The place of the crossover design in infertility trials: a maximum likelihood approach

Joseph McDonnell¹,²,³, Angelique J. Goverde² and Jan P.W. Vermeiden²

¹Institute for Medical Technology Assessment, Erasmus University, Rotterdam and ²Division of Reproductive Endocrinology and Fertility, Institute for Endocrinology, Reproduction and Metabolism, Vrije Universiteit Medical Centre, Amsterdam, The Netherlands
³To whom correspondence should be addressed at: Division of Reproductive Endocrinology and Fertility, Institute for Endocrinology, Reproduction and Metabolism, Vrije Universiteit Medical Centre, 1081 AV Amsterdam, The Netherlands. E-mail: j.mcdonnell@vumc.nl

BACKGROUND: For some years, there has been a debate as to the place of the crossover trial in assisted reproduction technology (ART). We aimed to investigate whether crossover and parallel designs result in different estimates of treatment effects. METHODS: We carried out computer simulation of cohorts of patients undergoing either intra-uterine insemination (IUI) or IVF under both parallel and crossover designs, under scenarios involving censoring and carryover effects. Results of the simulation were analysed using a maximum likelihood approach. RESULTS: No relevant difference was found between the designs. The crossover design resulted in slightly more pregnancies than the parallel design. Carryover effects may slightly distort the estimates of treatment effects. Crossover and parallel designs will produce essentially the same statistical estimates of treatment effect and percentage of pregnancies. The crossover design is an acceptable design in infertility research provided the data are analysed correctly.

Key words: carryover effects/crossover design/infertility/maximum likelihood estimation/parallel design

Introduction

Couples entering an assisted reproduction technology (ART) trial typically undergo a course of treatments. Each course may consist of the repeated administration of one type of treatment or a sequence in which the actual treatment changes over time. Two common designs are available to researchers, the parallel design and the crossover design. In the parallel design, couples are assigned randomly to one of two treatment regimes consisting of one type of treatment which is continued until pregnancy is achieved or the couple leaves the study of their own volition. In the crossover design, the treatment offered to patients varies between cycles. A number of possible designs are available for crossover studies. Under one design, patients may initially be randomized to one treatment, subsequently alternating treatment on each cycle. Alternatively, they may initially receive one treatment for several cycles before switching to another treatment.

Since 1993, an extended and sometimes heated debate has been conducted in Fertility and Sterility on the place of the crossover design in infertility trials. Daya (1993) opened the debate by stating that, in his opinion, the crossover design has no place in infertility trials. His concerns about the crossover design include the fact that some women will become pregnant at the first attempt and will therefore not be exposed to the second treatment, leading to possibly misleading results and a loss of statistical efficiency. Khan et al. (1996) conducted a meta-analysis to examine the hypothesis that a difference in the estimates of treatment effect exists between parallel and crossover designs. After considering 34 overviews, they came to the conclusion that the crossover design may greatly overestimate the treatment effect. Other authors did not fully agree with this conclusion. Olive (1997) agreed that the crossover trial will overestimate the treatment effects but suggested that this may be due to inadequate statistical analysis. He also stated that a crossover design may be more acceptable to patients, leading to easier accrual and reduced drop out, a point echoed by Mol and Bossuyt (1997). Ananth and Rhoads (1997) criticized the approach used by Khan, citing the method of pooling used by Khan and claiming that the statistical methods used were inappropriate. te Velde et al. (1998) also criticized the methods of analysis used by Khan and re-analysed the data used by Khan. They subsequently concluded that the observed differences were statistically insignificant. Cohlen et al. (1998) conducted a series of simulations which indicated that the crossover design did slightly overestimate the treatment effects, but that this overestimation was insignificant in comparison with random variation. Finally, Norman and Daya (2000) constructed a simple but ingenious model which showed an underestimation of effectiveness in the parallel arm and an overestimation in the crossover arm.
In this study, we extend the simulation analysis of Cohlen using a maximum likelihood approach based on a parametric model, similar in spirit to one derived from a study carried out in The Netherlands to compare the efficacy of intrauterine insemination (IUI) and IVF, to examine differences between the two designs and to examine the effect of censoring and carryover effects.

The question of interest is: does the design structure and/or the presence of carryover effects lead to a bias in the statistical analysis, leading to possible under- or overestimation of treatment effects?

