument on treatment and carry-over effect may be valid. This type of model will be referred to as "proportional" carry-over model in the sequel.

5.3 Optimality Criteria

For the comparison of two treatments, minimizing the variance of the treatment estimate is of direct interest to all parties involved in the study. There are always parameters of secondary importance (nuisance parameters), the estimation of which affects to a smaller or a larger extent the precision with which treatment effect is estimated. One such parameter is the carry-over effect, which affects the mean of our response. Another example is the variance components describing second order properties of a subject's repeated measurements vector. Adjusting for all secondary parameters is a typical precaution taken in all experimental design exercises.

The choice of optimality criterion becomes more laborious when more than two treatments are compared. Recall that if three treatments are studied, we are interested in estimating as precisely as possible the pair-wise differences $B - A$ and $C - A$, where A is the standard treatment and B, C the newly proposed therapies. The criterion used to decide the best design is that of D_s-optimality, in which the determinant of the relevant part of the variance-covariance matrix of the fixed model parameters is minimized. This criterion is a variant of the D-optimality rule, widely used in applications. Following Atkinson and Donev's or Fedorov and Hackl's notation (see [1],[13]), if we denote by $I(\xi)$ the variance-covariance matrix (inverse of the information matrix) of the fixed parameters in our regression model for a given plan ξ, then this matrix could be partitioned as follows:

$$\begin{pmatrix} I_{11}(\xi) & I_{12}(\xi) \\ I_{21}(\xi) & I_{22}(\xi) \end{pmatrix} \tag{5.1}$$

where $I_{11}(\xi)$ is the variance matrix for the treatment contrasts of interest. The objective is to find the plan ξ that minimizes $det\,(I_{11}(\xi))$. This is the rule used

throughout for the derivation of the best design plans presented in subsequent sections.

An alternative criterion is the one that minimizes the average variance of the pair-wise treatment contrasts. This is a special case of what is known as L-optimality criterion, a variant of the A-otimality rule. According to the L criterion designs that minimize the variance of linear combination(s) of the model parameters are declared optimum.

The previously presented criteria are the most widely used for practical applications. Many more criteria are available in the literature serving different purposes. The estimated variance-covariance matrix is the key element in all these optimization exercises. Both D and A-optimality criteria can be expressed in terms of the eigenvalues of the variance-covariance matrix. The D criterion is the product, while the A criterion is the sum of these eigenvalues. In both criteria the covariance between parameters of primary interest, as well as the covariance between primary and secondary model parameters, are taken into consideration during the optimization process. The geometric interpretation of the D criterion is to provide the experimenter with a confidence region of minimum content for the treatment contrasts of interest, while the A optimality criterion is mostly concerned with the length of the axes of that confidence region. It should be noted the A criterion gives more flexibility to the experimenter regarding the degree of interest he/she places to the various treatment comparisons (see, Jones and Donev [37]). For example, in a three-treatment cross-over trial minimizing the weighted average of the pair-wise treatment variances, with weights determined by the experimenter's interests, could lead to a different optimum plan, when compared to the situation where the average of pair-wise variances is minimized.

A computational note is in order. The results presented in subsequent sections have been derived by a full search over the design family where optimality is claimed. Some discussion of other search methods for finding best plans will be provided at the end of this chapter. Optimum experimental design results depend heavily on the assumed model and particularly on the type of carry-over for cross-over trials. Optimality criterion is a further dimension affecting our final decision. In this chapter, the sensitivity of results in accurately estimating

treatment contrasts of interest to changes in the type of carry-over assumed, is studied.

5.4 Two-treatment results

The majority of the cross-over examples reported in the literature deal with the comparison of two treatments. The typical 2x2 plan is widely used for that purpose, which explains away the fact that most research effort has been put into the study of that design. However, Ebbutt (see [9]) reports the results of a three-period cross-over experiment for the comparison of two treatments in asthma. Drug development sponsors have shown a strong interest in running two-sequence multi-period designs for comparing two treatments.

In that section designs that made up of dual sequences are only considered. Equal number of subjects are allocated in each sequence. Results in which different treatment sequences receive unequal number of subjects are provided in Laska and Meisner (see [51]), for relatively small design families. Optimum plans in two, four and six sequences are only presented. Four sequence designs are generated by different pairings of two-sequence plans. In a similar fashion by joining together two-sequence designs in triplets, six-sequence plans are produced. In order to make clearer the number of switches between A's and B's, design families with long treatment sequences are considered. For the two, four and six sequence plans up to ten, eight and six periods respectively are studied. Note, that three and four period designs will not bother us in what follows.

The computational effort needed to extend these results to larger families increases exponentially as the number of periods and/or sequences gets larger. To illustrate the point, suppose that cross-over designs in p periods are compared. There are $\kappa = 2^{p-1} - 1$ possible design plans. If combined in pairs, $\binom{\kappa}{2}$ four sequence designs are generated, while $\binom{\kappa}{3}$ six-sequence designs produced, if combined in triplets. To get a feel for the computational burden involved in the search for the optimal plan as the number of sequences and/or period grows, the following table provides the number of distinct designs for various combinations of sequences and periods. Bolded are the families for which optimal plans are

provided at the end of the chapter.

Table 5.1: Number of distinct designs

Sequences/Periods	5	6	7	8	9	10
2	**15**	**31**	**63**	**127**	**255**	**511**
4	**105**	**465**	**1953**	**8001**	32385	130305
6	**455**	**4495**	39711	333375	2731135	22108415

5.4.1 Practitioners's favourite model - No carry-over scheme

Most cross-over studies have been analyzed using that model, but it is not clear to me if this model has also been used for choosing the best plan as well. When there are no residual terms into the model, the optimum plan does not depend on the value of the correlation coefficient.

Looking at Table (5.2) the optimum design utilizes sequences in which switches between A's and B's are as frequent as possible, in fact the maximum number of switches occur in these designs. The design efficiency is unaffected by the correlation coefficient at the absence of any residual terms, although under anyone of the other carryover schemes higher treatment efficiency is achieved the closest the repeated measurements on each subject are collected. This may cut short the time a trial lasts but one has to bear in mind that if successive measurements are close in time, then the possibility of carry-over being present increases substantially, though the type of residual activity would be difficult to identify. Under the no carry-over scenario the fact that in the two sequence optimum plans frequent exchange between alternative treatments occurs, necessitates the need for formal testing for the presence of residual effects. Although regulatory bodies have recently argued unfavorably to the use of the 2x2 plan as lacking power for detecting carry-over effects (see Wang and Hung [90]), this should not be the case when more than two periods are used. Carry-over effects are now estimated with higher precision than in the 2x2 case, since within-sequence information is utilized. The last argument is in favor of using longer sequences when two treatments are compared, instead of avoiding running a cross-over experiment

altogether.

The optimum four and six sequence plans are made up of the best two sequence plan in conjunction with sequences, where again frequent switches between the two treatments occur. Note that if the experimenter decides to divide the trial time into six rather than five sub-period intervals then higher treatment precision is achieved. Similar argument is true if designs with six rather than four sequences are used. In other words by making our trial bigger, in either direction, more information is collected concerning efficacy of the two therapies and that would inevitably lead to more accurate treatment estimates. In the extreme scenarios the number of sequences ranges from one to the number of patients recruited to participate in the study. Similarly the number of periods could be made arbitrarily large by sub-dividing the trial-time into small time windows. The running of such studies is not recommended on financial grounds but also on difficulties concerning the management of large groups of people for long time.

5.4.2 Naive approaches for modeling residual activity - Simple Carry-over

This model has caused too much controversy in the cross-over literature, although carry-over effect is not really a major problem in cross-over trials (see Senn [76]). The additive carry-over term is difficult to be justified using pharmacological argument. Suppose in a typical trial the half life of the active compound is known to be T time units from Phase I studies. The clinical team has to make sure that the patient will be treated at time intervals of length at least $2T$, so that the possibility of carry-over being present diminishes. These scheduled visit arrangements are based purely on scientific reasons. However, it may be the case that the patient may scheduled his next visit in less than $2T$ time units. In that case, the carry-over effect to the next visit would be a known proportion of the previous-visit treatment effect and could be modeled as such. To keep track of the various patients visits and appropriately adjusting for any residual effects, would be an enormous task hardly affecting our treatment estimate. As a consequence the typical analyst models carry-over activity (if any) with a simple additive term, which most of the time is proved statistically unimportant and removed

from the model. We have already modeled residual effect as a proportion of the treatment one, and it seems to be the case that the non-linear model recovers treatment effect even under unrealistic carry-over assumptions.

Should this model being used for designing a study though? Clearly not in the 2x2 case, since imprecise first period data will only be used to generate a much larger sample size than it is actually needed (see Brown [3]). This may not be the case with longer sequences though. Results concerning optimum plans are presented in Tables (5.3)-(5.5). Correlation coefficient plays a minor role in deciding the best plan to go for. The number of switches between A's and B's are not that frequent as they were when no residual effect was present. In fact switches to the alternative therapy need only to be made every third measurement, instead of every second measurement as it was the case in the "no carry-over" scenario. For specific combinations of sequences and periods the number of optimal plans the experimenter has to choose from to run his trial, increases with the correlation coefficient. For example, in the 7-period 2-sequence family there is only one good plan when the correlation between successive measurements is low, while the number of optimum designs increases to three for medium or large values of ρ. Similarly for the 8-period 4-sequence family, the number of best designs for large values of ρ is twice as high as the number of designs when ρ takes values in the small/medium range.

5.4.3 Pharmacology matters - Fleiss carry-over

Under this type of carry-over scenario, the optimum plan depends heavily on the value of the auto-correlation coefficient ρ. The interesting result (see Table (5.6)) comes in the situation where repeated measurements on a subject are nearly uncorrelated ($\rho = 0.2$). The best treatment sequences are made of a long series of A's followed by a long series of B's. For instance, the subset of optimum plans in p-period cross-over families is made of p_1 series of A's followed by $p - p_1$ series of B's, where $2 \leq p_1 \leq (p-1)$. In these designs the minimum number of switches (i.e. one) occurs between the two competing therapies. Interestingly enough the four and six sequence best plans consists of all the possible combinations of two-sequence plans in pairs and triplets respectively.

Reasons have already been provided elsewhere why the use of such treatment sequences may be inappropriate to run a study. Firstly, if one of the two therapies is placebo, then it will be unethical to keep a patient untreated for such a long time. Secondly, the proposed treatment sequences and their duals are equivalent to an AB/BA design, the inefficiency of which have well been explored in the cross-over literature. To demonstrate the last point, suppose that an eight-week cross-over trial will be used to run a study. If the statistician believes that the most likely form of carry-over to be present is the Fleiss one, then one of the designs he could propose for running the study is the AAAABBBB and its dual. Repeated measurement will be obtained at weekly intervals. This is exactly the same as running an AB/BA design, where the total study-period has been divided into two four-week sub-periods. In the AB/BA case, measurements may be obtained on each subject on a weekly basis, but an appropriate summary for every four consecutive measurements, e.g. the mean, will be used as the analysis variable.

