Resolving methodological and clinical issues in the design of efficacy trials in assisted reproductive technologies: a mini-review

J.-C. Arce1, A. Nyboe Andersen2 and J. Collins3,4

1Ferring Pharmaceuticals A/S, Clinical Research and Development, Copenhagen, 2The Fertility Clinic, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark and 3McMaster University, Hamilton, Canada

4To whom correspondence should be addressed at: 400 Mader’s Cove Road RR1, Mahone Bay NS B3J 2E0, Canada.
E-mail: collinsj@aura.com

The validity, importance and relevance of randomized controlled trials depend on identifying an appropriate target population, ensuring adequate power, careful attention to the details of randomization and blinding, and selection of an endpoint that is important to the target population. With efficacy trials more than effectiveness trials, additional constraints are needed to reduce the variability that is typical of clinical practice: a narrowly defined sample, unvarying pre-randomization procedures and post-randomization treatments and follow-up that are as identical as possible for all patients. Efficacy trials comparing ovarian stimulation protocols should have strict protocol definitions, specific concomitant medications and minimal variability between centres with respect to stimulation goals and dose adjustments. Additionally, there should be narrowly defined criteria for administration of chorionic gonadotrophin, type of luteal support, embryo transfer and freezing policies. The goal of efficacy trials is to minimize the variability that is extrinsic to the comparison. When efficacy has been proven, effectiveness trials are needed to determine whether the effect of the new intervention is robust in the variability of typical clinical settings.

Introduction

All professionals who provide infertility treatment are continuously searching for better protocols to improve assisted reproductive technologies. Thus, assisted reproduction procedures are constantly under scrutiny, and the highest level of scientific evidence should guide the treatment strategies. The goal is to identify and implement those approaches that yield the highest birth rate of healthy newborns with the lowest possible risks.

Reflecting the global efforts to improve the quality of randomized controlled trials (RCTs), several recent publications have discussed methodological pitfalls in the design of assisted reproductive technology trials (Daya, 2003; Dickey, 2003; Vail and Gardener, 2003). These publications focused on the selection of a well-defined study population, adequate power and statistical principles, appropriate randomization and blinding, and consistent criteria for post-randomization interventions. This essay deals with comparative efficacy trials and extends the debate to consider the special requirements for efficacy designs compared with designs for effectiveness.

The different objectives of trials that address efficacy versus effectiveness were described by Cochrane (1972). Efficacy trials evaluate whether an intervention does more good than harm in an ideal setting (Can it work?), whereas effectiveness trials evaluate whether an intervention with proven efficacy does more good than harm in the normally turbulent clinical setting (Does it work?) (Haynes, 1999). Alternative terms for these different design objectives include fastidious versus pragmatic (Feinstein, 1983) and explanatory versus management (Sackett and Gent, 1979). With respect to pharmacological interventions, efficacy trials must precede effectiveness trials, because a drug cannot be used in clinical practice until efficacy (and safety) has been shown and regulatory authorities have approved the drug for marketing. On the other hand, effectiveness trials are still needed because, ‘Even if an intervention works astonishingly well in a ‘Can it work?’ study, it may not work well in usual care’ (Haynes, 1999). The design contrast is summarized in Figure 1. The pooled cycles of all efficacy trials in assisted reproductive technology might number a few thousand, while the annual number of cycles in clinical practice may be hundreds of thousands.

The present paper was initiated to address the design choices for planning new comparative trials to evaluate the efficacy of stimulation protocols in assisted reproduction. The competing requirements of effectiveness and efficacy designs add a complicating dimension to previous publications on
the issues involved in RCT design (Daya, 2003; Dickey, 2003; Vail and Gardener, 2003). The methodological issues for efficacy trials in assisted reproduction are discussed under headings that follow the sequence of the trial design: assembly of patients, experimental manoeuvre, outcomes, sample size, and analysis.

Methodological issues in trial design
Once the initial aim of a trial has been developed, consideration has to be given to the design architecture and overall objective of the study. Use of a multicentre approach for RCTs is almost mandatory in order to achieve an appropriate sample size. Additionally, the parallel design has a number of advantages in relation to studies of infertile couples (Daya, 2003). The present paper therefore deals with multi-centre, parallel trials.

Trial objectives
An early planning issue is whether the aim of the trial will be efficacy or effectiveness. Efficacy, fastidious or explanatory are terms which describe a trial designed to address whether an intervention does more good than harm in an ideal setting (Sackett and Gent, 1979; Feinstein, 1983; Haynes, 1999). In these trials, variability between centres, subjects, treatments and procedures is minimized as much as possible. Effectiveness, pragmatic and management are terms that describe trials in a typical clinical setting, with the range of subjects, co-treatments, changes of mind, judgements and fumbles that are normal in clinical practice. Effectiveness trials estimate whether an intervention that may have proven efficacy does more good than harm when used in day-to-day clinical practice.

Superiority, equivalence and non-inferiority
Another planning issue concerns whether the trial comparison will attempt to show superiority, equivalence or non-inferiority. The typical clinical trial design sets out to reject the hypothesis that the effects of two treatments are the same and thereby claim that one treatment is superior to the other. Both treatments may be active, or one may be a placebo. Equivalence trials, where the null hypothesis is that the treatments are clinically different (determined by a pre-specified margin), require a forbiddingly large sample size to show that the treatments only differ within a clinically unimportant margin (Piaggio and Pinol, 2001). Over the past two decades, the non-inferiority design has been proposed to reduce these sample size requirements. The non-inferiority design is often applied where two active treatments are under consideration and non-therapeutic benefits (fewer side-effects, lower cost) might determine the choice, provided that there was no difference in therapeutic benefit (Piaggio and Pinol, 2001). Non-inferiority usually involves ensuring that the two-sided confidence interval of the difference between treatments is entirely above a pre-specified non-inferiority threshold. Note that non-inferiority cannot be inferred from the results of a superiority trial simply because the null hypothesis is not rejected. However, a failed superiority trial design can be converted to a non-inferiority design if the non-inferiority limit has been pre-specified.