### Materials and methods

We used a simulation approach to follow the progress of two theoretical cohorts entering into a trial comparing IUI and IVF. The progress of the two cohorts was simulated under both the parallel and crossover designs. In the parallel arm, couples initially are assigned randomly to one treatment, staying with that treatment until leaving the trial. In the crossover arm, couples are assumed initially to be assigned randomly to one treatment, switching treatments in subsequent cycles.

We assumed that the carryover effects of a stimulated cycle (should they exist) will lead to a decreased chance of pregnancy in the following cycle. The existence of carryover effects (as postulated here) will have different consequences under the two designs. In a parallel study, IVF cycles following the first cycle will be subject to a decreased probability of success as compared with the ‘stand alone’ per cycle probability. In contrast, in the crossover study, it is the probability of conception on IUI cycles (which, by definition, follow IVF cycles) which is decreased.

This model differs from the models of Cohen and Norman in that it is a parametric model and estimates treatment effect by means of a regression model rather than simply counting the numbers of pregnancies achieved under the two designs. Moreover, it incorporates both carryover effects and censoring, both of which may effect the estimation of treatment efficacy.

Four baseline scenarios were examined, each involving a particular combination of censoring and carryover assumptions. The scenarios examined were: (i) no couples dropped out (no censoring) and the treatment had no effect on the following cycle (no carryover effect); (ii) no censoring, but there was a negative carryover effect; (iii) couples could be censored, but there was no carryover effect; and (iv) both censoring and carryover effects were present.

In 2000, we presented the results of a clinical trial carried out in The Netherlands which examined the efficacy and cost-effectiveness of IUI in a spontaneous cycle, IUI in a mildly stimulated cycle and IVF in a prospective, randomized study of 258 couples (181 with idiopathic subfertility and 77 with male subfertility) seeking treatment for infertility. After entry into the study, couples were randomized into one of the three treatment groups. Couples received a maximum of six treatment cycles. The design of the trial and patient characteristics are described elsewhere (Goverde et al., 2000). We analysed the results of this trial by directly examining the likelihood of the observed data (McDonnell et al., 2002). Our results indicated that there was no significant difference in the chance to conceive between the two IUI groups, and we therefore combined these groups. Subsequently, we developed a parametric model to examine the differences between two treatment groups (IUI and IVF). The approach used in this article is based on the methods and results of that trial. However, we extend this model to examine the differences between the parallel and crossover designs, including scenarios not encompassed by the trial.

In the clinical study, we estimated both the chance of achieving pregnancy and the chance of drop out using logistic functions of patient characteristics and treatment, and explicitly modelled the probability both of achieving pregnancy and of censoring. In the present study, we assumed that in the baseline scenarios, the (per cycle) probabilities of pregnancy and censoring were equal to those found in the clinical study, namely: (i) the probability of both pregnancy and drop out were logistic in form and are dependent on the clinical characteristics of the couple presenting for treatment; (ii) the (per cycle) probability of pregnancy was equal to that of patients undergoing that treatment in the clinical study; and (iii) the (per cycle) probability of drop out was equal to that of patients undergoing that treatment in the clinical study and that this probability was not dependent on the treatment given in earlier cycles.

We also assume that each couple is offered a maximum of six treatment cycles. Couples are censored if they leave the trial before the maximum number of cycles is reached, unless pregnancy is achieved. In the no carryover scenarios, we assumed that treatment in one cycle had no effect on treatment in the following cycle. In the scenarios involving carryover effects, we assumed that in a cycle which followed a stimulated (IVF) cycle, the log-odds of pregnancy associated with that cycle was reduced by ln(1.15), where ln(·) is the natural log function.

More exactly, we assumed that

(i) in the no-carryover scenarios, the per cycle probability of pregnancy was

\[ P[pregnancy] = \frac{\exp(\mu_1)}{1 + \exp(\mu_1)} \]

where

\[ \mu_1 = -0.321865 - 6.146251 \times \text{age} + 0.330317 \times (\text{treatment} = \text{IVF}) \]

where ‘age’ is the age of the female patient (divided by 100 for numerical reasons) and ‘treatment = IVF’ is a binary variable taking the value 1 if the treatment during that cycle is IVF and 0 if the treatment is IUI. The values of the coefficients are those derived from the clinical study (9).