The real question is if there are any experiments met in practice where treatment sequences, like AAAABBBB, are used. The answer is affirmative. The multiple-dosage regimen studies are good examples, where prolonged therapeutic activity is sought in order to achieve maximal clinical effectiveness. In these studies, drugs are released into the body by intravenous (IV) infusion at a constant rate. A loading dose (or bolus dose) usually precedes the IV infusion in order to obtain steady state concentrations as quickly as possible. If only one IV dose is administered, the time required to reach the steady-state drug-concentration in the plasma depends on the elimination rate of the drug from the body, but also on the half-life of the compound. For most drugs, the estimated time to reach 99% of the steady-state drug concentration after a single IV infusion is 6.6 half-lives. To get a deeper understanding for the drug concentration in plasma as a function of time after a single IV dose, the corresponding equation looks as follows (see, Shargel and Yu [84]):

$$C_p(t) = \frac{R}{V_p k}\left[1 - \left(\frac{k-b}{a-b}\right)e^{-at} - \left(\frac{a-k}{a-b}e^{-bt}\right)\right] \tag{5.2}$$

where a, b are known constants, R is the rate of infusion, k is the overall elimination constant and V_p is the volume of drug in plasma. It has been assumed that

our drug follow a two-compartment kinetic model. As time goes by (i.e. $t \to \infty$) then the steady state concentration is:

$$C_{ss} = \frac{R}{V_p k} \tag{5.3}$$

The above equation can predict at any time after the start of the IV dose the plasma drug concentration. In multiple-dosage regimes studies, it is assumed that earlier doses will not have an effect on later ones. This is the principle of superposition and essentially it makes sure that the pharmacokinetic profile of the repeated doses remains the same throughout the study. Obviously the size of the dose and the dose interval have been determined in such a way that the drug level in the blood increases at the end of each dose. In summary, the total plasma drug concentration would be equal to the sum of the residual drug concentrations of all the previous doses. The amount of drug in the body will increase and finally will reach a plateau. One should not wrongly assume that the steady-state drug concentration remains constant for the whole study period. In fact, depending on the type of the compound, it can fluctuate considerably between two values, which for sake of argument will be referred to as C_{min} and C_{max} in what follows. It is exactly this fluctuation that can generate the Fleiss and the Mixed types of carry-over, discussed in this and the next section respectively.

In a typical multiple-dosage study, which most of the time is a 2x2 cross-over study comparing equal doses of a test and reference products, patients usually maintained on the drug since the use of a wash-out period could place them at substantial risk. The patient continues on his own medication and blood samples are repeatedly collected at equal time-intervals. Once this process is completed the patient switches to the alternative therapy, where time is again allowed for the compound to reach its steady-state. Assume now, that the plasma drug concentration fluctuation at steady state is different for the two therapies. In that case, if multiple doses of a compound are followed by multiple doses of the same compound, then carry-over effect could be safely assumed to be negligible especially if the difference $C_{max} - C_{min}$ is small. But if repeated doses of the first therapy (A) are followed by repeated doses of the second one (B), then what remains from A may play a crucial role in deciding the response of B at steady state. One may argue that adequate time must be allowed, so that by the time

the second product reach his plateau level therapy A is completely eliminated from the body. However, one has to remember that the two therapies may react with each other and complete elimination of anyone of them may not be reached within the trial period.

The above reasoning justifies to some extent the point that using Fleiss carry-over model for designing purposes is a viable possibility, although considerable input may be required from the Phase I clinical team. The discussion for multiple-dosage studies was motivated by the fact that treatment sequences made by long series of A's (or B's) are optimum under the Fleiss carry-over model, when ρ is small. This does not seem to be the case when ρ lies in the middle or high range of its plausible values (see Tables (5.7)-(5.8)). If ρ lies in its middle range, designs made up of relatively short sequences, with three on average consecutive repetitions of A's (or B's), are the optimum ones. Most of these plans are optimum or have high efficiency under the simple carry-over scheme as well. When repeated measurements on the patients are highly correlated, then optimum plans are made up of even shorter sequences of the two therapies. Once more the proposed plans have excellent properties under the simple carry-over model. For a specific combination of sequences and periods the number of optimum plans decreases as ρ increases under the Fleiss carry-over model, contrary to the simple carry-over scheme.

In conclusion, designing a study with Fleiss carry-over in mind is equivalent to designing a study assuming that simple carry-over scheme applies, for the majority of the ρ values encountered in practice. For small values of ρ optimum plans are essentially multiple-dosage studies, widely used in bio-equivalence applications. Justification of Fleiss carry-over in practice is difficult, if not impossible, but examination of its applicability leads to a deeper understanding of the compound's activity in the human body.

5.4.4 Further pharmacology in action - Mixed Carry-over

This type of residual effect has already been introduced in the previous chapter as an intermediate scenario between the simple and the Fleiss type of carry-over. Recall that under this scheme, a treatment carries over to itself only a proportion

of what carries over to alternative therapies. This proportion is assumed constant for all treatments under study. If that proportion (ϕ) is high, then the simple carry-over model is recovered, while for small values of the same parameter the Fleiss carry-over model is retrieved.

A similar reasoning could possibly justify the mixed residual effect, although the introduction of an extra parameter to describe residual treatment activity makes the use of such a model, for both analysis and design purposes, difficult. Such a residual effect could possibly be met in multiple dosage studies. Suppose that the two therapies we are about to compare are compounds with similar pharmacological properties; a good example is when a low and a high dose of the same compound are compared. A direct consequence of that assumption could be that the therapeutic windows of the small dose is contained within the therapeutic window of the higher one. Suppose that when the higher dose precedes the lower one, then the carry-over effect is 10%. A fraction of that 10% would remain as residual effect when the order of administration for the two therapies is reversed, since the lower dose has a narrower therapeutic window compared to the higher one.

Comparing now the family of optimum plans for the mixed carry-over model when $\phi = 0.2$ (see Tables (5.9)-(5.13)) with the corresponding family for the Fleiss model, there seems to be some difference between the two families for small values of ρ. Recall that in the Fleiss model best treatment sequences were made up of long series of A's followed by long series of B's, in contrast to the mixed type model with $\phi = 0.2$, where the two-sequence plans contain frequent switches between the two therapies. However, in the four/six sequence optimum plans sequences that contain three or even four consecutive repetitions of the same therapy may be found. For other values of ρ the two models seem to propose designs which are not identical but have similar structure.

Comparing now designs among the simple carry-over model and the mixed model with $\phi = 0.8$, it seems to be the case that designs which are optimum in one model are also highly efficient under the alternative model. This finding is consistent over the whole range of ρ values. In summary, designing a cross-over study with the simple carry-over model in mind gives robust answers to model mis-

specification, if that model mis-specification is adequately described by the mixed model. This is not true for the Fleiss model. Note that the number of best plans under the mixed model for all values of ϕ and ρ is quite limited (usually one or two), contrary to the Simple and Fleiss model. Our purpose here is rather to explore similarities in the structure and efficiency of the best designs from the one end of the spectrum (Fleiss model) to the other end (Simple model), rather than using mixed model for designing a study.

5.5 More than two treatments

It is quite common in clinical trials to set-up a study for the comparison of three or more treatments. The analysis of a seven treatment cross-over study, for the comparison of three formulations of two asthma drugs and placebo, has already been presented.

Another example of such trials can be found in the pharmaceutical industry and it concerns the testing of combination of drugs. Common therapeutic area of application is HIV trials. Combination trials are set-up to explore how two or more factors affect a clinical response (see, Fletcher et al [19]). In the simplest situation where a low and a high dose of two drugs are considered, the four factorial combinations can be tested on each subject in four successive treatment periods. This calls for the use of a cross-over design for running the study. Medical researchers use the simple carry-over model for designing and analyzing combination studies. This is another example where residual effects at time $T+1$, if present, should depend on the treatments administered at times T and $T+1$. It is questionable if simple carry-over model is appropriate for modeling such residual effects, but even the Fleiss one may not be suitable for tackling the problem. The introduction of distinct carry-over terms depending on the order of treatments administered may be the appropriate course of action. For example, if we label the four combinations by A, B, C, D, and the carry-over from A to B is denoted by λ_{12}, then carry-over from B to A is λ_{21}, where $\lambda_{12} \neq \lambda_{21}$. Introducing a large number of residual terms may lead to parameter identification problems at the analysis stage. Appropriate parameter restrictions on these residual terms, based

on clinical knowledge, may overcome such problems.

With factorial experiments, like combination trials, the testing of treatment by treatment interactions is commonly reported in practice, regardless of any assumptions concerning residual terms. The analyst could also check the statistical significance of any carry-over by carry-over interactions, since in multiple sequence/period trials such terms are estimable. The testing of any important treatment by carry-over is usually overlooked. It is a similar situation to the one where the statistician includes fourth order terms in his linear model, without including third orders ones. Although in practice the presence of the above interactions is extremely unlikely, any model with carry-over by carry-over interactions should include treatment by carry-over ones as well.

5.6 Three treatment results

The design families considered in this section have at least three sequences and three periods, so that in a sxp arrangement each treatment occurs at least once in each row and at least once in each column. The within-subject covariance structure is AR(1) with $\rho = 0.7$ throughout. The sensitivity of results to departures from the chosen value of ρ is not studied, since this value is commonly met in practical applications.

Recall that we concentrate on the simultaneous comparison of several new therapies (B,C,...) to a control therapy (A) using the D-optimality criterion. Alternatively we might be interested in efficiently estimating all pair-wise treatment comparisons, although this will not be the case in what follows. For three-period designs, D-optimum plans are selected by performing a detailed search over the full listing of all possible distinct designs for that family. Due to the computational burden involved as the number of periods and sequences grows, in families with more than three periods our search has been restricted to the distinct cyclic designs for that family. Recall also that five carry-over schemes are studied: No carry-over, Simple, Fleiss, Simple2 and Proportional. In the Simple2 scheme one of our treatments (A) is Placebo with no treatment or residual effect. In that case our interest is focused on the comparison between the standard therapy (B)

and the new proposed treatments (C,...). Under the Proportional scheme multiple doses of a compound with proportional treatment and residual effects are administered to the study participants. Note that if the number of periods equal the number of treatments studied, then Fleiss carry-over does not apply, since each treatment should appear at least once in each treatment sequence implying that no treatment replication occurs.