Sample size
Virtually all clinical trials have a fixed sample size which is determined before recruitment begins, and recruitment ends when the required sample size is reached. Group sequential designs, however, re-calculate and adjust the sample size during the study based on the overall response variance, or the observed event rates. The group sequential design typically makes use of blinded trial data to determine a final sample size while the trial is in progress. Group sequential designs are distinct from the concept of sequential analysis, in which interim analyses are conducted by a data-monitoring group to determine whether evidence of efficacy or harm indicates that the trial should be terminated early. Note that these interim analyses have an effect on the sample size because each analysis ‘uses up’ a portion of the alpha allowance for significance.
Methodological issues in the assembly of patients

Methodological issues in the assembly of patients concern decisions about the target population, definitions of inclusion and exclusion criteria, and whether important prognostic variables should be considered as stratification factors in the randomization procedures or eventually as covariates in the analysis.

Population

The sample to be studied in a clinical trial should be drawn from the target population to which the findings will be extrapolated. An effectiveness study would include a broad spectrum of patients from a wide range of clinical practices. Effectiveness trial designs accept this variability or noise in order to generalize the results to the majority of clinical patients. Such heterogeneity usually means that outcome points will be estimated with less precision. In contrast, an efficacy study adopts explicit inclusion and exclusion criteria to establish a more narrowly defined study population. The consequently reduced noise allows a more precise estimate of the difference in success rates.

In both types of trials, randomization is expected to minimize systematic differences between the comparison groups with respect to important characteristics which could affect treatment outcome, such as age. Further steps can be taken to ensure balance in these prognostic variables, such as defining the eligibility criteria to eliminate some categories, such as older age. A useful tactic is to stratify the random allocation according to prognostic factors so that equal numbers of patients in each age group (for example) are allocated to each treatment group at the time of randomization. Of course, only a limited number of strata can be incorporated into the random allocation scheme in practice, even in a large trial. Therefore, especially in an efficacy trial, it is critical to select a well-defined homogeneous population and to carefully choose as stratification variables only those prognostic factors that are most likely to influence the primary endpoint. The number of strata should be in accordance with the sample size of the trial. While it is preferable to stratify on important covariates related to subject characteristics at randomization, if it is not done a multiple regression analysis can adjust for the factor should there be an imbalance between the comparison groups (Kernan et al., 1999).

Prognostic variables

In efficacy trials in assisted reproduction, age and basal FSH are among the variables usually considered for stratification or as covariates in multiple regression analyses. The relative importance of some population characteristics may change depending on whether the primary endpoint is ovarian response or pregnancy. While age and basal FSH are modest predictors of pregnancy in assisted reproduction cycles (Bancsi et al., 2003; Jain et al., 2004), antral follicle count is a better predictor of ovarian response (Bancsi et al., 2002; Frattarelli et al., 2003; Popovic-Todorovic et al., 2003a; Bancsi et al., 2004). Other covariates which are less frequently considered for stratification include fertilization procedure (IVF or ICSI), infertility diagnostic category, first cycle versus repeat cycle patients, the number of previous unsuccessful cycles, and body mass index (BMI). The reliability of the association between such variables and the ovarian response or pregnancy remains unproven.

Age of the female partner

Age is clearly a variable to consider for stratification. Available registry data from several countries have provided evidence that the chances of a live birth following IVF/ICSI treatment declines with ages >34 years (SART/ASRM, 2004). Selecting subjects aged <35 years in efficacy trials would decrease variability, but also inhibit the generalization of results to older women who form a large fraction of the target population (Nyboe Andersen et al., 2004). In the assisted reproduction cycles in the USA in 2001, 51% were in women aged ≧36 years and 11% in women aged ≧42 years (SART/ASRM, 2004). A more acceptable approach than exclusion would be to stratify for age according to the categories used by established registries (<table with age categories). In the assisted reproduction cycles in the USA in 2001, 51% were in women aged ≧36 years and 11% in women aged ≧42 years (SART/ASRM, 2004). As will be shown later, the inclusion of a comprehensive age span may complicate post-randomization uniformity because co-interventions may be proposed that are deemed to offset the age effect, such as the day of transfer and number of embryos transferred.

Baseline FSH level

Assessment of basal FSH is routinely done in most clinical settings to evaluate individual ovarian reserve. A recent meta-analysis concluded that basal FSH is a moderate predictor of ovarian poor response but a poor predictor of pregnancy, and that FSH threshold levels are not sufficiently sensitive to predict either poor response or non-pregnancy (Bancsi et al., 2003). Excluding the small proportion of subjects with abnormally high basal FSH levels avoids the need for stratification on basal FSH levels in efficacy trials in which the primary endpoint is pregnancy and the exclusion does little to compromise generalizability.

Antral follicle count

The number of basal antral follicles has frequently been reported to correlate with functional ovarian age (Tomas et al., 1997; Pellicer et al., 1998; Scheffer et al., 1999, 2003). In normally cycling women aged <40 years, antral follicle count has been proposed to be a much better predictor for poor ovarian response than basal FSH (Popovic-Todorovic et al., 2003a; Bancsi et al., 2004). Only 6–11% of patients are identified as abnormal, however, with the use of a threshold at four follicles (Bancsi et al., 2004). As a predictor of pregnancy in IVF cycles, antral follicle count is operationally similar to FSH (Chang et al., 1998; Frattarelli et al., 2003). Also, variability from cycle to cycle in antral follicle counts does not influence treatment outcome (Hansen et al., 2003). Moreover, there are only limited data on reproducibility of antral follicle count assessments (Scheffer et al., 2002; Hansen et al., 2003; Bancsi et al., 2004). Further data regarding intra- and inter-observer variability is needed to establish
whether basal antral follicle count is a robust predictor of pregnancy that should be considered as a variable on which to stratify the randomization procedure.

**Infertility diagnosis**

Different infertility diagnostic categories could contribute to heterogeneity in clinical trials. The diagnostic categories allowed in an efficacy trial should be clearly specified in the inclusion/exclusion criteria to ensure homogeneity of the study population. There is a problem with stratification by diagnostic category, however, in that the diagnostic categories do not necessarily reflect the true cause of the infertility (Smith et al., 2003). Another problem with stratification by diagnostic category is that many patients have more than one such category assigned.