(ii) In the carryover scenarios, the per cycle probability of pregnancy was

\[ P[pregnancy] = \frac{\exp(\mu_2)}{1 + \exp(\mu_2)} \]

where

\[ \mu_2 = -0.321865 - 6.146251 \times \text{age} + 0.330317 \times (\text{treatment} = \text{IVF}) - \ln(1.15) \times (\text{previous cycle was IVF}) \]

We also assume that the carryover affects only the following cycle.

(iii) In scenarios involving censoring, the probability of censoring was

\[ P[censoring] = \frac{\exp(\eta)}{1 + \exp(\eta)} \]

where

\[ \eta = -3.073850 + 1.361340 \times (\text{treatment} = \text{IVF}) \]

Again, these values are those derived from the clinical study.

We also assumed that there was no period effect (Senn, 2002). This is an important assumption as carryover effects will be confounded with the interaction of treatment and the period effect. Under our assumptions, good prognosis patients are more likely to become pregnant and leave the study. This will be reflected in a decrease in pregnancy rates in later cycles (as is seen in practice) and is not a result of any carryover effect.

### Sensitivity analysis

To examine the robustness of the results, the analyses were re-run under a number of other assumptions, namely (i) the carryover effect was much stronger than that presumed in the ‘baseline’
analyses; more precisely, the odds ratio associated with the simulated cycle in models involving carryover was 1.30 instead of 1.15; (ii) no difference in treatment effect existed, i.e. the per cycle probability of pregnancy for IVF was equal to that of IUI; (iii) the difference in treatment effect was much stronger than in the baseline scenarios; more exactly, the coefficient associated with IVF was equal to twice that of the value used in the baseline scenarios; (iv) there was no difference in the probability of censoring, i.e. the per cycle probability of censoring following IVF was equal to that of IUI; (v) the carryover effect was positive in nature; and (vi) the probability that a couple would conceive was not in fact constant but is scaled by a random variable distributed as a Beta(2,2) distribution.

**Simulation and statistical analysis**

Each cohort consisted of 100 couples, with each couple being simulated separately. The age of the female patient was randomly generated using the formula

\[ \text{age} = 25 \times 1.6^U \]

where \( U \) is a random variable uniformly distributed on \([0,1]\). This construction restricts the age range to \([25,40]\) with relatively more younger women than older women, a situation not dissimilar to that seen in practice. Subsequently, each couple ‘progresses’ through the treatment regimen. At each stage of treatment, the couple are randomly allocated to ‘pregnant’ or ‘not pregnant’, based on their (scenario-specific) probability of achieving pregnancy. Pregnant couples leave the study, while non-pregnant couples are then randomly assigned to ‘censored’ or ‘not censored’, again based on the scenario-specific probability of censoring. Censored couples leave the study while non-censored couples go on to receive another round of treatment, with couples having at most six rounds. For each couple, the number of IUI and IVF treatments, pregnancy and censoring status are recorded. We calculated the likelihood function associated with the progress of couples through this process. Details of the construction of the likelihood function are given in the Appendix. The resulting data were then analysed using the LE (likelihood estimation) module of BMDP. For both study designs and for all scenarios, the analysis was repeated 1000 times. The median values of the parameter values are reported.

It is important to stress that in the statistical analyses, we pretend to be unaware of the possibility of the existence of a carryover effect which may lead to a bias in the results. We wish to observe that bias, if it exists. If we ‘were’ aware of this possibility, we would adjust our analysis accordingly.

**Results**

The crossover design results in an increase in the pregnancy rate, albeit a slight one. For example, if we assume that there is neither censoring nor carryover, the pregnancy rate (per 100 couples) rises from 49.6 to 50.4 (see Table I). We would expect (roughly) just one couple more to achieve pregnancy as a result of the crossover design. Such an increase is to be expected as all couples are exposed to the most effective treatment. This increase remains small under other assumptions regarding degree of censoring and carryover effect. The presence of a negative carryover effect (unsurprisingly) reduces the pregnancy rates under both designs. The influence of the carryover effect is illustrated in Table II. Roughly speaking, the per cycle probability of pregnancy decreases by \(~1.3\%\) (dependent on age and treatment).