In the three-period, three-sequence family there is a variety of good plans under the simple carry-over scheme (see Table (5.14)). Under the "Simple2" and "Proportional" type of carry-over, optimum plans are identical with equal variances. These designs are optimum even when no carry-over terms included into the model, but with 90% lower variance for the estimation of contrasts of interest. This implies that in order to achieve a given treatment precision level, fewer patients need to be recruited under the model with no residual terms, compared to any model that contains carry-over effects.

By extending the number of sequences while keeping the number of periods fixed, the number of distinct designs is reduced. This makes easier the task for selecting a good plan, since the computational effort required is reduced. In the four-sequence three-period plans, not only the variance of the treatment contrasts reaches its lower value under the no-carryover model, but also the number of available plans under that scenario is at least twice as high as the number of plans under any model with residual terms. Note in passing that there are six treatment sequences in three periods, as a result of which only six five-sequence three-period plans exist. All of these plans are equally efficient for designing a study under the simple and the no-carryover scheme, while only five of them can be used under the two alternative residual-effect patterns.

The reader may wonder why the efficiency of designs with more sequences than periods are considered. It is the case that clinical trials are conducted in many different large recruitment centers all over the world. Assigning a treatment sequence to all subjects of a specific center is common practice. Under that scenario center and sequence effects are not separately estimable, the problem can be overcome by assigning more than one treatment sequence to the patients of any center. In other words, practical needs require the rate at which sequences

grow to be higher than the corresponding rate for periods and this necessitates special attention to designs with $s \geq p$.

5.7 Cyclic Designs

Suppose now that patients are scheduled to come to the clinic in six visits, although only three treatments will be tried on them. Due to the large number of distinct designs that can be chosen to run the study, attention will be restricted to cyclic designs. Although no formal mathematical proof has been given, a cyclic design must exist in the subset of best plans under any carry-over effect scenario. This conjecture is true in the three-period family but it is unclear if it can be extended to families with more than three periods. Generally speaking, cyclic plans tend to be highly efficient.Computer generation of such plans is straightforward. For example, a three-sequence cyclic design in three treatments and p periods can be generated from an initial treatment sequence by adding one and two to each element of that sequence and reducing modulo 3 when necessary. Up to eight period plans have been studied in three treatments. In that way the set of distinct treatment sequences is divided to mutual exclusive and exhaustive sub-families made of triplets of treatment sequences. The computational effort to search over the cyclic sub-families instead of searching over the range of all possible triplets is reduced considerably. In the presentation of results, only the initial sequence of the best cyclic plans are displayed. Incomplete block designs are good examples of designs produced by cyclic generation of an initial sequence. One can easily generate incomplete block designs for any number of sequences, periods and treatments. Special restrictions have to be imposed to the above parameters in order to get a balanced incomplete block design, in which pair-wise treatment comparisons are made with the same accuracy.

Further properties of cyclic designs can be found in John and Williams (see [36]). Mathematically speaking, when no residual terms are included and the error terms are uncorrelated with zero mean and constant variance, cyclic designs are attractive because both the information matrix and its inverse can be expressed as a linear combination of circulant matrices. A circulant is a symmetric matrix

having 1 in one of its minor diagonal and zero elsewhere. The eigenvalues of such matrices can be written down explicitly, allowing the variances of pair-wise treatment contrasts to be expressed analytically. This facilitates the task of deciding the properties a plan should possess in order to be optimum, but unfortunately it does not pick-up a good plan for the person who designs the study. A further key property of cyclic designs is the special form the concurrence matrix can take. This is a txt symmetric matrix (t being the number of treatments compared) with its $(i,j)^{th}$ element equals the number of sequences the treatment pair (i,j) appears.

Cyclic plans can be further classified and the concurrence matrix can take a special form in each case. One example are resolvable block designs, which are incomplete block designs where treatment sequences can be grouped so that each treatment appears once in each group. This type of designs can be quite helpful in multi-center studies, since groups of treatment sequences could be assigned into different centers. This implies that even in the scenario where some centers withdrawn from the study all treatments will have occurred equally often. Resolvable designs are good examples where recovery of inter-block information could result in more efficient treatment estimates. Another category of cyclic plans are the row-column designs. In such plans, the number of times each treatment can appear in each row/column can vary. In the special case where each treatment appears once in each row and once in each column the row-column plan is called Latin square. All cross-over experiments can be seen as row-column plans with row representing sequences while columns periods. Treatments comparisons are available form both rows and columns, but its the comparisons made within rows and columns that are expected to be of highest precision. When no residual terms included into the model, treatment effects can sometimes be independently estimated of any row/column effects. When residual terms of any sort are included, the orthogonality property is lost in the majority of the cases. Recall that lack of orthogonality between treatment and carry-over effects in the 2x2 case is the main reason for the deficiency of the two stage procedure.

Back to our results for three treatment plans, where three-sequence in more than three-period designs are examined (see Table (5.15)). A striking feature is the

fact that when residual terms excluded from the model, the number of best plans available for design purposes are much higher compared to the corresponding number of plans under models that include carry-over term of any kind. In addition under Simple2 and Proportional carry-over schemes, optimum plans are identical with the same efficiency. In these plans the design structure is quite interesting; for example the six period best plans are made of replicates of three period plans. For the Fleiss carry-over model there is a frequent exchange between the three treatments, on the contrary under simple carry-over model best treatment sequences are made of short successive repetitions of the same treatment. In fact, under the simple carry-over model, a six period plan is made of a three period plan followed by the same plan in reversed order. Furthermore the difference in efficiency between any model than contain residual terms and the model with no carry-over terms decreases with increasing number of periods used. Generally speaking, good plans under the model with no residual terms are not optimum under models with carry-over terms, in other words proposing robust solutions when analysis model is mis-specified becomes a difficult task. Obviously these observations generate hypothesis for future research. Finally, it has been reported that under the simple carry-over with additional restrictions in the design structure and model assumptions, a $t+1$-period design can be made of a t-period optimum plan by repeating the treatment of the last period. This rule does not seem to be justified in our case. Further research may be needed to find out under which circumstances the previous statement is true.
Practical experience suggests that routine follow-up can be easily implemented and it is not as costly as patient recruitment. In conclusion designs with many periods could be the future of cross-over trials. Worth noting that reporting of cross-over studies with few periods but with repeated measurements collected within each period is frequently met in practice.

5.8 Four, five and six treatment results

The use of cross-over plans for the comparison of more than three treatments is not currently favored neither by sponsors nor by regulatory authorities. Dose-

ranging trials are good examples of experiments where a large number of treatments are tested in a single trial. Such studies are usually conducted at later stages of drug development in order to determine the clinical effectiveness of a series of doses and appropriate adjustments to current dosage regimes made where necessary. Surprisingly enough, parallel designs are used for running such studies, implying that information on individual dose-response parameters is not recovered. This information will be available if a cross-over plan had been used instead, where several doses are tested on each subject. The problem with a cross-over study for dose ranging is not that of carry-over but rather the dose timetable. This timetable has to be chosen so that toxicity problems are avoided. Sheiner et al (see [85]), describe simulation studies where at each period the dose level is increased as long as the response remains above a threshold level and there is no toxicity. During the first period placebo is administered to all subjects. Obviously these strict guidelines not only reduces the number of available plans, but also decreases the number of observations offered from each subject. Modern statistical methodology can easily predict missing values of follow-up (i.e. measurements typically collected if we had continued to monitor the subject for the entire study duration), but in addition these predictions can be used for drawing inference about population parameters.

Despite the questionable usefulness of these studies, due to the practical limitations discussed above, results will be presented for cyclic families only. For the four treatment comparison, designs up to seven periods are tested (see Table (5.16)). The initial intention was to generate and compare up to ten-period cyclic plans, but the upper bound of available computational memory was soon reached, and the idea abandoned. Nevertheless, insight into the structure of optimum plans when more than seven periods are used, can still be gained. As ever, treatment effects are more precisely estimated when residual terms of any kind are excluded from the model. A nice property of the four-period, four-sequence family is that the same set of designs are optimum irrespective of any assumptions made concerning the carry-over effect. In the five period plans the above property is valid under all carry-over schemes, but a different set of designs are optimum when presence of carry-over is ruled out. When we move on to six periods, the

simple and Fleiss carry-over model offer identical and equally efficient plans for running a study. This is evidence supporting a conjecture made by Matthews (see [65]) that optimum plans under simple carry-over model are usually good choices under the Fleiss model as well. This property is also justified for the six and seven period families. Generalization of that conjecture to p-period families for comparing four treatments is worth investigating. Six and seven period families, offer one of the few occasions where the Simple2 and Proportional schemes provide us with different solutions for designing a study. The Simple2 scheme seems to propose the same designs, as Simple and Fleiss models do in the 6-period family. This is not the case for the seven period family though.

Once more computational restrictions did not allow the study of long treatment sequences when five or six treatments are compared (see Table (5.17)). All carry-over scenarios seem to agree on the set of best plans, apart from the Proprtional scheme. Also the Proportional scheme offers a limited number of solutions compared to the other carry-over scenarios. As has been noted in other occasions, the number of proposed plans when carry-over terms are not included is at least twice as high as the number of plans when carry-over terms included. In conclusion some interesting hypotheses have been generated by comparing plans in families where more than two treatment compared, although it is a difficult task to provide theoretical justification for these hypotheses.

5.9 Non-linear Designs for two treatments

Cross-over trials have been widely used for the comparison of hypertension or asthma drugs. In such studies wash-out periods are not allowed for ethical reasons. Under these circumstances, presence of residual effects are likely. In the majority of cross-over studies treatment periods are usually long time windows. So, depending on the time each measurement is collected, the carry-over effect to the next period, is usually a proportion of the treatment activity in the current period. For the sequence AB, which might be part of a longer treatment sequence, carry-over from A to B can be written as $\lambda_A = \tau_A \rho_A$ and similarly from B to A is $\lambda_B = \tau_B \rho_B$. This is the Simple carry-over model with two additional non-

linear terms. In the sequel it will be assumed that $\rho_A = \rho_B$. Data have already been analyzed using this model and it has been shown that treatment effect can truly be recovered irrespective of inclusion or not of any residual terms in the model. The assumption that the two treatments carry-over the same proportion can be defended on the basis that similar pharmacological properties govern the therapeutic activity of two beta blockers, two ace inhibitors, or any other agents belonging to the same group from a pharmacological point of view.

A further assumption made, without explicitly stated, is that proportion of treatment persisted to the next period is assumed constant throughout the whole study duration. Violation of that assumption can frequently be met when different doses administered at different treatment periods. Higher doses expected to carry-over more than lower ones. But even if the same dose is administered for each compound (say 10mg), fluctuation around this value (e.g. overdose) may result to different proportion of treatment persisted to the next period.