The EISG study illustrates another reason why diagnostic properties may not be a sound basis for stratification (Table I) (Platteau et al., 2004). In that study, ongoing pregnancy rates appear to differ between treatment groups according to the infertility diagnostic category; however, a major post-randomization intervention—the fertilization procedure (i.e. IVF or ICSI) which was stratified at randomization in that study—may explain the apparent differences in success rate by diagnosis (Platteau et al., 2004). Because the choice of ICSI is usually determined by the presence of severe male factor, this treatment is more likely to be a specific treatment than IVF, where there are numerous diagnostic categories and the treatment is empiric unless there is tubal obstruction. The method of fertilization *per se* has been emphasized as an important covariate which should be considered as a stratification variable (Daya, 2003).

Efficacy trial designers might consider limiting eligibility to a single diagnostic category or a single type of fertilization (IVF only or ICSI only). This would improve homogeneity of the population and limit the variety of post-randomization procedures so that they might be identical for all patients. On the other hand, an effectiveness trial would not restrict by fertilization method (IVF or ICSI) if the target population included patients needing both procedures. Also, an effectiveness trial would not restrict by diagnostic category, but limiting inclusion or exclusion criteria to ensure homogeneity of the study population. There is a problem with stratification on the number of cycles, or use cycle number as a time-dependent covariate in a proportional hazards analysis. The concern is that an inappropriate response in previous cycles might be deemed to warrant a change in some aspect of the post-randomization management during study cycles, introducing bias. Bias could also occur if repeat cycle patients previously had not responded optimally to one of the study treatments. There is no evidence to support these hypotheses, however, and opposing arguments are also persuasive. For one thing, excluding only the patients with a previous poor or excessive response from participation in efficacy trials would improve the homogeneity of the sample studied. For normal or good responders, bias would be minimized in the study cycle if the trial protocol called for consistent criteria in the critical post-randomization procedures. Also, on average, success does not decline with repeat assisted reproduction cycles until after the fourth cycle (Meldrum et al., 1998).

The data from controlled clinical trials, which usually exclude previous poor responders, support the registry reports that a limited number of previous assisted reproduction cycles do not affect pregnancy rates. Table II displays the ongoing pregnancy rate in EISG (2000) patients with no previous and with previous treatment cycles, as well as according to the number of previous consecutive unsuccessful treatment cycles. The ongoing pregnancy rate did not differ between naive and non-naive patients, and was not affected in patients with up to three previous consecutive unsuccessful cycles. If strict criteria for the major post-randomization procedures are outlined in the protocol, the inclusion of women with a previous normal response and less than four previous consecutive treatment failures would have no more than a negligible impact on pregnancy rates.

**Other baseline characteristics**

Characteristics such as body mass index (BMI) could influence treatment response, but since most of the relevant trials have prohibited inclusion of women with extremely low or

---

**Table I.** Ongoing pregnancy rate by reason for infertility and method of fertilization (EISG, 2000)

<table>
<thead>
<tr>
<th>Main reason for infertility</th>
<th>HP-hMG (%)</th>
<th>rFSH (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tubal</td>
<td>33</td>
<td>18</td>
</tr>
<tr>
<td>Unexplained</td>
<td>37</td>
<td>15</td>
</tr>
<tr>
<td>Male factor</td>
<td>19</td>
<td>23</td>
</tr>
<tr>
<td>Other</td>
<td>24</td>
<td>15</td>
</tr>
</tbody>
</table>

**Table II.** Ongoing pregnancy rate by previous assisted reproductive technology cycles (EISG, 2000)

<table>
<thead>
<tr>
<th>Total (%)</th>
<th>HP-hMG (%)</th>
<th>rFSH (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve patients (n = 402)</td>
<td>23</td>
<td>–</td>
</tr>
<tr>
<td>Non-naïve patients (n = 323)</td>
<td>21</td>
<td>–</td>
</tr>
<tr>
<td>Previous consecutive unsuccessful cycles</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 cycles (64% of patients)</td>
<td>24</td>
<td>23</td>
</tr>
<tr>
<td>0–1 cycles (85% of patients)</td>
<td>23</td>
<td>24</td>
</tr>
<tr>
<td>0–2 cycles (97% of patients)</td>
<td>22</td>
<td>23</td>
</tr>
<tr>
<td>0–3 cycles (100% of patients)</td>
<td>22</td>
<td>23</td>
</tr>
</tbody>
</table>

*Unpublished data.

*Platteau et al. (2004).*
high BMI, only limited data are available regarding the actual impact of BMI on pregnancy rates. Furthermore, body-weight does not seem to be a major predictor of the ovarian response in ovulatory women to a fixed starting dose of rFSH (Popovic-Todorovic et al., 2003a).

Methodological issues in the manoeuvre

Although randomization is the key element of the manoeuvre, RCTs necessarily involve many other treatments and procedures. Protocol features before randomization are designed to minimize differences between the groups at randomization. Protocol features after randomization ensure that the groups have equivalent observation, management and follow-up. In efficacy trials especially, it is important to standardize all concomitant medications and procedures (pre-randomization and post-randomization) and minimize co-interventions during follow-up. The methodological issues during the manoeuvre are discussed in temporal order.

Pre-randomization procedures

For an efficacy trial, when a homogeneous trial sample has been selected, interventions that could introduce biases in the study should be avoided. None of the interventions occurring from screening to randomization should be allowed to bias the study population in favour of one of the treatments evaluated. For example, in trials comparing gonadotrophin products with and without LH activity, medication administered before randomization, such as GnRH agonist, could induce relative LH deficiency and theoretically could introduce bias into the comparison of gonadotrophins with and without LH activity.