The median coefficients from the statistical analyses are presented in Table III. Where no carryover effect is present, there is little evidence of bias in the coefficients. In other words, use of the crossover design does not result in over- or underestimation of treatment effects, provided no carryover effect exists. However, the existence of such a crossover effect may apparently ‘bias’ some of the coefficients. As stated previously, the treatment (negatively) affected in the parallel design will be IVF, while in the crossover design, it will be IUI. In the parallel study, we see that the coefficient associated with treatment is indeed smaller than the ‘true’ value, while under the crossover design, both the constant associated with the regression and the coefficient associated with treatment are biased.

These coefficients (and their apparent biases) are better interpreted in terms of their effect on the per cycle probability of pregnancy. Table IV compares these estimates with the ‘true’ values used in the simulation. Designs not involving carryover effects produce essentially unbiased estimates, while designs which include such an effect apparently introduce a small negative bias in the probability of the cycle following a stimulated cycle (see Table IV). These biases are numerically small, the most extreme being for the per cycle estimate for IVF in a parallel study involving both censoring and carryover, with an absolute difference of 1.44% (13.84 versus 15.28%) at age 28. These differences are smaller for older women.

**Sensitivity analysis**

The sensitivity analyses confirmed the conclusions drawn from the baseline scenarios. The results of the scenarios, all involving censoring and one incorporating a carryover effect,
are reported in Table V for the IVF variables relating to probability of pregnancy and censoring. When the (negative) carryover effect is stronger, the regression coefficient associated with IVF is larger in the crossover design and decreased in the parallel design, in relation to both the no carryover situation and the milder (baseline) carryover effect scenario. In other words, in the parallel design, the effect of IVF is decreased further, while in the crossover design, the effect of IUI is diminished further. If the carryover is positive, the effects are in the opposite direction. These results are in line with expectation. In the other scenarios reported, estimates for both parameters were very much in line with the underlying values, indicating that the estimates are robust under different conditions as regards treatment effect and censoring.

**Discussion**

In this study, we used a likelihood-based approach to the statistical analysis of pregnancy data from both parallel and crossover designs. This approach was based on a parametric model incorporating both censoring and carryover effects.
In both the baseline and the alternative scenarios, estimates of treatment effect and the effect of treatment on censoring were essentially unbiased where no carryover effects are present. This was true over a range of treatment effects and degree of censoring. In contrast, the existence of carryover effects may bias the estimate of treatment effect if the analysis does not allow for the possibility of such effects. These effects may be ‘substantial’ in terms of the regression coefficients; in the baseline scenarios, this bias amounted to ~30% of the true value, while in scenarios involving a stronger carryover effect, this bias was much more substantial, being >50% of the true value. Moreover, these effects occur in both designs. These results indicate that carryover effects may be of more importance than study design.

Based on our calculations, the estimated pregnancy rates and the statistical estimates of treatment effects obtained from a crossover trial (and the conclusions which follow from these estimates) are largely the same as those from the parallel design. The results of this study do not support the conclusion of Daya that the crossover design should be avoided as an inappropriate design.

Cohlen et al. (1998) and Norman and Daya (2000) both used models to examine the role of trial design. Of the two models, that of Cohlen looks more like ours. Cohlen examined the progress of a cohort of couples, drawn from a heterogeneous population in which fecundity was assumed to follow a $\beta$ distribution, under both designs. However, they did not explicitly examine the effect of study design on the estimation of treatment effect within a parametric model, which was our focus. Nor did they examine the effects of differences in treatment effect, censoring and carryover effect. Thus, our results are more general than theirs. However, our conclusions are essentially the same. We disagree somewhat with them when they state that ‘crossover designs tend to overestimate the effect of the best treatment’, although they mitigate this statement by stating that the over-estimation is ‘clinically irrelevant’. In the Appendix, we indicate algebraically why the crossover design should produce more pregnancies, at least in the absence of censoring. A non-algebraic ‘argument’ is the following: suppose treatment $A$ is very effective (99% per cycle probability of pregnancy) while treatment $B$ is very ineffective (1% chance of pregnancy), and there are just two treatment cycles with no censoring. In the parallel design, almost all women receiving treatment $A$ fall pregnant while few receiving treatment $B$ do so. If the two groups are equal in size, ~50% of the women become pregnant. In the crossover design, almost all the women receiving treatment $A$ become pregnant, while few women receiving treatment $B$ fall pregnant. In the second cycle, those from the latter group will receive treatment $A$, with most becoming pregnant. Therefore, after just two cycles, almost all the women in the crossover design will be pregnant, compared with just 50% in the parallel design. Of course, these probabilities are extremely implausible but the conclusion remains unaltered if more realistic values are substituted. These conclusions apply even if it is not known which is the more effective treatment.