Recall in the 2x2 case that the treatment effect estimate is biased by half the difference of the unknown residual effect ($\lambda = \lambda_A - \lambda_B$) between the two treatments. Introducing the non-linear term makes this bias dependent on the unknown treatment effect ($\lambda = \rho\tau$). The consequence of that assumption for designing a study is minimal. In fact, for two treatment comparison and when two sequence designs are considered, results are similar regardless if residual term is a non-linear function or completely unrelated to the treatment effect. This result cannot be extended in the four and six sequence design families. For example, in the seven-period four-sequence family, under the simple carry-over model in which carry-over is modeled using a non-linear term, the optimum plan is (ABBAABB, ABBAABA, duals). Similar argument holds for the Fleiss carry-over model.

It is also true in the non-linear case that the efficiency of the various plans do not depend on the proportion of treatment persisted to the next period. Obviously things would change if the assumption $\rho_A = \rho_B$ is removed. But how easy is to assume otherwise? For designing a study plausible values of ρ_A, ρ_B have to be provided. Physicians are usually unaware of such information, though results from previous studies could help in the derivation of any unknown quantities. Uncertainty of this kind can also be incorporated into the design problem by

using the Bayesian approach, although implementation of this method can be difficult. This is because priors imposed on unknown parameters affect results to some extent. Robust designs to the choice of prior are desirable. Finally the assumption $\rho_A = \rho_B$ does not make much sense if more than two treatments are compared. This is why there is a lack of research attention to this kind of problem.

The mathematics for tackling the non-linear design problem is a direct extension of the linear approach. More specifically the choice of best design depends on a number of unknown parameters. In our case, this is the unknown treatment effect τ and the proportion of treatment ρ that carries over to the next period. It is assumed that $\tau = 2.5$ and ρ ranges from zero to one. The nonlinear mean response can be expressed as follows:

$$E\left(y_{ijk}\right) = \mu + \tau_{d(i,j)} + \tau_{d(i,j-1)}\rho \tag{5.4}$$

where the mean μ includes overall mean, period or any other effects that distinguish among cells of the cross-over plan. A modified version of this equation has already been used at chapter 3 and explanation of the $d(i, j-1)$ and the other subscripts is given there. To derive an expression for the dispersion matrix of contrasts of interest, the matrix of partial derivatives of the mean response equation with respect to all parameters needs to be evaluated at the selected values of ρ and τ. Because all other terms, apart from ρ and τ, enter linearly into the model, specification of other parameters is not necessary. The partial derivatives of interest are:

$$\frac{\partial E(y_{ijk})}{\partial \tau_{d(i,j)}} = 1 + \rho\frac{\partial \tau_{d(i,j-1)}}{\partial \tau_{d(i,j)}} \tag{5.5}$$

$$\frac{\partial E(y_{ijk})}{\partial \rho} = \tau_{d(i,j-1)} \tag{5.6}$$

In case where the two treatments carry-over a different proportion to the next period the mean equation can be written as:

$$E\left(y_{ijk}\right) = \mu + \tau_{d(i,j)} + \tau_{d(i,j-1)}\rho_{d(i,j-1)} \tag{5.7}$$

and the partial derivatives with respect to $\tau_{d(i,j)}$ and $\rho_{d(i,j)}$ are:

$$\frac{\partial E(y_{ijk})}{\partial \tau_{d(i,j)}} = 1 + \rho_{d(i,j-1)} \frac{\partial \tau_{d(i,j-1)}}{\partial \tau_{d(i,j)}} \tag{5.8}$$

$$\frac{\partial E(y_{ijk})}{\partial \rho_{d(i,j)}} = \tau_{d(i,j-1)} \frac{\partial \rho_{d(i,j-1)}}{\partial \rho_{d(i,j)}} \tag{5.9}$$

Note that mean response expressed by equation (5.4) is a special case of the mean response model described by equation (5.7). Model (5.7) covers all possibilities, i.e. the carry-over effects of the two treatments may be equal (equation (5.4)), or related in some mathematical way, or completely unrelated. Results when the mean response is described by equation (5.7) are not given due to the lack of information concerning $\rho_{d(i,j)}$. Contrary, results for the first scenario (equation (5.4)) are straightforward to derive and have already be presented.

Assume there are n study participants, in a p period cross-over study. Assume further that our model has k unknown parameters (in our case $k = p + 2$). Once the vector of partial derivatives of the mean response with respect to every unknown parameter has been evaluated, these vectors are joined together in a $npxk$ matrix, denoted as X_β for sake of reference. The variance matrix is simply

$$V = \left(X_\beta^T \Sigma^{-1} X_\beta\right)^{-1} \tag{5.10}$$

where Σ is a block diagonal $npxnp$ matrix, each block being an AR(1) type correlation matrix. The variance matrix for any set of linear/nonlinear contrasts (a good example of a non-linear contrast is the overall treatment effect $\tau + \rho\tau$) is given by:

$$V_A = AVA^T \tag{5.11}$$

where the i^{th} row of A is a vector of partial derivatives of the i^{th} contrast with respect to the unknown model parameters evaluated at specific values of these parameters where necessary.

5.10 Computational approaches in searching for optimum plans

It has already be mentioned that current computing limitations do not allow fast detection of the best cross-over design for any number of sequences and

periods. In the results presented so far, we had to confine our search to special sub-families (cyclic plans), but even then only for moderate values of the number of sequences/periods exact results were derived. Special numerical algorithms have been devised in order to tackle the high-dimensional optimization problem, in the case where p (number of periods) and s (number of sequences) are large. Our problem can be expressed in a straightforward manner: in a set of s 1xp vectors, select the best subset consisting of l such vectors (l can vary from 2 to $s-1$). The winning subset will be the one that minimizes some function defined by the needs of the experimenter. An initial attempt to solve the problem is to identify a smaller group of highly efficient designs and then search within that sub-family. Theoretical results can limit the number of designs under consideration, but usually easily programmed counting rules are more efficient in finding plans worth further attention. An example of such a rule is to minimize the sum of squares

$$\sum_i \sum_j \lambda_{ij}^2 \tag{5.12}$$

where λ_{ij} is the number of treatment sequences containing both i^{th} and j^{th} treatments. Note though, that this rule may produce a large sub-class of plans. Other rules need to be implemented in that sub-class so that further reduction in the candidate design set is achieved.

Once the experimenter defines both the dimension of the problem (i.e. p and s), and the sub-family of plans selected from his screening procedure is deemed appropriate, then he can either do a full search or use one of the interchange (or exchange) algorithms to find the best plan. Assume for the moment that there are N_{cand} candidate treatment sequences and we are interested for the best triplet. Any exchange algorithm usually start the search from a design that is optimum in a family with fewer number of periods than the family we are interested in. This may not necessarily be the case and the starting point can be a design randomly chosen from the family under consideration. The next step of the algorithm is to improve the starting design by exchanging treatment sequences of that design with treatment sequences that they belong to the candidate set but they are not included in the starting design. The first sequence is exchanged with the one from the candidate set that leads to the greatest reduction in the determinant of

the variance matrix for the contrasts of interest. The same process is repeated for the second and third sequence of the starting plan and at the end of that cycle a new starting design is proposed. The process starts all over again and at the end of each cycle a new starting point is reached, far more efficient than the one recommended at the beginning of the cycle. The process terminates when no further exchanges can be made that will improve upon the design at the end of the current cycle.

Various modifications of the above algorithm are available. For example instead of doing the best current exchange we could simply update the design with any exchange that improves its efficiency as soon as it is discovered. Another modification is to accept an exchange that may not improve the objective function with small probability. One should always remember that all these modifications are made in order to increase our chances of locating the global rather than a local optimum plan. There are currently routines available for generation of all possible treatment sequences for any number of periods and sequences and also it is not difficult to program an algorithm for listing all l possible subsets of these sequences. As computational power increases rapidly, exact results can be made possible as s and p grow. This will not replace the use of exchange algorithms but will probably improve their performance as well as the accuracy of results they provide.

Table 5.2: Optimum two-treatment designs. Model: No carry-over

Periods	Designs (Variance $x10^{-2}$ when $\rho = 0.2, 0.5, 0.8$)
	Two Sequence Designs
5	ABABA (7.44, 5.12, 3.75)
6	ABABAB (6.12, 4.16, 3.02)
7	ABABABA (5.20, 3.50, 2.52)
8	ABABABAB (4.52, 3.03, 2.17)
9	ABABABABA (4.00, 2.66, 1.90)
10	ABABABABAB (3.59, 2.38, 1.69)
	Four Sequence Designs
5	ABABA with any of: ABAAB, AABAB, ABBAB, ABABB Variances: 3.95, 2.85, 2.13
6	ABABAB with any of: ABABAA, ABAABA, AABABA, ABBABA, ABABBA Variances: 3.22, 2.27, 1.67
7	ABABABA with any of ABABAAB, ABAABAB, AABABAB, ABBABAB, ABABBAB, ABABABB Variances: 2.71, 1.88, 1.37
8	ABABABAB with any of ABABABAA, ABABAABA, ABAABABA, AABABABA, ABBABABA, ABABBABA, ABABABBA Variances 2.34, 1.61, 1.16
	Six Sequence Designs
5	Define: 5=ABABA, 9=ABAAB, 10=AABAB, 11=ABBAB, 13=ABABB The following triplets are optimal: (5 9 10), (5 9 11), (5 10 11), (5 9 13), (5 10 13), (5 11 13) Variances: 2.69, 1.98, 1.48
6	Define: 5=ABABAA, 9=ABAABA, 10=AABABA, 11=ABBABA, 13=ABABBA, 21=ABABAB. The following triplels are optimal: (5 9 21), (5 10 21), (9 10 21), (5 11 21), (9 11 21), (10 11 21), (5 13 21), (9 13 21), (10 13 21), (11 13 21) Variances: 2.18, 1.56, 1.24

Table 5.3: Optimum two-treatment designs. Model: Simple carry-over. Within-subject error structure AR(1) ($\rho = 0.2$)

Periods	Design 1	Design 2	Design 3	Design 4
Two Sequence Designs (Variance $x10^{-2}$)				
5	AABBA (9.77)			
6	ABBAAB (7.96)			
7	AABBAAB (6.94)			
8	ABBAABBA (5.92)			
9	ABBAABBAB (5.38)	ABBAABAAB	ABAABBAAB	ABBABBAAB
10	ABBAABBAAB (4.73)			
Four Sequence Designs (Variance $x10^{-2}$)				
5	ABBAA			
	AABBA (4.89)			
6	ABBAAB	ABBAAB		
	ABAABB (4.04)	ABBABB		
7	AABBAAB	AABBAAB	AABBAAB	
	ABBAABA (3.43)	ABAABBA	ABBABBA	
8	ABBAABBA	ABBAABBA	ABBAABBA	
	AABBAABA (2.98)	AABAABBA	AABBABBA	
Six Sequence Designs (Variance $x10^{-2}$)				
5	ABBAA	ABBAA		
	AABBA	AABBA		
	ABAAB (3.23)	ABBAB		
6	AABBAA	AABBAA		
	ABBAAB	ABBAAB		
	AABAAB (2.68)	AABBAB		

