Associate editor’s commentary on the article ‘Common statistical errors in the design and analysis of subfertility trials’ by A.Vail and E.Gardner.

Pitfalls in the design and analysis of efficacy trials in subfertility

Salim Daya

Departments of Obstetrics and Gynecology, and Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada

1To whom correspondence should be addressed at: McMaster University, 1200 Main Street West, Hamilton, Ontario, Canada, L8N 3Z5. E-mail: dayas@mcmaster.ca

Introduction

The introduction, more than half a century ago, of the randomized trial in which allocation to the experimental and control interventions occurs by chance has been a pivotal point in the evaluation of therapeutic efficacy. It is now well accepted that the gold standard in such evaluation is the well-designed controlled, clinical experiment that has high methodological rigour so that bias can be minimized and the magnitude of treatment effect can be estimated reliably and confidently. The acceptance of the randomized controlled trial (RCT) in the field of reproductive medicine is evident by the increasing numbers of such trials being published.

In the mid-1990’s the importance of improving the quality of reports of RCTs led to the publication of the Consolidated Standards of Reporting Trials (CONSORT) statement (1996) which was developed by an international group of experts, comprising clinical trialists, statisticians, epidemiologists and biomedical editors. The CONSORT statement, which consists of a checklist and flow diagram for reporting RCTs, has gained widespread support by many journals and organizations representing science editors. This format prompts investigators to ensure that important elements in clinical trials design and reporting have been addressed so that the information conveyed has more value. Unfortunately, this approach seems to be lacking in journals in reproductive medicine and science as evidenced by the findings of the research conducted by Vail and Gardener (2003). In their review of published controlled trials in subfertility, (published in the journals Human Reproduction and Fertility and Sterility in the year 2001), they found methodological errors including fatal flaws in design, a misunderstanding of the intention-to-treat principle and statistical errors arising from the unit of analysis being inappropriate.

What steps can be taken to reduce such errors so that the results of the trials can be interpreted reliably?

First, the CONSORT statement should be reviewed, and modified where necessary, so that it addresses the design and analytical areas that are frequently handled incorrectly in subfertility research. For example, the use of assisted reproductive techniques has focused our attention inappropriately on evaluating outcomes on a per-cycle of treatment basis rather than a per-patient basis. These two outcome measures are synonymous only in situations in which each subject contributes data from only one cycle of treatment. It is not unusual to see reports of trials in which some subjects contribute data from several cycles of treatment and the analysis of outcomes is undertaken on a per-cycle basis. Completing the updated CONSORT statement should be a requirement for all authors submitting manuscripts describing the findings from their trials.

Preliminary data indicate that the CONSORT approach is associated with an improvement in the quality of reports of RCTs (Egger et al., 2001; Moher et al., 2001a). This observation indicates that improvements in the methods of the trials have resulted, thereby enhancing their validity. By continuously monitoring the process and modifying, adding to or deleting the CONSORT items accordingly, its usefulness is increased making the CONSORT statement a continually evolving instrument through this iterative process (Moher et al., 2001b). It is hoped that such an approach would positively influence the manner in which RCTs are conducted in subfertility research.

A second step involves organizing workshops on evidence-based medicine in which clinical trial methods and critical appraisal of the literature are taught. The European Society of Human Reproduction and Embryology (ESHRE), having embarked on such a project several years ago, should be encouraged to increase the frequency of these workshops so that more investigators can equip themselves with the necessary skills to conduct clinical trials. As more investigators become familiar with the methodological elements of clinical trials, the peer review process can be strengthened, thereby ensuring the literature is not populated by results from less valid studies. Additionally, the quality of presentations at scientific meetings would be considerably improved.
A third step involves highlighting the elements in methodology that are of particular relevance to subfertility research as Vail and Gardener (2003) have tried to do in their paper. In addition to reiterating these points, several others will be highlighted below.