The model of Norman and Daya (2000) is rather more simple. They compare two groups, each consisting of two subgroups, one (constituting 80% of the group) with ‘low’ fecundity (equal to 10% chance of pregnancy per cycle under a control treatment), while the other 20% have a much higher fecundity (40% chance per cycle under the same treatment). An experimental treatment is assumed to double the per cycle chance of pregnancy (20 and 80% chance per cycle, respectively). Both groups undergo treatment under both the parallel and crossover designs. Their conclusions, based on this model, are rather different from ours. They found that a constant sequence design (parallel) consistently underestimates the treatment effect in all but the first cycle, whereas an alternating sequence design (crossover) overestimates the treatment effect in even cycles, but correctly estimates the treatment effect in odd cycles. However, we take issue with their method of calculation. They assume that the experimental treatment has a relative risk (RR; in this context, probability of pregnancy) of 2 compared with the control treatment. The question is as to how this assumption should be interpreted. Our interpretation is that a couple undergoing the experimental treatment have twice the probability (per cycle) of achieving pregnancy than an identical couple undergoing the control treatment. Translating this to the aggregate level, a given group of couples would experience twice as many pregnancies (per cycle) under the experimental treatment as they would under the control treatment. The point of this statement is that the value of the RR is relevant only when the groups being compared are essentially identical. In Norman’s model, this is the initial situation but this is not true for subsequent cycles in the parallel arm nor is it true in even cycles in the crossover arm. For example, in the second cycle in the parallel arm, 14% of patients have ‘high fecundity’ compared with just 6% in the experimental group. The calculated RR of 1.75 applies to a comparison of two groups who differ markedly in their average fecundity, and it is difficult to know how to interpret this value. In the crossover arm, the situation is reversed, with the experimental group containing 14% high fecundity patients compared with 6% in the control group. The comparison of the per cycle RR with the true RR is therefore not valid. This situation is not dissimilar to that of the well known Simpson’s paradox: at the patient level, the treatment RR is still 2, but this is masked at the group level due to differences in group composition. We therefore disagree with the method of calculation used by Norman and argue that the concept of relative risk should not be applied on a per cycle basis.

The existence and degree of both carryover and period effects in ART trials have not been investigated. The presence of both can lead to confounding between the carryover effect and a treatment by period interaction; indeed they are the same in the crossover trial of the form AB/BA (11). Whether this holds for ART trials is unclear. In pharmacological trials, treatment is not according to a timetable devised by physicians. In contrast, ART trials are shaped largely by patient decisions: patients may decide to delay a treatment cycle for any number of reasons and can choose the length of the delay. One such reason may well be
the nature of the treatment itself. For example, IVF is a much more physically and psychologically demanding treatment than IUI. Data from the clinical study indicate that the time between successive IVF cycles is significantly longer than that between cycles involving IUI with or without ovarian stimulation (4 months as opposed to 1.5 months, unpublished observations). These differences may largely allow for ‘washout’, with a subsequent reduction in carryover effects.

In this model, we make fairly strong assumptions about the strength of the carryover effects. We assume a ‘simple’ effect which persists only for the following cycle and which can (in principle) be observed. In practice, this may not be the case. The carryover effect may vary from cycle to cycle (e.g. due to the time difference between them) or may affect subsequent cycles (higher order carryover). Gauging the nature and extent of carryover (if it exists) is difficult and has consequences for the modelling of treatment effect. Ideally, a randomized trial comparing the crossover and parallel designs and investigating any carryover effect could be carried out but, for reasons both practical and ethical, such a trial will probably never be carried out.

Relatively few crossover trials have been carried out in the ART field. Few crossover trials with a binary outcome have been carried out in other areas. Taylor and Dominik (1999) report data from a study on condom failure using a crossover design. In this study, the outcome measure was condom failure. Unlike ART trials, participating couples did not leave the study.