Table 5.4: Optimum two-treatment designs. Model: Simple carry-over. Within-subject error structure AR(1) ($\rho = 0.5$)

Periods	Design 1	Design 2	Design 3	Design 4
Two Sequence Designs (Variance $x10^{-2}$)				
5	AABBA (8.72)			
6	ABBAAB (6.55)			
7	ABAABBA (5.85)	ABBAABA	ABBABBA	
8	ABBAABBA (4.88)			
9	ABBAABBAB (4.36)	ABBAABAAB	ABAABBAAB	ABBABBAAB
10	ABBAABBAAB (3.90)			
Four Sequence Designs (Variance $x10^{-2}$)				
5	AABBA	AABBA		
	ABAAB (4.23)	ABBAB		
6	ABBAAB	ABBAAB		
	AABAAB (3.40)	AABBAB		
7	AABBAAB	AABBAAB	AABBAAB	
	ABBAABA (2.87)	ABAABBA	ABBABBA	
8	ABBAABBA	ABBAABBA	ABBAABBA	
	AABBAABA (2.49)	AABAABBA	AABBABBA	
Six Sequence Designs (Variance $x10^{-2}$)				
5	ABBAA	ABBAA		
	AABBA	AABBA		
	ABAAB (2.81)	ABBAB		
6	AABAAB			
	ABBAAB			
	AABBAB (2.30)			

Table 5.5: Optimum two-treatment designs. Model: Simple carry-over. Within-subject error structure AR(1) ($\rho = 0.8$)

Periods	Design 1	Design 2	Design 3	Design 4
Two Sequence Designs (Variance $x10^{-2}$)				
5	AABBA (7.23)			
6	ABBAAB (5.13)			
7	ABAABBA (4.52)	ABBAABA	ABBABBA	
8	ABBAABBA (3.80)			
9	ABBAABBAB (3.36)	ABBAABAAB	ABAABBAAB	ABBABBAAB
10	ABBAABBAAB (3.03)			
Four Sequence Designs (Variance $x10^{-2}$)				
5	AABBA	AABBA		
	ABAAB (3.38)	ABBAB		
6	ABBAAB	ABBAAB		
	AABAAB (2.69)	AABBAB		
7	AABBAAB	AABBAAB	AABBAAB	
	ABBAABA (2.25)	ABAABBA	ABBABBA	
8	ABBAABBA	ABBAABBA	ABBAABBA	ABBAABBA
	ABAABAAB (1.92)	ABBABAAB	ABABBAAB	ABBAABAB
8(con't)	ABBAABBA	ABBAABBA		
	ABAABBAB	ABBABBAB		
Six Sequence Designs (Variance $x10^{-2}$)				
5	ABBAA			
	AABBA			
	ABAAB (2.29)			
6	AABAAB			
	ABBAAB			
	AABBAB (1.83)			

Table 5.6: Optimum two-treatment designs. Model: Fleiss carry-over. Within-subject error structure AR(1) ($\rho = 0.2$)

Periods	Design 1	Design 2	Design 3	Design 4
Two Sequence Designs (Variance $x10^{-2}$)				
5	AAABB (15.2)	AABBB	ABBBB	
6	AAAABB (12.7)	AAABBB	AABBBB	ABBBBB
7	AAAAABB (10.9)	AAAABBB	AAABBBB	AABBBBB
7 (con't)	ABBBBBB			
8	AAAAAABB (9.61)	AAAAABBB	AAAABBBB	AAABBBBB
8 (con't)	AABBBBBB	ABBBBBBB		
9	AAAAAAABB (8.56)	AAAAAABBB	AAAAABBBB	AAAABBBBB
9 (con't)	AAABBBBBB	AABBBBBBB	ABBBBBBBB	
10	AAAAAAAABB (7.71)	AAAAAAABBB	AAAAAABBBB	AAAAABBBBB
10 (con't)	AAAABBBBBB	AAABBBBBBB	AABBBBBBBB	ABBBBBBBBB
Four Sequence Designs (Variance $x10^{-2}$)				
5	All possible combinations in pairs of 5-period 2-sequence designs, are optimal Number of optimal designs: 3 - Variance: 7.61			
6	All possible combinations in pairs of 6-period 2-sequence designs, are optimal Number of optimal designs: 6 - Variance: 6.37			
7	All possible combinations in pairs of 7-period 2-sequence designs, are optimal Number of optimal designs:10 - Variance: 5.48			
8	All possible combinations in pairs of 8-period 2-sequence designs, are optimal Number of optimal designs:15 - Variance: 4.80			
Six Sequence Designs (Variance $x10^{-2}$)				
5	All possible combinations in triplets of 5-period 2-sequence designs, are optimal Number of optimal designs: 1 - Variance: 5.08			
6	All possible combinations in triplets of 6-period 2-sequence designs, are optimal Number of optimal designs: 4 - Variance: 4.25			

Table 5.7: Optimum two-treatment designs. Model: Fleiss carry-over. Within-subject error structure AR(1) ($\rho = 0.5$)

Periods	Design 1	Design 2	Design 3	Design 4
Two Sequence Designs (Variance $x10^{-2}$)				
5	ABBAA (15.3)			
6	ABBBAA (14.2)	ABBAAA	AABBAA	
7	ABBAABB (11.1)			
8	AABBAABB (10.5)	ABBAAABB	ABBBAABB	ABBAABBB
9	ABBAABBAA (8.69)			
10	AABBAABBAA (8.33)	ABBAABBBAA	ABBBAABBAA	ABBAAABBAA
Four Sequence Designs (Variance $x10^{-2}$)				
5	ABBAA	ABBAA	ABBAA	
	AAABB (8.69)	AABBB	ABBBB	
6	ABBAAA	ABBAAA	AABBAA	
	AABBAA (7.14)	ABBBAA	ABBBAA	
7	ABBAABB	ABBAABB	ABBAABB	ABBAABB
	ABBAAAA (6.06)	AABBAAA	ABBBAAA	AAABBAA
7 (con't)	ABBAABB	ABBAABB		
	AABBBAA	ABBBBAA		
8	All combinations in pairs of 8-period, 2-sequence designs are optimal			
	Number of designs: 6 - Variance : 5.62			
Six Sequence Designs (Variance $x10^{-2}$)				
5	ABBAA	ABBAA	ABBAA	
	AAABB	AAABB	AABBB	
	ABBBB (6.06)	AABBB	ABBBB	
6	ABBAAA			
	AABBAA			
	ABBBAA (4.76)			

Table 5.8: Optimum two-treatment designs. Model: Fleiss carry-over. Within-subject error structure AR(1) ($\rho = 0.8$)

Periods	Design 1	Design 2	Design 3	Design 4
Two Sequence Designs (Variance $x10^{-2}$)				
5	ABBAA (13.7)			
6	ABBAAB (12.1)			
7	ABBAABB (9.46)			
8	ABBAABBA (8.60)			
9	ABBAABBAA (7.22)			
10	ABBAABBAAB (6.68)			
Four Sequence Designs (Variance $x10^{-2}$)				
5	ABBAA	ABBAA		
	AABBA (8.04)	ABBBA		
6	ABBAAB	ABBAAB	ABBAAB	
	ABBAAA (6.34)	AABBAA	ABBBAA	
7	ABBAABB	ABBAABB	ABBAABB	
	ABBAAAB (5.24)	AABBAAB	ABBBAAB	
8	ABBAABBA	ABBAABBA	ABBAABBA	ABBAABBA
	ABBAAABB (4.47)	AABBAABB	ABBBAABB	ABBAABBB
Six Sequence Designs (Variance $x10^{-2}$)				
5	ABBAA			
	ABAAB			
	ABBAB (5.71)			
6	ABBAAA	AABBAA	ABBAAA	
	ABBBAA	ABBBAA	AABBAA	
	ABBAAB (4.31)	ABBAAB	ABBAAB	

Table 5.9: Optimum two-treatment designs. Model: Mixed. Within-subject error structure AR(1) ($\rho = 0.2$)

Periods	Mixed with $\phi = 0.2$		Mixed with $\phi = 0.5$	
	Design 1	Design 2	Design 1	Design 2
Two Sequence Designs (Variance $x10^{-2}$)				
5	ABBAB (9.44)		ABBBA (9.76)	
6	ABBABB (7.68)		ABBABB (7.83)	
7	ABBABBA (6.29)		ABBABBA (6.51)	
8	ABBABBAB (5.60)		ABBAABBA (5.83)	
9	ABBABBABA (4.97)	ABBABABBA	ABBABBABB (5.14)	
9 (con't)	ABABBABBA			
10	ABBABBABBA (4.37)		ABBABBABBA (4.54)	
Four Sequence Designs (Variance $x10^{-2}$)				
5	ABBBA		ABBBA	
	ABBAB (4.60)		ABBAB (4.74)	
6	ABBABB	ABBABB	ABBABB	
	ABBABA (3.78)	ABABBA	ABBBAB (3.96)	
7	ABBABBA	ABBABBA	ABBABBA	
	ABBBABA (3.20)	ABABBBA	ABBAABB (3.34)	
8	ABBABBAB	ABBABBAB	ABBABBAB	ABBABBAB
	ABBBABBA (2.80)	ABBABBBA	ABBBABBA (2.90)	ABBABBBA
Six Sequence Designs (Variance $x10^{-2}$)				
5	ABBBA		ABBBA	
	ABBAB		ABBAB	
	ABABB (3.11)		AABBA (3.17)	
6	ABBABA		ABBABA	ABABBA
	ABABBA		ABBABB	ABBABB
	ABBBBA (2.53)		ABBBBA (2.63)	ABBBBA

Table 5.10: Optimum two-treatment designs. Model: Mixed. Within-subject error structure AR(1) ($\rho = 0.5$)