GnRH agonists

Regarding medication initiated pre-randomization, it has been postulated that the type, route, dose and duration of GnRH agonist for down-regulation (in a long protocol) could affect the percentage of patients having LH levels below a certain threshold where fertilization, embryo development and pregnancy would be compromised and early pregnancy loss would be more pronounced (Westergaard et al., 2000, 2001). There is no solid empirical evidence to support this LH threshold concept when using GnRH agonists for down-regulation in ovulatory women. This may in part be due to the fact that the clinical development of GnRH agonists was not aimed at their use in assisted reproductive treatment. The products have been used for this clinical application despite incomplete information on dose–response relationships and the pharmacodynamic impact on folliculogenesis and clinically relevant endpoints such as pregnancy and live birth rates. Among the widely used GnRH agonist products available (i.e. leuprolide acetate, nafarelin, buserelin and triptorelin), dose-finding studies for prevention of premature LH surge during assisted reproduction treatment are available only for subcutaneous daily triptorelin (Janssens et al., 2000).

The lack of evidence about the LH threshold effect on pregnancy does not rule out the need for more explicit study designs to test the hypothesis. For example, dose–response studies are important because higher doses of GnRH agonist may increase the likelihood of very low LH levels and affect the likelihood of pregnancy, an issue that is relevant to planning an efficacy trial to compare gonadotrophins with and without LH activity. In a placebo-controlled trial involving 240 cycles, low dose (15 μg) triptorelin given subcutaneously fully prevented premature LH surges even though it did not suppress the LH area under the curve (AUC) (Janssens et al., 2000). Higher doses of triptorelin (50–100 μg) did decrease the LH AUC, but generated significantly more oocytes than the 15 μg dose. Although the study was not powered to evaluate pregnancy, the pregnancy rates were similar (13, 15 and 12%) from low to high triptorelin doses. This underlines several issues: (i) the optimal dose of triptorelin and other GnRH agonists has not been defined with respect to pregnancy rates; (ii) a cause–effect relationship between GnRH agonist-induced LH decrease and pregnancy outcome has not been established; (iii) there may be other mechanisms by which the type, dose or route of GnRH agonist can impact pregnancy rates.

In general, the potential impact on overall pregnancy rates by the type, dose and route of administration of GnRH agonist should be assessed prior to initiating further efficacy trials in assisted reproduction using the long protocol. Furthermore, when comparing gonadotrophin preparations it needs to be evaluated if bias is introduced in favour of those with LH activity. The data available from the two largest trials indicate that the dose/route of administration of GnRH agonist may affect the overall pregnancy rate (Westergaard et al., 2001; EISG, 2002); however, the effect seems to be similar for various gonadotrophin preparations (Figure 2). The pregnancy rate in both studies was affected by the dose/route of GnRH agonist, but the differences between the two gonadotrophin preparations remained fairly constant by dose/route of GnRH agonist within each study.

Other examples of pre-randomization interventions could be the use of oral contraceptives for programming the cycle or more rarely other medications. Some other pre-randomization conditions (diet and physical activity, for example) may not be easily controlled in practice, but at least the information should be captured. All potential pre-randomization interventions which could affect either the characteristics of the study population or the differential effect of the treatments on the primary endpoint should be evaluated in both efficacy and effectiveness trials comparing types of gonadotrophins.

Randomization

Random allocation and concealment

Random allocation to treatment groups is necessary to ensure that there is no systematic difference between the groups and to validate the statistical methods of analysis. Open allocation methods such as even versus odd birth dates or insurance numbers, and administrative allocations, such as day of the week or clinic location, are liable to systematic bias. They are also insecure because all trial personnel are aware of the allocation sequence (Daya, 2003; Vail and Gardner, 2003).
Concealment of the allocation sequence (What will the next patient receive?) is important to prevent conscious or unconscious steering of patients into or away from the study by clinical staff who might deem without proof that one treatment is more effective than the other. To reduce post-randomization withdrawals, randomization should take place immediately before the beginning of the treatment under evaluation. Arranging for randomization to occur on the first day of treatment reduces post-randomization losses (Schulz and Grimes, 2002).

Blinding
Blinding of treatment allocation minimizes the risk of bias during the trial and helps to assure that the assessment and management of both groups will be equivalent. Blinding is important because trial personnel are naturally susceptible to hunches about the effectiveness of one or both trial treatments and only if they are blinded can anyone be confident that decisions and assessments are not affected by such intuitive influences. Double-blinding in assisted reproductive technology trials is infrequently attempted, and most gonadotrophin trials are either not blinded, or the outcome assessors are blinded to allocation as a reasonable compromise. It is logical to assume that double-blinding would bring about an increased confidence in clinical trial results and in theory this is simply a matter of making equivalent preparations for each drug. However, in reality the investigational drug would need to have indistinguishable primary packaging material compared to the approved comparator. This is very difficult to arrange and most likely would require new qualification studies. True double-blinding would be optimal, but in practice it remains very difficult. A double dummy design is most feasible, but even that would require identical primary packaging material indistinguishable from the approved comparator preparation.

Co-intervention and contamination
In assisted reproduction cycles the likelihood of co-intervention is small because few women are likely to start a cycle when they are temporarily unwell and less likely to accept prescriptions for intercurrent diseases that may coincide with the cycle. Contamination, or choosing to be treated with the comparison treatment, is also unlikely in assisted reproductive technology studies when the randomization and initiation of treatment are on the same day and the interventional treatment is of short duration.

Post-randomization concomitant medications
After the initiation of ovarian stimulation, which would be post-randomization in gonadotrophin trials, assisted reproduction cycle protocols require several concomitant medications to achieve fertilization and implantation. The list includes hCG for triggering final follicular maturation and progesterone for luteal support. Within the same trial, different drug types, formulations, doses, routes of administration and duration of use, might be possible because of local availability, investigator/centre preference or patient’s choice. Potential pharmacodynamic or pharmacokinetic differences among these products are frequently overlooked or not taken into consideration during the design, analysis and interpretation of trials.

While effectiveness trials would embrace variable monitoring protocols and allow different concomitant medication protocols, efficacy trials need to ensure that this source of variability is minimized. Thus interventions should be specified and similar for all treatment groups and all patients from randomization until the end of the study (or at least until the assessment of the primary endpoint). In multicentre trials, differences in post-randomization protocols are a major potential source of variation. Usually, even in efficacy trials, critical decisions have been left to the local investigators or loosely outlined; subject and embryology procedures have been executed in a non-standardized manner according to
centres; and concomitant medication protocols have been allowed to differ among centres and countries, and even among patients at the same centre. This variability and background noise could possibly overwhelm all but the most robust differences between the treatments being compared. This background noise would be likely to favour the less effective drug because noise reduces the chance of observing a difference. The steps taken to minimize the variability in post-randomization procedures must be carefully planned to avoid any type of bias in efficacy trials.