In this article, we investigated possible bias introduced as a result of either study design or carryover effect. Study design had little effect on the estimation of treatment effect. On the other hand, carryover effects may well introduce bias in the estimation of treatment effects. However, is the ‘bias’ associated with carryover effects actually a bias? The answer is ‘yes’ and ‘no’. If the treatment given in one cycle affects the outcome of a given treatment in the following cycle, we are indeed observing some form of distortion in the estimate of the effect of that treatment in comparison with its ‘stand alone’ form. However, we must realize that, should carryover effects exist, the estimates of treatment effect gathered relate to that type of treatment given the total treatment regime. The estimation of a particular treatment within an overall design may not give a true picture of its actual efficacy.

Opponents of the use of the crossover design in ART trials often argue that couples achieving pregnancy should be considered as having been censored. This strikes us as being bizarre since the main outcome measure of an ART trial is pregnancy! Indeed, it is interesting to speculate what they consider is the main outcome measure if it is not pregnancy. This apparent contradiction suggests that the consideration of a ‘standard’ crossover trial (whatever that may be) is inappropriate. However, this does not imply that the design itself is inappropriate, just that another form of analysis is required.

In the Amsterdam study, censoring following IVF was greater than that following IUI. As several authors have pointed out, a crossover design may be more acceptable to couples, with possibly reduced censoring following IVF. We believe that a crossover design will lead to essentially the same conclusions as a parallel design, does not lead to over- or underestimate treatment effects, and may be more attractive to couples, possibly leading to fewer drop outs and more pregnancies. We recommend consideration of the crossover design in future studies.

References


Submitted on December 17, 2002; revised on April 16, 2004; accepted on July 23, 2004.

Appendix

Construction of the likelihood function

We define a likelihood function which is applicable to both the parallel and crossover studies. For each couple, we have data of the form \((n_{IUI}, \delta_{IUI}, \rho_{IUI}, n_{IVF}, \delta_{IVF}, \rho_{IVF})\) where: \(n_{IUI}\) = the number of IUI cycles undergone by the couple; \(\delta_{IUI} = 1\) if the couple achieved pregnancy as the result of an IUI cycle and 0 if they did not; and \(\rho_{IUI} = 1\) if the couple dropped out following an IUI cycle and 0 if they did not, with IVF parameters being defined similarly. Some of these parameters may, by definition, be equal to zero for a given pair. For example, couples in the IUI arm of a parallel study will necessarily have \(n_{IVF} = \delta_{IVF} = \rho_{IVF} = 0\).

We define \(\rho_{IUI} = P[\text{achieving pregnancy during a IUI cycle}]\) and \(\tau_{IUI} = P[\text{becoming censored following a IUI cycle}]\), with similar definitions for IVF probabilities. We assume these probabilities are constant over cycles.
Since the process is essentially discreet, each couple makes a contribution to the likelihood equal to the probability of their observed progress.

To see how the likelihood function is defined, first consider couples undergoing IUI treatment in a parallel design. There are three possible treatment progress scenarios.

(i) The couple achieve pregnancy as the result of IUI treatment ($n_{IUI} > 0$, $\delta_{IUI} = 1$, $\rho_{IUI} = 0$): i.e there were $n_{IUI} - 1$ unsuccessful cycles, none being censored, followed by a successful IUI cycle. The probability of that event is $(1 - p_{IUI})^{n_{IUI} - 1}(1 - \pi_{IUI})^{\delta_{IUI} - 1}p_{IUI} = [(1 - p_{IUI})(1 - \pi_{IUI})]^{n_{IUI} - \delta_{IUI} - 1}p_{IUI}$.

(ii) The couple do not achieve pregnancy, and subsequently drop out ($n_{IUI} > 0$, $\delta_{IUI} = 0$, $\rho_{IUI} = 1$): their likelihood contribution is $(1 - p_{IUI})^{n_{IUI}}(1 - \pi_{IUI})^{\delta_{IUI} - 1}\pi_{IUI}$.

(iii) The couple complete six cycles without achieving pregnancy or dropping out ($n_{IUI} = 6$, $\delta_{IUI} = 0$, $\rho_{IUI} = 0$): their likelihood contribution is $[(1 - p_{IUI})(1 - \pi_{IUI})]^{n_{IUI} - \delta_{IUI} - 1}$.