Periods	Mixed with $\phi = 0.2$		Mixed with $\phi = 0.5$	
	Design 1	Design 2	Design 1	Design 2
Two Sequence Designs (Variance $x10^{-2}$)				
5	ABBAB (7.23)		ABBAB (7.91)	
6	ABBABA (6.08)	ABABBA	ABBABB (6.66)	
7	ABBABBA (4.91)		ABBABBA (5.12)	
8	ABBABBAB (4.24)		ABBABBAB (4.60)	
9	ABBABBABA (3.74)	ABBABABBA	ABBABBABB (4.19)	
9 (con't)	ABABBABBA			
10	ABBABBABBA (3.36)		ABBABBABBA (3.52)	
Four Sequence Designs (Variance $x10^{-2}$)				
5	ABBBA		AABBA	
	ABBAB (3.84)		ABBAB (3.98)	
6	ABBABB	ABBABB	ABBABB	ABBABB
	ABBABA (3.02)	ABABBA	ABBABA (3.28)	ABABBA
7	ABBABBA	ABBABBA	ABBABBA	ABBABBA
	ABBABAB (2.46)	ABABBAB	ABBAABA (2.68)	ABAABBA
8	ABBABBAB	ABBABBAB	ABBAABBA	
	ABBABABB (2.21)	ABABBABB	ABBABBAB (2.30)	
Six Sequence Designs (Variance $x10^{-2}$)				
5	ABBBA		ABBAA	
	ABBAB		ABBAB	
	ABABB (2.63)		AABBA (2.71)	
6	ABBABA		ABBABB	ABBABB
	ABABBA		ABBAAB	ABBAAB
	ABBABB (2.00)		ABBABA (2.19)	ABABBA

Table 5.11: Optimum two-treatment designs. Model: Mixed ($\phi = 0.8$). Within-subject error structure AR(1)

Periods	$\rho = 0.2$		$\rho = 0.5$	
	Design 1	Design 2	Design 1	Design 2
Two Sequence Designs (Variance $x10^{-2}$)				
5	AABBA (9.77)		AABBA (8.71)	
6	ABBAAB (8.02)		ABBAAB (6.61)	
7	ABBABBA (6.89)		ABBABBA (5.51)	
8	ABBAABBA (5.88)		ABBAABBA (4.85)	
9	ABBABBAAB (5.34)		ABBABBAAB (4.31)	ABBAABBAB
10	ABBAABBAAB (4.74)		ABBABBABBA (3.82)	
Four Sequence Designs (Variance $x10^{-2}$)				
5	ABBAA		AABBA	
	AABBA (4.88)		ABBAB (4.13)	
6	ABBAAB		ABBAAB	ABBABB
	ABBABB (4.01)		ABBABB (3.39)	ABABBA
7	ABBABBA		ABBABBA	ABBABBA
	ABBAABB (3.39)		ABBAABA (2.83)	ABAABBA
8	ABBAABBA		ABBAABBA	
	AABBABBA (2.96)		ABBABBAB (2.41)	
Six Sequence Designs (Variance $x10^{-2}$)				
5	ABBAA		ABBAA	
	AABBA		ABBAB	
	ABBAB (3.22)		AABBA (2.77)	
6	AABBAA		ABBABB	
	ABBAAB		ABBAAB	
	ABBABB (2.68)		AABBAB (2.29)	

Table 5.12: Optimum two-treatment designs. Model: Mixed. Within-subject error structure AR(1) ($\rho = 0.8$)

Periods	Mixed with $\phi = 0.2$		Mixed with $\phi = 0.5$	
	Design 1	Design 2	Design 1	Design 2
Two Sequence Designs (Variance $x10^{-2}$)				
5	ABBAB (5.48)		ABBAB (6.05)	
6	ABBABA (4.64)	ABABBA	ABBAAB (5.22)	
7	ABBABBA (3.76)		ABBABBA (3.92)	
8	ABBABBAB (3.17)		ABBABBAB (3.46)	
9	ABBABBABA (2.79)	ABBABABBA	ABBABBABA (3.16)	ABABBABBA
9 (con't)	ABABBABBA		ABBABABBA	
10	ABBABBABBA (2.53)		ABBABBABBA (2.66)	
Four Sequence Designs (Variance $x10^{-2}$)				
5	AABBA		AABBA	
	ABBAB (3.04)		ABBAB (3.15)	
6	ABBABA		ABBAAB	ABBAAB
	ABABBA (2.32)		ABBABA (2.59)	ABABBA
7	ABBABBA	ABBABBA	ABBABBA	ABBABBA
	ABBABAB (1.85)	ABABBAB	ABBAABA (2.06)	ABAABBA
8	ABBABBAB	ABBABBAB	ABBAABBA	
	ABBABABA (1.65)	ABABBABA	ABBABBAB (1.76)	
8 (con't)	ABBABBAB			
	ABABABBA			
Six Sequence Designs (Variance $x10^{-2}$)				
5	ABABA		ABBAA	
	AABBA		AABBA	
	ABBAB (2.12)		ABBAB (2.19)	
6	ABBABA		ABBABB	ABBABB
	ABABBA		ABBAAB	ABBAAB
	ABBABB (1.56)		ABBABA (1.73)	ABABBA

Table 5.13: Optimum two-treatment designs.Model:Mixed ($\phi = 0.8$). Within-subject error structure AR(1) ($\rho = 0.8$)

Periods	Two-sequence Designs		Four-sequence Designs	
	Design 1	Design 2	Design 1	Design 2
Two Sequence Designs (Variance $x10^{-2}$)				
5	ABBAB (6.81)		ABBAB	
			AABBA (3.29)	
6	ABBAAB (5.17)		ABBAAB	
			AABBAB (2.70)	
7	ABBABBA (4.23)		ABBABBA	ABBABBA
			ABBAABA (2.18)	ABAABBA
8	ABBAABBA (3.78)		ABBAABBA	
			ABBABBAB (1.85)	
9	ABBABBAAB (3.30)	ABBAABBAB		
10	ABBABBABBA (2.90)			
	Six Sequence Designs (Variance $x10^{-2}$)			
5	AABBA			
	ABBAB			
	ABBAA (2.26)			
6	ABBAAB	ABBAAB		
	AABBAB	AABBAB		
	ABBABA (1.84)	ABABBA		

Table 5.14: Optimum three-treatment designs. Full Design Listing. Within-subject error structure AR(1) ($\rho = 0.7$)

Carryover Scheme	Design 1	Design 2	Design 3	Design 4	Design 5	Design 6
Three-period, three-sequence designs (Variances $x10^{-2}$)						
No	ABC	ACB				
	BCA	BAC				
	CAB (1.43)	CBA				
Simple	ABC	ACB	ABC	ACB	ABC	ACB
	BAC	BCA	BCA	BCA	BAC	BAC
	CBA (10.0)	CAB	CBA	CBA	CAB	CAB
Simple2 and Propo-rtional	ABC	ACB				
	BCA	BAC				
	CAB (15.4)	CBA				
Three-period, four-sequence designs (Variances $x10^{-2}$)						
No	ACB	ACB	ABC	ABC	ABC	ABC
	BAC	BAC	BCA	ACB	BAC	ACB
	BCA	CAB	CAB	BCA	BCA	BAC
	CBA (0.92)	CBA	CBA	CAB	CAB	CBA
Simple	ACB	ABC	ABC			
	BCA	BAC	ACB			
	CAB	BCA	BAC			
	CBA (2.70)	CBA	CAB			
Simple2 and Propo-rtional	ABC					
	ACB					
	BAC					
	CAB (10.5)					
Three-period, five-sequence designs (Variances $x10^{-2}$)						
No carry-over: All 6 possible designs are optimal with variance 0.56						
Simple carry-over: All 6 possible designs are optimal with variance 1.52						
Simple2: (ABC,ACB,BAC,BCA,CAB) and (ABC,ACB,BAC,CAB,CBA) - variance 9.22						
Proportional: (ABC,ACB,BAC,BCA,CBA) and (ABC,BAC,BCA,CAB,CBA) - variance 9.22						

Table 5.15: Optimum three-treatment designs. Cyclic Designs considered only.

Within-subject error structure AR(1) ($\rho = 0.7$)

No Carryover	Simple Carryover	Fleiss Carryover	Simple2 Carryover	Proportional Carryover
Cyclic 4-period, 3-sequence designs (Variances $x10^{-2}$)				
ABAC	ABBC	ABAC	ABCA	ABCA
ABCB	ACCB	ACAB	ABBC	ACBA
ACBC	(1.82)	(4.45)	ACBA	(10.6)
ACAC (0.68)			ACCB (10.6)	
Cyclic 5-period, 3-sequence designs (Variances $x10^{-2}$)				
ABCBC	ABBAC	ABCBA	ABCAB	ABCAB
ABABC	ACCAB	ACBCA	ACBAC	ACBAC
ACBCB	(0.99)	(1.35)	(7.80)	(7.80)
ACACB (0.39)				
Cyclic 6-period, 3-sequence designs (Variances $x10^{-2}$)				
ABACBC	ABCCBA	ABCBAC	ABCABC	ABCABC
ABCABC	ACBBCA	ACBCAB	ACBACB	ACBACB
ABCACB	(0.55)	(0.75)	(6.05)	(6.05)
ABCBAC				
ACABCB				
ACBCAB				
ACBACB				
ACBABC (0.25)				
Cyclic 7-period, 3-sequence designs (Variances $x10^{-2}$)				
ABABCBC	ABCCBAC	ABCACBA	ABCABCA	ABCABCA
ABCBABC	ACBBCAB	ACBABCA	ACBACBA	ACBACBA
ACACBCB	(0.38)	(0.46)	(5.14)	(5.14)
ACBCACB (0.17)				
Cyclic 8-period, 3-sequence designs (Variances $x10^{-2}$)				
ABCACBCB	ABCAACBA	ABCACBAC	ABCABCAB	ABCABCAB
ACBACBCB	ACBAABCA	ACBABCAB	ACBACBAC	ACBACBAC
ACABCBCB (0.13)	(0.28)	(0.33)	(0.43)	(0.43)
plus 25				
other designs				

Table 5.16: Optimum four-treatment designs. Cyclic Designs considered only. Within-subject error structure AR(1) ($\rho = 0.7$)

No Carry-over	Simple Carry-over	Fleiss Carry-over	Simple2 Carry-over	Proportional Carry-over
Cyclic 4-period, 4-sequence designs (Variances $x10^{-2}$)				
ABDC ADBC (0.09)	ABDC ADBC (0.29)		ABDC ADBC (1.66)	ABDC ADBC (4.13)
Cyclic 5-period, 4-sequence designs (Variances $x10^{-2}$)				
ABCBD ACBCD ACDCB ADCDB (0.04)	ABDCB ADBCD (0.12)	ABDCB ADBCD (0.12)	ABDCB ADBCD (0.93)	ABCDB ADCBD (3.16)
Cyclic 6-period, 4-sequence designs (Variances $x10^{-2}$)				
ACBCBD ACDCDB (0.02)	ABCADB ADCABD (0.06)	ABCADB ADCABD (0.06)	ABCADB ADCABD (0.60)	ACBADC ACDABC (2.52)
Cyclic 7-period, 4-sequence designs (Variances $x10^{-2}$)				
ABADBDC ADBABDC (0.01) plus 14 other designs	ABDCBDA ADBCDBA (0.03)	ABDCBDA ADBCDBA (0.03)	ACBDABC ACDBADC (0.42)	ABCDABC ADCBADC (2.08)

Table 5.17: Optimum five, six-treatment designs. Cyclic Designs considered only.