(i) Intention-to-treat analysis
The intention-to-treat approach to analysis (also referred to the ‘as randomized’ analysis) involves comparison of outcomes in subjects in the groups to which they were originally assigned. This rule is generally interpreted as including all subjects, regardless of whether they actually received the treatment or withdrew from the treatment programme or crossed over to the alternative intervention group. In subfertility research, and particularly in pharmaceutical trials, post-randomization exclusion of subjects is common practice for a variety of reasons, including protocol violation. Unfortunately, the bias introduced by such exclusion will affect the magnitude and direction of the effect-size and may lead to erroneous conclusions. Additionally, many investigators have misinterpreted intention-to-treat as ‘actually starting treatment’ and only include in the denominator those commencing the treatment protocol. For example, in some trials comparing the GnRH antagonist with the agonist for use in assisted reproduction, only subjects commencing FSH for ovarian stimulation were included in the analysis; patients who had been randomized to one group or the other and did not commence FSH stimulation, despite having received the agonist treatment, were incorrectly dropped from the analysis. The resulting estimate of the effect of the experimental treatment is biased and may convert a null result into a positive or negative one, especially when the results of the trials are pooled using meta-analysis. In most clinical trials, because it is expected that there will be some degree of non-compliance, the intention-to-treat analysis will tend to underestimate the treatment effect produced when subjects are fully compliant with the intervention being compared. Maintaining group similarity and preserving the balance among prognostic factors in the study groups produces the most cautious approach to evaluation, and minimizes the likelihood of a type I error in hypothesis testing. Another strategy to avoid errors introduced by subjects failing to start treatment in the group to which they were allocated is to conduct the randomization exercise as late as possible in the study design; the dictum of ‘select subjects early but randomize late’ is particularly relevant in subfertility research.

(ii) Concealment of allocation
By allocating subjects randomly to the experimental and control groups, the effect of the experimental intervention can be estimated more reliably because it is less likely to be influenced by prognostic factors, which tend to be similarly distributed in the two groups. Thus, randomization produces study groups that are comparable with respect to both known and unknown risk factors. It also removes investigator bias in the allocation of subjects.

An additional, and very important, component of the randomization process is that of concealment of allocation whereby the investigators are unaware of the group to which the subjects are allocated because the sequence of allocation is hidden from them. An example of achieving concealment very effectively is the use of a central facility for randomization, such as the pharmacy or a telephone operator who has access to the randomization table. Concealment of allocation is to be differentiated from blinding, another very important methodological criterion that ensures the subject, investigators and outcome assessors are not aware of the identity of the intervention that the subject is receiving. Thus, concealing treatment allocation, which is possible to undertake in every randomized trial, is intended to eliminate selection bias by not influencing which patient gets into the trial and the intervention she may be assigned to. In contrast, blinding is not possible to perform in all trials because it is undertaken after randomization and is intended to reduce ascertainment bias by removing any influence on the assessment of outcome (Altman and Schultz, 2001).

There is sufficient empirical evidence confirming that the effect of an experimental intervention can be overestimated (by as much as 40%) if the randomization sequence is not concealed from the investigators when consent is being obtained from the trial participants (Chalmers et al., 1983; Schulz et al., 1995). In the study by Vail and Gardener (2003), concealment of allocation using a third party or a sealed envelope method was undertaken in only one third of the subfertility trials. Given the size of the overestimation in treatment effect that is associated with lack of concealment, and the simplicity of instituting a method of ensuring that the allocation sequence is concealed, it is difficult to understand why more attention is not given to this methodological issue. It is hoped that in the 50% or so of trials reviewed, the lack of mention of concealment of allocation was merely an oversight in reporting and not an oversight in study design and execution.