All three possibilities are included in the following expression, as can be seen by substituting the appropriate values of $(n_{IUI}, \delta_{IUI}$ and $\rho_{IUI}$: 

$$ [(1 - p_{IUI})^{n_{IUI} - \delta_{IUI}}(\pi_{IUI})^{\delta_{IUI}}][(1 - \pi_{IUI})^{n_{IUI} - \rho_{IUI}}(\pi_{IUI})^{\rho_{IUI}}].$$

Similarly, couples undergoing IVF treatment in a parallel design have a likelihood contribution of the form

$$ [(1 - p_{IVF})^{n_{IVF} - \delta_{IVF}}(\pi_{IVF})^{\delta_{IVF}}][(1 - \pi_{IVF})^{n_{IVF} - \rho_{IVF}}(\pi_{IVF})^{\rho_{IVF}}].$$

In the crossover design, the likelihood contribution is constructed similarly. For each couple, there is an IUI and an IVF component. For example, consider a couple who achieve pregnancy as a result of an IVF cycle. Such a couple have undergone $n_{IUI} (\geq 0)$ IUI attempts without success and without censoring, and $n_{IVF} (\geq 1)$ IVF attempts, $n_{IVF} - 1$ without success and without censoring, followed by a successful IVF cycle. Their contribution is, therefore

$$ [(1 - p_{IUI})(1 - \pi_{IUI})]^{n_{IUI} - \delta_{IUI}}[(1 - p_{IVF})(1 - \pi_{IVF})]^{n_{IVF} - \delta_{IVF}}p_{IVF}^{-1}.$$  

The contribution of other couples is constructed similarly. For each couple, irrespective of the trial design, the likelihood contribution can be written as

$$ [(1 - p_{IUI})^{n_{IUI} - \delta_{IUI}}(\pi_{IUI})^{\delta_{IUI}}][(1 - \pi_{IUI})^{n_{IUI} - \rho_{IUI}}(\pi_{IUI})^{\rho_{IUI}}] \times [(1 - p_{IVF})^{n_{IVF} - \delta_{IVF}}(\pi_{IVF})^{\delta_{IVF}}][(1 - \pi_{IVF})^{n_{IVF} - \rho_{IVF}}(\pi_{IVF})^{\rho_{IVF}}] \times (\pi_{IVF})^{\delta_{IVF}}.$$}

The likelihood itself is the product of the individual contributions from each couple. This (log-)likelihood is subsequently maximized to achieve the parameter estimates.

This likelihood function can also be used in situations not described above. For example, the protocol for the crossover arm might stipulate that the initial treatment is given for three cycles before a switch is made.

**Estimating the difference in proportion of couples achieving pregnancy under the parallel and crossover designs**

We can estimate the extra number of pregnancies due to the use of a crossover design, at least in homogenous (with respect to fecundity) populations in which no drop out occurs. We define $q_{IUI} = 1 - p_{IUI} = P[\text{no pregnancy in an IUI cycle}]$ and $q_{IVF} = 1 - p_{IVF} = P[\text{no pregnancy in an IVF cycle}]$.

In a parallel trial, the expected proportion of couples in the IUI arm not pregnant at the end of the trial is $(1 - p_{IUI})^{6} = q_{IUI}^{6}$, while in the IVF arm the expected proportion is $(1 - p_{IVF}) = q_{IVF}^{6}$. The proportion for the whole trial (assuming equal sample sizes in both arms) is therefore $(q_{IUI}^{6} + q_{IVF}^{6})/2$. In the crossover trial, the proportion of couples not pregnant at the end of the trial is $(1 - p_{IUI})^{3} = q_{IUI}^{3}$. The difference between the trials in the proportion of couples not pregnant is

$$ \Delta = P[\text{not pregnant in the parallel study}] - P[\text{not pregnant in the crossover study}] = (q_{IUI}^{6} + q_{IVF}^{6})/2 - q_{IUI}^{3}q_{IVF}^{3},$$

Denoting $q_{IUI}^{3}$ by $A$ and $q_{IVF}^{3}$ by $B$, $\Delta = (A^2 + B^2)/2 - AB = (A - B)^2 = (q_{IUI} - q_{IVF})^2$. $\Delta \geq 0$ since it is proportional to a perfect square and equals 0 if and only if $q_{IUI} = q_{IVF}$, i.e. the two treatments are equally effective. Therefore, the overall proportion of couples not pregnant in the parallel trial is (in general) greater than the proportion in the crossover trial, irrespective of the values of $q_{IUI}$ and $q_{IVF}$ (and hence $p_{IUI}$ and $p_{IVF}$).