**Ovarian stimulation**

The objective of ovarian stimulation is to produce mature oocytes that when fertilized will yield one or more healthy embryos for transfer to the uterus. Because the number of oocytes is just one step toward the desired outcome, even efficacy trials in assisted reproductive technology cannot take the number of oocytes as the primary outcome. Thus, the strategy of the stimulation must shift from obtaining the maximum number of follicles and retrieved oocytes to focus on stimulation protocols that balance the ultimate efficacy (ongoing pregnancy rate/live birth) and safety aspects [risk of ovarian hyperstimulation syndrome (OHSS)]. As shown in Table III, in the EISG (2002) trial the ongoing pregnancy rate was higher when more oocytes were retrieved, but stabilized at ~25% when seven to 15 oocytes were retrieved. At the same time, moderate/severe OHSS did not occur when fewer than seven oocytes were retrieved and the likelihood was only 0.28% when seven to 15 oocytes were retrieved. Notably, the rate of moderate to severe OHSS was 10-fold higher when 16–19 oocytes were retrieved and 20-fold higher when ≥20 oocytes were retrieved. In line with the proposed strategy, it would be appropriate in efficacy trials for ovarian stimulation to target seven to 15 oocytes per patient. Criteria for dose adjustments in order to achieve this target need to be established. In one study on individual versus fixed doses of FSH for the first week of stimulation, the FSH dose was increased up to 250 IU/day if the leading follicle was <10–11 mm, or in case of asynchrony, defined as >4 mm difference between the leading follicle and the next one in the pool. The FSH dose from day 8 was decreased if an excess of 20 follicles was observed (Popovic-Todorovic et al., 2003b). The predefined stimulation target in that study was 5–15 oocytes. Further refinements of stimulation goals could be suggested as data become available and the goal of stimulation should be clearly established a priori in the study protocol.

**Table III.** Ongoing pregnancy rate and moderate/severe ovarian hyperstimulation syndrome (OHSS) rate by number of oocytes retrieved (EISG, 2000)

<table>
<thead>
<tr>
<th>No. of oocytes retrieved</th>
<th>Ongoing pregnancy (%)</th>
<th>Moderate/severe OHSS (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1–4 (n = 52)</td>
<td>7.7</td>
<td>0</td>
</tr>
<tr>
<td>5–6 (n = 61)</td>
<td>14.8</td>
<td>0</td>
</tr>
<tr>
<td>7–15 (n = 363)</td>
<td>24.5</td>
<td>0.28</td>
</tr>
<tr>
<td>16–19 (n = 99)</td>
<td>25.3</td>
<td>2.02</td>
</tr>
<tr>
<td>≥20 (n = 124)</td>
<td>26.6</td>
<td>4.89</td>
</tr>
</tbody>
</table>

A defined stimulation goal will be achieved for most patients if a minimum of flexibility is incorporated in the study protocol. Excess flexibility, however, could lead to heterogeneous stimulation within each treatment tested, statistical noise, and the results could be difficult to interpret. Some authors have proposed that efficacy trials in assisted reproductive technology comparing different gonadotrophin preparations should be performed using a fixed dose regimen throughout the study (van Wely et al., 2003). There are, however, several limitations associated with the use of completely fixed dose protocols: (i) a low fixed dose will result in a higher rate of cancellations because of inadequate response; (ii) a high fixed dose will lead to many withdrawals due to excessive response or risk of OHSS and also in an increased incidence of OHSS; (iii) the outcome is conditioned by the proportion of patients who respond to the dose selected; and (iv) the focus becomes mis-directed to comparing the effects of drug potency on ovarian response. A strict fixed dose approach throughout the stimulation may work with trials having ovarian response as the outcome, but may not be the ideal for those evaluating pregnancy rates. For comparative fixed dose studies, it may be useful to have as an inclusion criterion that the subject had responded adequately to that dose level in a previous cycle. For the average efficacy study, however, the drawbacks of a fixed dose regimen for the whole stimulation period outweigh any theoretical advantage in reducing variability. A fixed starting dose for at least 5–7 days should be proposed for all efficacy trials.

Providing a minimum level of flexibility to achieve the stimulation goal while minimizing the potential for post-randomization variability is a major challenge. The stimulation protocol should also incorporate pharmacological knowledge about the tested compounds, so that the frequency of dose adjustments is tied to the clinical pharmacology of the drugs, ensuring that potential differences do not bias key decisions within the stimulation protocol. Based on the pharmacokinetics of currently available gonadotrophins, adjustments should not be made more often than around every 4th day, as more frequent dose adjustments would make it difficult to evaluate the actual impact of the previous dose level. It would also be reasonable to suggest that the magnitude of these adjustments should be standardized as much as possible to minimize variability caused by the subjective responses of the clinical investigators. Adjustments based on indirect biological responses (i.e. hormonal changes such as estradiol) should not be proposed unless the pharmacodynamic responses for those parameters are known to be similar for each treatment being compared. Overall, goal-oriented fixed regimens could be a compromise when trying to minimize variation in gonadotrophin dose and still allow some patient individualization.

Selection of the starting dose is less critical in regimens with a fixed starting dose than in continuously fixed dose regimens. The choice of starting dose should be based on the product labelling information about efficacy and safety for the dose level(s), after evaluating additional pharmacokinetic/pharmacodynamic considerations for the specific study. For example, a starting dose of 225 IU for a comparative trial
would be consistent with the labelling for gonadotrophins both in Europe and the USA. That starting dose would also suit most patients, as the majority of patients undergoing ovarian stimulation require doses >150 IU to achieve an optimal response (Popovic-Todorovic et al., 2004).