(iii) Cross-over trial
The widespread availability of assisted reproductive techniques has produced a shift in focus for reporting event rates (such as clinical pregnancy) from a per-patient (or per-couple) to a per-treatment cycle basis. The variability in the numbers of treatment cycles per couple and the length of time couples may have to wait between successive cycles of treatment make evaluation of treatment efficacy more complicated. The more common clinical trial has a parallel design in which a group of subjects receiving a new treatment (the experimental intervention) is compared with another group of equally eligible subjects receiving a standard treatment or placebo (the control intervention). Thus, subjects receive their respective treatments simultaneously (i.e. in parallel fashion) and each subject receives only one of the possible treatments so that a between-patient comparison in outcome can be conducted. If the unit of analysis is the cycle, then the experimental treatment is administered randomly in one cycle in one group of subjects and the control treatment is administered randomly in one cycle in a second group of subjects. Since each subject only contributes data from only one cycle of treatment, the outcome can be expressed on a per-patient or per-cycle basis.
Over the past decade, the cross-over study design has gained popularity and is being recommended for subfertility research by many investigators. In the classical case of a two-period cross-over trial design, the responses are observed in the same subject in two different cycles, one with the experimental treatment and the other with the control treatment, the assignment of the order (experimental treatment first or control treatment first, then the alternative treatment) being determined randomly. The treatment effect is then ascertained by comparing the summary effect measures in the two groups using a within-subject comparison method. Apart from the advantages of statistical efficiency and reducing bias, it is more appealing to patients because they all have the opportunity to receive the experimental treatment, if not in the first cycle (or period) then in the second cycle (or period). The initial impression is that cross-over trial is straightforward and sensible, but in the area of subfertility, when pregnancy is the outcome of interest, it is an inappropriate choice and should be avoided (Daya, 1993, 1999, 2001a).

In subfertility cross-over trials, the subject who conceives with one treatment in the first period will be classified as a dropout and will not have the opportunity to receive the alternative treatment in the second period. Hence, a within-subject comparison is not possible. Consequently, the treatment evaluation becomes uncontrolled, unless the analysis is restricted to the first period before the cross-over point, in which case the trial has a parallel design but with inadequate power because the sample size is insufficient. By pooling data over two study periods (and ignoring the within-patient comparison) as is usually (and incorrectly) done in subfertility trials, one obtains an estimate of treatment effect that is much larger than with a one-cycle parallel design trial (Khan et al., 1996; Norman and Daya, 2000). This issue of bias has been explored further by extending the numbers of cycles of treatment per subject to more than two to simulate the common practice of offering patients multiple cycles of the same treatment or alternating between two different treatments (Norman and Daya, 2000). Both approaches produce biased estimates and are not appropriate for subfertility research. The inability to estimate both the period effect and carry-over effect further complicates the analysis rendering the cross-over design inappropriate for use in subfertility.

If compliance is an issue, then the cross-over design can be used, but the analysis should be restricted to the data from only the first period and the sample size should be sufficiently large. In this manner, the trial is really a parallel design, but subjects failing to conceive after receiving the control treatment in the first period are assured that they will receive the experimental treatment in the next cycle. The data from these additional cycles after the cross-over would be discarded from the analysis.

(iv) First cycle enrolment

Ideally, when evaluating therapeutic efficacy in subfertility when treatment is undertaken on a per-cycle basis, the subjects who are enrolled should be receiving treatment for the very first time. This approach of using the first cycle of treatment reduces any potential bias that may result from the experience of treatment in a previous cycle. In many respects, the issues are similar to those raised in the discussion of the cross-over trial. It is possible that subjects, having failed to conceive in previous cycles, belong to a different prognostic category than those undergoing treatment in their first cycle. For example, if one were interested in comparing recombinant FSH with urinary FSH for ovarian stimulation, patients having had previous cycles of FSH treatment may not respond in the same manner as those undergoing treatment for the first time. Alternatively, the trial should be undertaken with ad-hoc stratification for the cycle number with randomization being performed in each stratum. Unfortunately, the literature is replete with trials in which previous cycle performance is not taken into consideration. Only when the research question calls for evaluation in patients with previous failures (e.g. poor responders) should women having had previous cycles be enrolled in efficacy trials.