Equivalently, the proportion of couples pregnant in the crossover trial exceeds the proportion of couples pregnant in the parallel trial. Whether this holds in practice depends on the difference in censoring between the different types of trial. However, we can estimate the difference in the proportions falling pregnant under the two designs for various values of $p_{IUI}$ and $p_{IVF}$ (see Table VI). If $p_{IUI} = 0.10$ and $p_{IVF} = 0.15$, the crossover trial has just 0.66% pregnant couples more than the parallel trial, i.e. <1 couple, if both arms contain 100 couples.

**Table VI.** Difference in proportions pregnant in the two trials as a function of $p_{IUI}$ and $p_{IVF}$ (crossover relative to parallel, maximal six cycles)

<table>
<thead>
<tr>
<th>$p_{IUI}$</th>
<th>0.05</th>
<th>0.075</th>
<th>0.10</th>
<th>0.125</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_{IVF}$</td>
<td>0.15</td>
<td>2.96</td>
<td>1.57</td>
<td>0.66</td>
</tr>
<tr>
<td>0.17</td>
<td>0.16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.175</td>
<td>4.38</td>
<td>2.64</td>
<td>1.4</td>
<td>0.59</td>
</tr>
</tbody>
</table>
If the groups are heterogeneous with respect to fecundity, the situation is rather more complex. Assume fecundity (under IUI) can be described by a variable $s$. For an IUI patient, the probability of failing to become pregnant after six rounds is

$$P(\text{failure in one IUI cycle} \mid \text{fecundity} = s) = q_{\text{IUI}}^6(s)$$

Similarly, for an IVF patient with fecundity $t$ (under IVF), the probability of failure after six cycles of IVF is

$$P(\text{failure in one IVF cycle} \mid \text{fecundity} = t) = q_{\text{IVF}}^6(t).$$

In a parallel trial, the expected proportion of couples failing to achieve pregnancy in the IUI group is

$$\int q_{\text{IUI}}^6(s)f_{\text{IUI}}(s)ds$$

where $f_{\text{IUI}}(s)$ is the distribution of fecundity under IUI and the integration is carried out over the range of $s$. The corresponding proportion in the IVF arm is

$$\int q_{\text{IVF}}^6(t)f_{\text{IVF}}(t)dt$$

The total proportion is therefore

$$\frac{1}{2}\left( \int q_{\text{IUI}}^6(s)f_{\text{IUI}}(s)ds + \int q_{\text{IVF}}^6(t)f_{\text{IVF}}(t)dt \right)$$

In the crossover trail, the proportion failing after six trials (three IUI and three IVF and assuming no carryover effects) is

$$\left( \int q_{\text{IUI}}^3(s)f_{\text{IUI}}(s)ds \right) \left( \int q_{\text{IVF}}^3(t)f_{\text{IVF}}(t)dt \right)$$

Denote $q_{\text{IUI}}^3(s)$ by $A(s)$ and $q_{\text{IVF}}^3(t)$ by $B(t)$, then

$$\Delta = P[\text{not pregnant in the parallel study}]$$

$$- P[\text{not pregnant in the crossover study}]$$

$$= \left( \int q_{\text{IUI}}^3(s)f_{\text{IUI}}(s)ds + \int q_{\text{IVF}}^3(t)f_{\text{IVF}}(t)dt \right) / 2$$

$$- \left( \int q_{\text{IUI}}^3(s)f_{\text{IUI}}(s)ds \right) \left( \int q_{\text{IVF}}^3(t)f_{\text{IVF}}(t)dt \right)$$

$$= \left( \int A^2(s)f_{\text{IUI}}(s)ds + \int B^2(t)f_{\text{IVF}}(t)dt \right) / 2$$

$$- \left( \int A(s)f_{\text{IUI}}(s)ds \right) \left( \int B(t)f_{\text{IVF}}(t)dt \right)$$

$$= \frac{1}{2}\int [A(s) - B(t)]^2f_{\text{IUI}}(s)f_{\text{IVF}}(t)dsdt$$

(using the fact that $\int f_{\text{IUI}}(s)ds = 1$ and $\int f_{\text{IVF}}(t)dt = 1$)

Again, $\Delta \geq 0$ since $[A(s) - B(t)]^2$ is non-negative and the crossover design will (assuming no drop out) result in more pregnancies than the parallel design.