**Protocol for hCG administration**

The duration of the ovarian stimulation phase, a common secondary endpoint in assisted reproductive technology trials, is determined in part by the criteria for administration of hCG. Most protocols indicate that hCG should be administered to trigger final follicular maturation when a minimum number of follicles of a pre-determined diameter are observed (often with additional estradiol concentration criteria). Basing an action on ‘at least’ a minimum number is permissive, introducing variability because extended ovarian stimulation occurs in some patients. Prolonging follicle growth in this way may, however, impact the pregnancy rates (Clark et al., 1991; Kolbicianakis et al., 2004). To ensure a more comparable clinical situation among all patients in a single trial, a stricter criterion for the timing of hCG administration should be established: ‘when there are X follicles of Y or greater diameter’, not ‘when there are X or more follicles of Y or greater diameter.’ Criteria for triggering hCG based on either estradiol concentration levels, total or per follicle of Y or greater diameter, should not be proposed. The rationale behind this position is that many factors, intrinsic (endogenous LH levels) or extrinsic (type of gonadotrophin preparation with or without LH activity) could affect the levels of estradiol.

Triggering of final follicular maturation could be done with urinary hCG 5000–10 000 IU or recombinant hCG 250 µg. Although some studies report similar therapeutic efficacy of 5000–10 000 IU of urinary hCG with 250 µg of recombinant hCG for triggering ovulation (Driscoll et al., 2000; ERHCGSG, 2000; Chang et al., 2001), there are differences between these two types of preparations in progesterone and hCG levels for 6–7 days after administration (Driscoll et al., 2000; ERHCGSG, 2000). Whether this is clinically relevant is unknown, but the difference is another factor that could increase post-randomization variability.

**Excessive follicular response**

The management of excessive follicular response to ovarian stimulation is another potential source of post-randomization variability. The choice between cancellation and coasting is not made clear by the current evidence. Withholding gonadotrophin treatment (coasting) may in the future be found to reduce rates of OHSS with no important effect on implantation or pregnancy rates (Levinsohn-Tavor et al., 2003), but there is no evidence from RCTs. An effectiveness trial would embrace variable protocols for the 5–10% of cycles at high risk. For an efficacy trial, however, a strict approach would be needed to minimize variability. One tactic is to prohibit coasting and cancel cycles that meet criteria indicating increased risk for severe OHSS, such as estradiol above a certain level and more than a specified number of follicles. An alternative approach would be to clearly define a priori the policy for coasting.

**Luteal support**

The range of products available for luteal support could contribute to post-randomization variability. A meta-analysis of RCTs assessing the efficacy of luteal supplementation has identified effects of different formulations, routes and doses of progesterone and hCG on outcomes (Pitts and Atwood, 2002). Pregnancy rates were higher with intramuscular progesterone than with vaginal progesterone. Unless intramuscular progesterone is commercially available, however, and produced according to good manufacturing practices, it is not suitable for comparative efficacy trials. The optimal formulation and dose of vaginal products may need further research because few trials compare different vaginal formulations and they involve small numbers of patients. Consistency in the duration (Nyboe Andersen et al., 2002) of luteal support also varies. Despite the evidence suggesting that the type of luteal support may be a confounding factor on pregnancy rates, luteal support has not yet been standardized in efficacy trials for assisted reproductive technology.

**Variability due to concomitant medications**

Unlike some post-randomization procedures, the concomitant medications are a source of variability within a trial that can be minimized by the use of prescribed protocols for sites and patients. Thus, for each indication concomitant medication should be limited to a single product and dose level. The products used should be commercially available, with readily available efficacy and safety information. Steps are needed to ensure that the proposed concomitant medications do not introduce any bias in favour of one of the treatments tested in the trial. Only recommended doses, routes of administration and durations stipulated by product labelling should be implemented in study protocols.

**Post-randomization embryology procedures**

Post-randomization procedures in the embryology laboratory may contribute to variability in the management of patients after randomization. This variability could diminish the contrast between the treatment groups and obscure true treatment effects. If the variability caused the groups to be treated differently, it would be a source of bias.

**Embryo assessment**

Variability in embryo laboratory procedures could obscure differences between ovulation stimulation protocols in assisted reproductive technology trials. Retrospective data indicate that embryo quality is associated with pregnancy rate (Giorgetti et al., 1995) and that embryo score seems to be the best predictor of pregnancy (Terriou et al., 2001). There is, however, no information from large assisted reproductive technology trials about the impact of different treatments on embryo quality. If all trials included an evaluation of how the treatments under evaluation affect embryo quality, the results would help to better interpret the effects of the treatments on pregnancy rates.
Selection of the embryos with the best chance of implantation is based on cleavage stage and morphology. Cell stage (Van Royen et al., 1999; Racowsky et al., 2000), extent and localization of fragments (Alkani et al., 1999; Antezak and Van Blervrok, 1999), blastomere symmetry (Giorgetti et al., 1995; Hardason et al., 2001), and multinucleation (Van Royen et al., 2003) predict the likelihood of implantation. Combining some of these parameters may identify so-called ‘top quality embryos’ which appear to show the highest implantation potential (Van Royen et al., 1999). Furthermore, early cleavage, defined as complete first meiotic division within 25–27 h, correlates with embryo quality and pregnancy rates (Shoukir et al., 1997; Lundin et al., 2001). It is hypothesized that early cleaving embryos are derived from oocytes with better synchronized cytoplasmic and nuclear maturation and represent a higher metabolic fitness and competence (Lundin et al., 2001).

Embryo quality variables

Efficacy trials could provide valuable data for prediction research if they included clearly defined embryo quality variables as intermediate outcomes. Trial logistics should include daily evaluations from oocyte retrieval to day of transfer. Of course, such assessments present various obvious methodological problems: ensuring consistent laboratory processing, estimating the accuracy of the assessments, determining the extent of inter- and intra-observer variability, and timing of the assessments.

Laboratory procedures for handling oocytes and embryos are subject to continuing quality assessment and increasingly strict regulatory control. Given that the quality of assisted reproductive technology laboratories must be certified, it makes sense that during the conduct of efficacy trials in assisted reproductive technology, the focus should be on consistent data collection, rather than on harmonizing procedures in a group of already quality-assured laboratories. Key potential factors such as type and batch of culture media used, which could theoretically interfere with the interpretation of morphological variables, could be included among the variables assessed. Consistent with the obligation to use approved concomitant medications, centres should use only commercially available culture media (or even better if feasible, the same culture media) in efficacy trials. All embryos generated in the study should be evaluated and followed until transfer during fresh or frozen cycles or disregarded.