(v) Sample size estimation

A fundamental, and often overlooked, question facing investigators of any therapeutic trial is to determine the number of subjects that would be required to adequately test the hypothesis regarding treatment efficacy. Easy to use software programmes are widely available to assist investigators in calculating the sample size needed. An estimation of the control event rate is required as a starting point and is often obtained from the literature or from pilot studies. The experimental event rate is selected based on the difference in the two rates the investigator wishes to detect that will have clinical value. The hypothesis testing thresholds to avoid type I and type II errors are selected (usually 0.05 to 0.2 respectively for most trials, although more conservative limits can be assigned depending on the level of assurance required by the investigators) and the testing method of using a one-sided or two-sided method is chosen. Unfortunately, for the event rates that are commonly expected in subfertility research, the calculated sample size is often prohibitively large. For example, if the control event rate is 25% and the experimental event rate is 30% (producing a treatment difference of 5% which, in assisted reproduction is a meaningful difference), a sample size of 2500 would be required using a two-tailed test with the probabilities of type I and type II error set at 0.05 and 0.2 respectively. Accruing this number of subjects would require several years to complete the trial.

Consequently, in everyday practice, smaller trials are usually conducted because they are easier to complete in a shorter period of time. Unfortunately, they are insufficiently powered to test the null hypothesis of no difference between the two interventions, leading to erroneous inferences. Alternatively, investigators will select different outcome events on which to base the sample size calculations. A commonly used strategy is to select the number of oocytes retrieved as the desired endpoint; a clinically important difference that is often selected is two oocytes. The desired sample size is much smaller and is easily achieved because each patient contributes 8–10 oocytes to the total pool. However, this approach is methodologically incorrect because the subject being randomized is the patient and not the oocyte.
It is important when designing the trial, and especially when reporting the results, that a clear indication be provided on how the sample size was calculated. In this manner, the inferences that are drawn from the results can be more appropriately assessed.

(vi) Lack of superiority versus equivalence

Efficacy trials are generally designed to answer the primary question of whether the new (experimental) intervention is superior in some way to the standard (control) intervention. The aim in such superiority trials is to rule out equality between the interventions by rejecting the null hypothesis that there is no difference between the two treatments. A common mistake, when a superiority trial fails to reject the null hypothesis, is to conclude that the two interventions being compared are equivalent. Take, for example, a new treatment that results in a pregnancy rate of 28% compared with the rate of 20% with the standard treatment evaluated in a trial of 100 subjects. This difference of 8% is not statistically significant with this sample of subjects and may lead one to conclude that the two treatments are the same (i.e. equivalent). This is an incorrect interpretation because although lack of proof of superiority may be consistent with equivalence, it not proof that equivalence is present. If the sample size was increased to 1000 subjects, the same magnitude in effect size would be statistically significant.

The goal in an equivalence trial is to rule out differences of clinical importance in the primary outcome between the two treatments. The null hypothesis (in contrast with that in a superiority trial) is stated differently as a minimum difference that is acceptable and would render the two treatments interchangeable (Daya, 2001b). The execution of an equivalence trial becomes an ambitious exercise because, by necessity, it requires a much larger sample size than a superiority trial and is less feasible to conduct, especially when pregnancy (a relatively less common event) is the primary outcome of interest.

The increasing availability of therapeutic choices resulting from advances in subfertility research poses a problem in determining whether these options are equivalent for use in clinical care. The current practice of conducting small comparative trials that fail to show superiority of the new intervention should be avoided when trying to evaluate consistency. A common mistake, when a superiority trial fails to reject the null hypothesis, is to conclude that the two treatments are the same (i.e. equivalent). This is an incorrect interpretation because although lack of proof of superiority may be consistent with equivalence, it not proof that equivalence is present. If the sample size was increased to 1000 subjects, the same magnitude in effect size would be statistically significant.

The goal in an equivalence trial is to rule out differences of clinical importance in the primary outcome between the two treatments. The null hypothesis (in contrast with that in a superiority trial) is stated differently as a minimum difference that is acceptable and would render the two treatments interchangeable (Daya, 2001b). The execution of an equivalence trial becomes an ambitious exercise because, by necessity, it requires a much larger sample size than a superiority trial and is less feasible to conduct, especially when pregnancy (a relatively less common event) is the primary outcome of interest.