Within-subject error structure AR(1) ($\rho = 0.7$)

No Carry-over	Simple Carry-over	Fleiss Carry-over	Simple2 Carry-over	Proportional Carry-over
Five-treatment designs				
Cyclic 5-period, 5-sequence designs (Variances $x10^{-2}$)				
ABDCE	ABEDC		ABEDC	ACDEB
ACBDE	ACDBE		ACDBE	ADCBE
ADECB	ADCEB		ADCEB	(1.57)
(0.005)	AEBCD		AEBCD	
plus 9	(0.02)		(0.19)	
other designs				
Cyclic 5-period, 4-sequence designs (Variances $x10^{-2}$)				
ABDBCE	ABEDCE	ABEDCE	ABEDCE	ABDEAC
ACBCED	ACDBED	ACDBED	ACDBED	AECBAD
ADCECB	ADCEBC	ADCEBC	ADCEBC	(1.29)
(0.002)	AEBCDB	AEBCDB	AEBCDB	
plus 13	(0.008)	(0.008)	(0.072)	
other designs				
Six-treatment designs				
Cyclic 6-period, 6-sequence designs (Variances $x10^{-4}$)				
ABDECF	ACBEFD		ACBEFD	ACEFBD
ACFEDB	AEFCBD		AEFCBD	AECBFD
ADFECB	(0.11)		(0.88)	(72.88)
(0.02)				
plus 15				
other designs				

Chapter 6

Thesis Close-out

6.1 The 2x2 case revisited

A short account of the main thesis results will be provided in this chapter. A thorough examination of the analysis strategies of the 2x2 design with continuous data has been presented. Depending on the inclusion or not of the carry-over term, two test statistics can be proposed for testing treatment effectiveness: the never pooled test using the first period data only (PAR) and the more powerful pooled test based on data from both periods (CROS). Under the simple carry-over model, the condition needed to be satisfied so that the more powerful CROS is selected instead of PAR depends upon the unknown carry-over effect. Similarly, the best weighted combination of PAR and CROS, places weight on CROS which depends not only on the unknown carry-over effect but also on the variance of the test statistic for checking the significance of that term.

The properties of the two stage procedure (TS), where CROS is selected with probability p and PAR with probability $1-p$, have been reviewed. It is well-known, that TS has worst performance in terms of power for treatment effect estimation in comparison to CROS. This comparison though is not statistically appropriate, since the Type I error rate of TS is 8.7%, while that of CROS is 5%. Two strategies for fixing the Type I error rate of TS are presented. The new improved TS scheme still performs worse in terms of power when compared to CROS. Both the original and improved TS strategy perform worst in terms of MSE, when compared to CROS. A 2x2 trial in asthma is then analyzed from

a Frequentist and a Bayesian point of view. In both approaches the residual treatment effect seems to be unimportant. The Bayesian approach has the advantage of concluding that the newly proposed therapy is more effective than the standard treatment, regardless of the inclusion or not of the carry-over term.
The inclusion of baselines in the 2x2 cross-over experiment is then considered. A three stage procedure is proposed for evaluation of the treatment effect. Two modifications of that scheme are studied. The Type I error rate is over the nominal 5% level for both strategies. Both schemes perform worse in terms of power when compared to CROS. One of the two schemes (strategy 1) handles carry-over terms in a more rationale way than the other one (strategy 2). Overall strategy 1 has always a better performance when compared to strategy 2, in terms of power and MSE for estimating treatment effect. The same trial in asthma is re-analyzed, but now baseline measurements included in the analysis. Similar conclusions to the ones drawn by the analysis where baselines ignored, are reported. The inclusion of demographic information, e.g. sex, in a 2x2 experiment affects only the terms estimated using between subject information (e.g. carry-over). The impact of these terms on treatment effect or other within subject contrasts is rather minimal. Finally, since carry-over is related to the treatment effect, the analysis of the 2x2 trial in asthma was repeated by introducing appropriate non-linear terms in order to describe the mathematical association between treatment and carry-over. That analysis stresses in an even more emphatic way that the newly proposed therapy is more effective compared to the standard therapy, even when carry-over effects are not handled in the best way. A model selection exercise based on the AIC criterion is then performed, and the model with no residual terms seems to be favored as the most appropriate for having generated the observed data.

6.2 Selecting a design

The problems with the 2x2 design can possibly be overcome, if the estimation of the carry-over effect is made by using within-subject information. Multi-period, multi-sequence designs can be used to that purpose. Initially four types of carry-

over are examined: No carry-over, Simple, Fleiss and second order carry-over. In the three-period, two-sequence family a design with excellent properties for estimating both treatment and carry-over effect is the (ABB/duals). This choice is quite robust to the type of carry-over assumed. In the three-period four-sequence family a design with good properties over all carry-over schemes is the (ABB,AAB,duals). When more periods are used, then the choice of a good plan becomes less clear. In the four-period family when two sequence plans are considered, design (ABBA/duals) is optimum for estimating treatment and carry-over difference over most of the carry-over scenarios. Firm recommendations cannot be made if the number of sequences increases to four or six.

Optimum plans under a decision rule that may sound appealing to practitioners who design cross-over trials, are derived. More specifically during the planning stage the statistician is unaware of the carry-over mechanism (if any) that will generate the observed data. In addition he is unaware if the model fitted at the analysis stage will correctly identify true carry-over activity. However, the statistician should write down in the protocol clearly the type of carry-over he is prepared to adopt in his analysis, without analyzing the data. In other words, the "analysis" model may completely miss the "true" model and interest focuses in identifying designs with minimum MSE for estimating treatment effect, under that scenario. Four types of carry-over are now considered: No carry-over, Simple, Fleiss and the Mixed one. The mixed carry-over is an intermediate scenario between the Simple and the Fleiss types. The selected plan does not seem to depend on the correlation between successive responses on a subject. The decision though, is heavily affected by the assumptions made regarding "true" and "analysis" models. But how much of a problem is the identification of the correct carry-over type during the planning stage? The clinical team usually allows for adequate wash-out interval which make sure that presence of carry-over is highly unlikely. This point is illustrated with the analysis of data from a cross-over trial with 7 treatments, where carry-over types that sound reasonable in the outset have no effect on our inferences for treatment, simply because carry-over is not present anyway.

For two treatment comparison, up to four-period cross-over designs are used to

run a trial. Practical considerations suggest that routine follow-up is not that costly as recruitment of new patients. The future of cross-over studies lies on using multi-period designs and evaluation of best plans in these design families is worth investigating. For two treatment comparison, under the no carry-over model, frequent switches between the two therapies are needed. The number of switches is moderate for the simple carry-over scenario, while it is minimal if treatment residual activity is described by the Fleiss carry-over scenario. It has to be noted that conclusions heavily depend on within-subject correlation structure. If in the two treatment scenario, carry-over is modeled as a proportion of the treatment effect this has minimal impact on the design choice. In the case where more than two treatments compared results are less clear. To begin with, full listing of the design family is not possible, since the number of distinct plans grows fast as number of sequences and/or periods increases. Cyclic families are studied. A general comment worth made is that the number of optimum designs under the model with no carry-over terms is usually much higher compared to the number of best plans under any carry-over scenario. Five different carry-over types have been considered.

In conclusion, modeling carry-over activity has future only in Phase I trials. The derivation of algorithms for efficient planning of cross-over designs with unlimited number of sequences and periods needs further development. Multi-stage procedures for other data types (e.g. binary) need to be studied. Efficient assessment of specific interaction terms of interest to sponsors could be seen as another research direction.

Bibliography

[1] A. C. Atkinson and A. N. Donev. *Optimum Experimental Designs.* Clarendon Press, Oxford, 1992.

[2] G. E. Box and G. C. Tiao. *Bayesian Inference in Statistical Analysis.* John Wiley & Sons, New York, 1992.

[3] B. Brown. The crossover experiment for clinical trials. *Biometrics*, 36:69–79, 1980.

[4] K. C. Carriere and G. C. Reinsel. Optimal two-period repeated measurement designs with two or more treatments. *Biometrika*, 80:924–929, 1993.

[5] C. S. Cheng and C. F. Wu. Balanced repeated measurements designs. *The Annals of Statistics*, 8:1272–1283, 1980.

[6] M. E. Chi. Recovery of inter-block information in cross-over trials. *Statistics in Medicine*, 10:1115–1122, 1991.

[7] M. E. Chi. Analysis of cross-over trials when within-subject errors follow an AR(1) process. *Biometrical Journal*, 34:359–365, 1993.

[8] W. S. Cleveland. Robust locally weighted regression and smoothing scatter plots. *Journal of the American Statistical Association*, 74:829–836, 1979.

[9] A. F. Ebbutt. Three-period crossover designs for two treatments. *Biometrics*, 40:219–224, 1984.

[10] B. Efron. Forcing a sequential experiment to be balanced. *Biometrika*, 58:403–417, 1971.

[11] B. S. Everitt. *The Analysis of Contingency Tables.* Chapman & Hall, London, 1992.

[12] F. Ezzet and J. Whitehead. A random effects model for ordinal responses from a crossover trial. *Statistics in Medicine*, 10:901–907, 1991.

[13] V. V. Fedorov and P. Hackl. *Model-Oriented Design of Experiments.* Springer-Verlag, New York, 1997.

[14] M. Feingold and B. W. Gillespie. Cross-over trials with censored data. *Statistics in Medicine*, 15:953–967, 1996.

[15] V. Fidler. Change-over clinical trials with binary data: mixed model based comparisons of tests. *Biometrics*, 40:1063–1070, 1984.

[16] R. A. Fisher and F. Yates. *Statistical Tables for Biological, Agricultural and Medical Research.* Oliver and Boyd, Edinburgh, 1963.

[17] J. L. Fleiss. *The design and analysis of clinical experiments.* John Wiley & Sons, New York, 1986.

[18] J. L. Fleiss. A critique of recent research on the two treatment cross-over design. *Controlled Clinical Trials*, 10:237–243, 1989.

[19] S. M. Fletcher, D. J. Lewis and J. N. S. Matthews. Factorial designs for crossover clinical trials. *Statistics in Medicine*, 9:1121–1129, 1990.

[20] A. L. France, A. J. Lewis, and R. Kay. The analysis of failure time data in crossover studies. *Statistics in Medicine*, 10:1099–1113, 1991.

[21] P. R. Freeman. The performance of the two-stage analysis of two-treatment two-period cross-over trials. *Statistics in Medicine*, 8:1421–1432, 1989.