Selection of embryos for transfer

There is no relevant literature on the most fundamental issue, intra- or inter-observer variability in embryo assessment. Measures of the accuracy of the prediction of implantation of an embryo include sensitivity and specificity, likelihood ratios and the area under the receiver operating characteristics curve. All of these measures indicate whether the variable or the prediction score has any value beyond chance alone. In order for the embryo quality data from efficacy trials to be useful in a prediction, they must first be comparable across centres and embryologists. This requires clear definitions of embryo parameters and endpoints to be established prior to the study. Lack of consistency in the timing of embryo assessments can contribute to erroneous interpretations. As displayed in Figure 3, a variance of a few hours in the timing of the assessment can produce different results for cell stage on that day in the same embryo (Sakkas et al., 2001). Pooling embryology data from different sites would be of limited value unless all study sites were instructed to evaluate the embryos at the same specific time post-insemination. An atlas with representative images to serve as visual aids for key morphological features would also facilitate harmonization of embryo scoring. Because the method of fertilization by IVF or ICSI could affect the fertilization time and thereby the early cleavage state, separate embryo analysis for IVF and ICSI embryos is advisable (Lundin et al., 2001). Thus, efficacy trials should focus on ensuring that embryo quality assessments are made at narrow time-windows in all centres. Innovations in laboratory data imaging now allow for central evaluation of embryos, which can overcome one source of heterogeneity of scoring among centres and embryologists and additionally provide needed data on inter-observer variability.

Post-randomization embryo transfer procedures

As with embryology laboratory procedures and embryo assessments, embryo transfer procedures can be a source of variability that could introduce bias or diminish a difference in pregnancy rates between the treatment groups.

Embryo transfer

Technical factors of embryo transfer that may have an impact on pregnancy rates include the method of embryo transfer, use of dummy transfer, use of ultrasound guidance, position during transfer, difficulty of transfer, prophylactic use of antibiotics, day of transfer, type of catheter, cleansing of the cervix, use of general anaesthesia, use of fibrin sealant, bed rest after transfer, experience of the clinician and presence of uterine contractions (Sallam, 2004).

An evidence-based assessment is difficult because few RCTs have addressed these issues (Sallam, 2004). If efficacy trials were to include details of the transfer techniques, however, the role of embryo transfer variables could be elucidated. For example, in a meta-analysis of
four trials involving 2051 patients, abdominal ultrasound-guided transfer significantly increased the ongoing pregnancy rate over the clinical touch method (Sallam and Sadek, 2003).

For efficacy trials in assisted reproductive technology, the day of embryo transfer should be harmonized among centres. With prolongation of embryo culture, more information can be collected about cleavage and morphological quality. In a meta-analysis clinical pregnancy rate was higher with day 3 embryo transfer compared to day 2 transfer, but the difference was not large and the live birth rate was similar (Blake et al., 2004). A similar clinical pregnancy rate was also found when comparing cleavage stage (day 2–3) and blastocyst stage (day 5–6) transfers (Blake et al., 2004). At this time it appears to be most appropriate in multicentre trials to harmonize embryo transfer on a specific day. Day 3 could be favoured as this gives additional data on the cleavage, which is more easily assessed than morphology. Data on factors associated with the embryo transfer technique, such as whether it was ultrasound-guided, involved dummy transfer, was a difficult transfer, location of transfer, type of catheter and use of routine antibiotics could be collected for a post hoc evaluation to see if these factors had an effect on treatment outcome.

**Number of embryos transferred**

The number of embryos transferred is a key issue in the design of effectiveness and efficacy trials in assisted reproductive technology. The present worldwide trend towards lowering the number of transferred embryos reflects a desire to reduce multiple birth rates (IFFS Surveillance, 2004). Cycles with single embryo transfers currently account for ~12% of the IVF/ICSI cycles in Europe, but the proportion is expected to increase, in part through legislation. In women aged <38 years, presence of a single high quality embryo for transfer in the first IVF/ICSI cycle is associated with relatively high pregnancy rates (Martikainen et al., 2001; De Neubourg et al., 2002). While trial designs should specify a fixed number of embryos to be transferred for all patients in both effectiveness and efficacy trials in assisted reproductive technology, a policy limited to single embryo transfer, albeit ideal, would drastically curtail recruitment at this time. The number of embryos transferred is conditioned by several factors, all of which result in a choice which most clinicians prefer to be elective rather than non-elective. Because the optimal conditions are not known until the time of transfer, stratification of the random allocation by the number of embryos to be transferred is not feasible and many violations would occur in practice. A more realistic approach is to adjust for this variable at the time of statistical analysis, equivalent to post-hoc stratification, which is easily implemented in the design and analysis of the trial. However, this approach has its limitations as patients with no embryo transfer would not be included. Ensuring that criteria of embryo quality are collected at embryo transfer would allow for a balanced comparison within each stratum of the number of embryos transferred.

**Frozen embryo transfer cycles**

To strengthen the database arising from trials, it would be important to collect information on the following parameters: embryo quality at freezing, day of freezing, freezing procedure, embryo cryosurvival, embryo quality after thawing, and type of frozen embryo replacement cycle (e.g. natural versus an estrogen/progesterone substitution cycle).

Despite the limitations of data on secondary outcomes, cumulative data from fresh and frozen embryo transfer cycles could be more easily interpreted if the procedural details and endpoints were consistent in both fresh and frozen cycles, if post-randomization protocols were comparable in the frozen cycle and if timing for the additional evaluation of treatment outcome for the frozen cycles was pre-defined in the protocol. A reasonable period of observation could be set at 1 or 2 years from the initial cycle.

**Methodological issues concerning the outcomes**

An RCT can evaluate many intermediate and ultimate outcomes, but the sample size and validity of the trial to evaluate a difference between the interventions relates to the primary outcome. Normally, one primary outcome should be pre-specified. However, a small set of pre-determined primary outcomes can also be managed. Adjustments for multiplicity should then be considered (e.g. by ranking outcomes or adjusting the significance level). Other outcomes are secondary and analyses involving secondary outcomes are no more than hypothesis-generating observations that may suggest the need for further research.