The increasing availability of therapeutic choices resulting from advances in subfertility research poses a problem in determining whether these options are equivalent for use in clinical care. The current practice of conducting small comparative trials that fail to show superiority of the new intervention should be avoided when trying to evaluate consistency. A common mistake, when a superiority trial fails to reject the null hypothesis, is to conclude that the two treatments are the same (i.e. equivalent). This is an incorrect interpretation because although lack of proof of superiority may be consistent with equivalence, it not proof that equivalence is present. If the sample size was increased to 1000 subjects, the same magnitude in effect size would be statistically significant.

(vii) Definition of pregnancy and the implantation rate

There is much variability and no consensus on the definition of pregnancy that is of most clinical relevance. In part, the lack of agreement stems from the perspective of the interested party (such as the investigator, patient, policy maker, insurance provider and so on). Market forces and the quest to be the best also add to this problem.

Reporting of biochemical pregnancy (i.e. detection of hCG in the blood) creates confusion because of the possible carry-over effect from exogenously administered hCG that could be detected in the blood test. The more relevant and clinically appropriate definition is the clinical pregnancy rate which requires the detection of rising hCG titles followed by ultrasonographic demonstration of a gestational sac. In the event of an abnormal pregnancy (e.g. miscarriage, ectopic pregnancy, hydatidiform mole and so on), confirmation by pathological examination is desirable. There should be agreement on when the ultrasonographic examination is carried out. One option for consideration is 6 weeks gestation (i.e. 4 weeks after embryo transfer) using a transvaginal ultrasound transducer.

Another relevant outcome event is the ongoing pregnancy rate, which has not been defined consistently but is understood to describe a pregnancy that is less likely to miscarry. The likelihood of having a miscarriage declines with gestational age and is significantly reduced when there is evidence of fetal viability at a gestation of at least 10 weeks. To provide more assurance, it may be appropriate to select a slightly higher gestational age to define ongoing pregnancy as, perhaps, one that has evidence of fetal cardiac activity on ultrasonography at 12 weeks gestational age.

A third option to consider is the live birth rate, which has more relevance to patients. The problem with this endpoint is the additional time and effort required to collect the data, given that most patients would have been referred to their own physicians or obstetricians for antenatal care after the first trimester. In many situations, the patients may have had to travel long distances to access treatment, making it more difficult to obtain reliable outcome information after they have been discharged from care from the fertility clinic. Nevertheless, birth outcome data are very important to have and every effort must be taken, especially in clinical trials, to obtain this information.

The high rates of multiple pregnancy with assisted reproduction have raised many concerns, some of which are being addressed by the trend to transfer only one embryo in a cycle. Multiple pregnancy, as an outcome event, is now viewed as an undesirable consequence of trying to improve pregnancy rates with the transfer of several embryos. This adverse effect should be highlighted by selecting a singleton live birth as the outcome of interest, defined as the birth of a single, live baby.

By using all these definitions for outcome events, a consistent approach to reporting outcomes can be instituted so that reviewing the evidence can be undertaken with more consistency.

Increasingly, investigators are including implantation rate in their summary statistics on outcomes. This endpoint is calculated by simply aggregating the numbers of embryos transferred in the subjects in one arm of the study and using this figure as the denominator. The aggregate number of gestational sacs seen is used as the numerator to obtain the implantation rate. Clearly, this is incorrect use of the data because the unit of analysis is wrong. It is the patient, and not the embryo, who is being randomized and treated. Consequently, the use of implantation rate should be discontinued because it is a misleading outcome indicator.

Summary

The paper by Vail and Gardener (2003) is timely and raises important issues in trial methodology and analysis. The
approach in subfertility research should be to adhere to the CONSORT statement when conducting the trial and reporting the results. This statement will require periodic modification to incorporate changes that occur in the field. Elements of trial methods that are relevant to subfertility research (such as intention-to-treat analysis, concealment of random allocation, avoiding the cross-over design, focusing on the first cycle of treatment or stratifying for the number of the treatment cycle, accruing sufficient subjects to reach the appropriate sample size and clearly defining the outcome event) need to be addressed in all trials. Only with these approaches can the quality of evidence be improved so that the estimate of the treatment effect becomes more reliable and useful to assist clinicians in determining what is the most appropriate care of their patients.

References