[22] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in practice.* Chapman and Hall, New York, 1995.

[23] P. S. Gill and G. K. Shukla. Optimal changeover designs for correlated observations. *Communication in Statistics A*, 16:2243–2261, 1987.

[24] J. M. Grender and W. D. Johnson. Analysis of crossover designs with multivariate response. *Statistics in Medicine*, 12:69–89, 1993.

[25] J. M. Grender and W. D. Johnson. Fitting multivariate polynomial growth curves in two-period crossover designs. *Statistics in Medicine*, 13:931–943, 1994.

[26] A. P. Grieve. A bayesian analysis for the two-period crossover design for clinical trials. *Biometrics*, 41:979–990, 1985.

[27] A. P. Grieve. Extending a bayesian analysis of the two-period crossover to allow for baseline measurements. *Statistics in Medicine*, 13:905–924, 1994.

[28] A. P. Grieve. Extending a bayesian analysis of the two-period crossover to accomodate missing data. *Biometrika*, 82:277–286, 1995.

[29] A. P. Grieve and S. J. Senn. Estimating treatment effects in crossover trials. *Journal of Biopharmaceutical Statistics*, 8:191–247, 1998.

[30] J. E. Grizzle. The two-period change over design and its use in clinical trials. *Biometrics*, 21:467–480, 1965.

[31] O. Guilbaud. Exact inference about the within-subject variability in 2x2 crossover trials. *Journal of the American Statistical Association*, 88:939–946, 1993.

[32] K. B. Hafner, G. G. Koch, and A. T. Canada. Some analysis strategies for three-period changeover designs with two treatments. *Statistics in Medicine*, 7:471–481, 1988.

[33] A. Hedayat and K. Afsarinejad. Repeated measurements designs II. *Annals of Statistics*, 6:619–628, 1978.

[34] P. J. Huber. *Robust Statistics.* John Wiley & Sons, New York, 1981.

[35] C. Jennison and B. W. Turnbull. *Group Sequential Methods with Applications to Clinical Trials.* Chapman & Hall / CRC, London, 1999.

[36] J. A. John and E. R. Williams. *Cyclic and Computer Generated Designs.* Chapman and Hall, London, 1995.

[37] B. Jones and A. N. Donev. Modelling and design of cross-over trials. *Statistics in Medicine*, 15:1435–1446, 1996.

[38] B. Jones and M. G. Kenward. Modeling binary data from a three-period cross-over trial. *Statistics in Medicine*, 6:555–564, 1987.

[39] B. Jones and M. G. Kenward. *Design and analysis of cross-over trials.* Chapman & Hall, London, 1989.

[40] B. Jones and A. J. Lewis. The case for cross-over trials in phase III. *Statistics in Medicine*, 14:1025–1038, 1995.

[41] B. Jones and J. Wang. Comments on estimating treatment effects in clinical crossover trials. *Journal of Biopharmaceutical Statistics*, 8:235–238, 1998.

[42] M. G. Kenward and B Jones. The analysis of data from 2x2 cross-over trials with baseline measurements. *Statistics in Medicine*, 6:911–926, 1987.

[43] R. P. Kershner and W. T. Federer. Two-treatment crossover designs for estimating a variety of effects. *Journal of the American Statistical Association*, 76:612–619, 1981.

[44] J. Kiefer. Construction and optimality of generalized youden designs. In J. N. Srivastava, editor, *A survey of Statistical designs and Linear Models*, pages 333–353, Amsterdam, 1975. North Holland.

[45] G. G. J. Koch. The use of non-parametric methods in the statistical analysis of the two-period change-over design. *Biometrics*, 28:577–584, 1972.

[46] W. J. Krzanowski. *Principles of Multivariate Analysis - A User's Perspective.* Clarendon Press, Oxford, 1988.

[47] J. Kunert. Optimal design and refinement of the linear model with applications to repeated measurements designs. *The Annals of Statistics*, 11:247–257, 1983.

[48] J. Kunert. Optimality of balanced uniform repeated measurements designs. *The Annals of Statistics*, 12:1006–1017, 1984.

[49] J. Kunert. Optimal repeated measurements designs for correlated observations and analysis by weighted least squares. *Biometrika*, 72:375–389, 1985.

[50] J. Laird, N. M. Skinner and M. Kenward. An analysis of two-period crossover designs with carry-over effects. *Statistics in Medicine*, 11:1967–1979, 1992.

[51] E. M. Laska and M. Meisner. A variational approach to optimal two treatment crossover designs: Application to carryover effect models. *Journal of the American Statistical Association*, 80:704–710, 1985.

[52] E. M. Laska, M. Meisner, and H. B. Kushner. Optimal crossover designs in the presence of carryover effects. *Biometrics*, 39:1087–1089, 1983.

[53] V. Lasserre. Determination of optimal designs using linear models in crossover trials. *Statistics in Medicine*, 10:909–924, 1991.

[54] M. W. J. Layard and J. N. Arvesen. Analysis of poisson data in crossover experimental designs. *Biometrics*, 34:421–428, 1978.

[55] W. Lehmacher. Analysis of the crossover design in the presence of residual effects. *Statistics in Medicine*, 10:891–899, 1991.

[56] J. K. Lindsey. *Models for Repeated Measurements.* Clarendon Press, Oxford, 1993.

[57] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data.* John Wiley & Sons, New York, 1987.

[58] S. L. Lohr. Optimal bayesian design of experiments for the one-way random effects model. *Biometrika*, 82:175–186, 1995.

[59] G. C. Magda. Circular balanced repeated measurement designs. *Communications in Statistics - Theory and Methods*, 9:1901–1918, 1980.

[60] J. T. Mardia, K. V. Kent and J. M. Bibby. *Multivariate Analysis.* Academic Press, New York, 1979.

[61] J. N. S. Matthews. Optimal crossover designs for the comparison of two treatments in the presence of carryover effects and autocorrelated errors. *Biometrika*, 74:311–320, 1987.

[62] J. N. S. Matthews. Recent developments in crossover designs. *International Statistical Review*, 56:117–127, 1988.

[63] J. N. S. Matthews. Estimating dispersion parameters in the analysis of data from crossover trials. *Biometrika*, 76:239–244, 1989.

[64] J. N. S. Matthews. The analysis of data from crossover designs: The efficiency of least squares. *Biometrics*, 46:689–696, 1990.

[65] J. N. S. Matthews. Modelling and optimality in the design of crossover studies for medical applications. *Journal of Statistical Planning and Inference*, 42:89–108, 1994.

[66] P. McCullaph and J. A. Nelder. *Generalized Linear Models.* Chapman & Hall, London, 1989.

[67] R. McHugh and O. Gomez-Marin. Randomization and additivity in the two-period crossover clinical trial. *Biometrical Journal*, 29:961–970, 1987.

[68] H. I. Patel. Analysis of incomplete data in a two-period crossover design with reference to clinical trials. *Biometrika*, 72:411–418, 1985.

[69] J. G. Pigeon and D. Raghavarao. Crossover designs for comparing treatment with a control. *Biometrika*, 74:321–328, 1987.

[70] J. Pilz. *Bayesian estimation and experimental designs in linear regression problems.* Wiley, New York, 1989.

[71] A. B. Richardson and F. V. Flack. The analysis of incomplete data in the three-period two-treatment cross-over design for clinical trials. *Statistics in Medicine*, 15:127–143, 1996.

[72] J. L. Schafer. *Analysis of Incomplete Multivariate Data.* Chapman & Hall, London, 1997.

[73] S. R. Searle. *Variance Components.* John Wiley & Sons, New York, 1992.

[74] S. J. Senn. Cross-over trials, carry-over effects and the art of self-delusion. *Statistics in Medicine*, 7:1099–1101, 1988.

[75] S. J. Senn. The use of baselines in clinical trials of bronchodilators. *Statistics in Medicine*, 8:1339–1350, 1989.

[76] S. J. Senn. Is the simple carryover model useful ? *Statistics in Medicine*, 11:715–726, 1992.

[77] S. J. Senn. *Crossover trials in clinical research.* John Wiley & Sons, New York, 1993.

[78] S. J. Senn. The AB/BA crossover: past, present and future. *Statistical Methods in Medical Research*, 3:303–324, 1994.

[79] S. J. Senn. The AB/BA cross-over: How to perform the two stage analysis if you can't be persuaded that you shouldn't. *Liber Amicorum Roel van Strik*, pages 93–100, 1996.

[80] S. J. Senn. *Statistical Issues in Drug Development.* John Wiley & Sons, New York, 1997.

[81] S. J. Senn and H. Hildebrand. Crossover trials, degrees of freedom, the carryover problem and its dual. *Statistics in Medicine*, 10:1361–1374, 1991.

[82] S. J. Senn and D. N. Lambrou. Robust and realistic approaches to carry-over. *Statistics in Medicine*, 17:2849–2864, 1998.

[83] S. J. Senn, J. Lillienthal, F. Patalano, and D. Till. An incomplete blocks cross-over in asthma: a case study in collaboration. *Cross-Over Trials - L. Hothnor, Ed. Fischer, Stuttgart*, 1996.

[84] L. Shargel and A. B. C. Yu. *Applied Biopharmaceutics and Pharmacokinetics.* Appleton and Lange, Stamford, 1999.

[85] Y. Sheiner, L. B. Hashimoto and S. L. Beal. A simulation study for comparing designs for dose ranging. *Statistics in Medicine*, 10:303–321, 1991.

[86] R. C. Shumaker and C. M. Metzler. The phenytoin trial is a case study of individual bioequivalence. *Drug Information Journal*, 32:1063–1072, 1998.

[87] D. J. Spiegelhalter. A language and program for complex bayesian modelling. *The Statistician*, 43:169–178, 1995.

[88] T. K. Tsai and L. H. Patel. Robust procedures for analysing a two-period cross-over design with baseline measurements. *Statistics in Medicine*, 15:117–126, 1996.

[89] J. Vuorinen and J. Turunen. A simple three-step procedure for parametric and nonparametric assessment of bioequivalence. *Drug Information Journal*, 31:167–180, 1997.

[90] S. Wang and H. M. J. Hung. Use of two-stage test statistic in the two-period crossover trials. *Biometrics*, 53:1081–1091, 1997.

[91] A. R. Willan. Using the maximum test statistic in the two period cross-over trial. *Biometrics*, 44:211–218, 1988.

[92] A. R. Willan and J. R. Pater. Using the baseline measurements in the two-period crossover clinical trial. *Controlled Clinical Trials*, 7:282–289, 1986.

[93] H. Zimmermann and W. Rahlfs. Model building and testing for the change-over designs. *Biometrical Journal*, 22:197–210, 1980.