**Primary outcome**

For effectiveness trials, live birth or singleton live birth is the outcome of interest to couples and the appropriate primary outcome of trials (Min et al., 2004). Recognition of the importance of live birth has brought about an increased regulatory demand for products in fertility treatment to show efficacy in terms of pregnancy or live births. This section concerns which primary outcome is most appropriate for efficacy studies.

**Primary outcome of efficacy trials**

Many studies comparing gonadotrophin preparations in assisted reproductive technology have incorporated the number of oocytes retrieved as the primary endpoint (Out et al., 1995; Bergh et al., 1997; Hoomans et al., 1999; Dickey et al., 2002). With the exception of poor responders, the number of oocytes retrieved is not a good predictor for pregnancy. Several observations indicate that the number of oocytes retrieved is an inadequate endpoint for efficacy in assisted reproduction, although it may be an explanatory intermediate outcome that assesses the ovarian response. The number of oocytes does not take into consideration potential differences in the pharmacological actions of the treatments tested. Differences in pregnancy rate have been reported with different treatment regimens, although the number of oocytes retrieved was similar (Ganirelix dose-finding study group, 1998). This finding suggests that other important aspects for implantation may be affected by the treatments tested which
are not accurately reflected by just evaluating ovarian response. In the Ganirelix dose-finding trial the mean number of replaced embryos in the different dose groups ranged from 2.3 to 2.7 (Ganirelix dose-finding study group, 1998). The number of embryos transferred is generally lower at present, however, and pregnancy rate is therefore likely to be even less affected by the number of oocytes. The number of oocytes retrieved, although more proximal than pregnancy to the main pharmacological action of gonadotrophins, is not an appropriate primary endpoint for efficacy trials comparing stimulation protocols unless the objective is limited to an assessment of the impact on ovarian response, an objective usually found in the early clinical path of a drug development programme.

A significant logistic problem when live birth is an endpoint for efficacy trials evaluating assisted reproduction protocols is the lag time from exposure to the measured outcome. During this lag time, many factors not related to the effect of the gonadotrophin used may cause intermediate events that affect the primary endpoint. Unless there is a pharmacodynamic or pharmacokinetic reason to expect a differential effect of the specific protocol (in this case, gonadotrophins) on live birth, pregnancy rate in one cycle is a suitable and realistic surrogate outcome. Pregnancy endpoints include positive β-hCG, clinical pregnancy (often defined as an intrauterine sac with heart beat 4–6 weeks after embryo transfer) and ongoing pregnancy late in the first trimester (viable fetus at 10–12 weeks after embryo transfer). The choice among these pregnancy endpoints should be based on the likelihood of live birth. In the large EISG trial, the probability of live birth (95% CI) was 67% (61–73%) after positive β-hCG, 84% (78–89%) after confirmation of clinical pregnancy with a living fetus at the 7th week of gestation and 92% (88–96%) after ongoing pregnancy with a viable fetus at the 12th week of gestation. The relative rate of pregnancy loss was 20% from β-hCG to clinical pregnancy and a further 9% from clinical pregnancy to ongoing pregnancy (Figure 4).

Given that ongoing pregnancy rate is usually within 10% of the live birth outcome of interest, it is an appropriate primary endpoint for efficacy trials: it best predicts live birth; it is a practical surrogate for live birth; and it is definable within a time-period in which other potential confounding factors are more easily controlled. Live birth data as well as data from related cryopreservation cycles should be collected as follow-up information to support the findings on the primary endpoint.

**Secondary outcomes**

*Types of secondary outcomes*

Assisted reproductive treatment trials involve numerous secondary outcomes such as the following: alternatives to the primary outcome, process variables, intermediate outcomes, adverse events, and costs. Alternatives to the primary outcome might be satisfaction, adoption or resolution. Examples of process variables include stimulation protocol outcomes such as number of gonadotrophin ampoules, days of stimulation, late follicular phase estradiol, progesterone, cancellation rate, number of follicles on hCG day and acceptability of the interventions. Intermediate outcomes in assisted reproductive treatment trials include the number of oocytes retrieved, fertilization rate, number of embryos, implantation rate and pregnancy rate. Note that some intermediate outcomes, such as implantation rate and pregnancy rate, are in the pathway to the live birth outcome. Analyses of some process and intermediate outcomes may help to explain the primary outcome results. Adverse events include pregnancy losses, complications such as OHSS, and multiple pregnancies.

**Analysis of secondary outcomes**

Since most fertility trials have multiple endpoints, numerous analyses may be done, each with the potential for apparently significant findings. With α set at 5%, the probability that a P-value <0.05 might occur simply by chance (a false positive result) is about one in 20 comparisons. Frequently
clinical claims are made as a result of the high proportion of false positive findings that may be associated with multiple endpoints. As noted above, such seemingly significant findings are no more than hypothesis-generating; they merely point to a possible clinical question to be tested in a future RCT. With respect to intermediate outcomes that are in the pathway to the primary outcome, these variables should be adequately described in the study protocol and ranked according to their clinical relevance. Claims about statistically significant differences on pathway secondary endpoints are legitimate if the primary objective of the trial is achieved, a hierarchical order for testing hypothesis related to secondary endpoints has been pre-specified in the trial protocol and as long the precedent hierarchical endpoint is also significant. For all other variables, the apparently significant findings should be considered exploratory in nature.

Methodological issues concerning sample size

Estimating the sample size needed for a trial involves a process whereby the investigator justifies assumptions about the clinical outcome, significance and power as recently described by Hughes (2003). The present discussion will focus on the clinical aspects.

Clinical assumptions

Dealing with assisted reproductive treatment trials and accepting ongoing pregnancy as an appropriate endpoint, efficacy trials are unlikely to show large increments in success. Thus a large sample size is required to demonstrate superiority or non-inferiority of one type of treatment compared with another. Furthermore, to detect a possible difference between treatment outcomes which can be quantified (number of oocytes) requires a smaller sample size than binomial outcomes (such as proportion of patients with number of oocytes above a certain threshold).

The number of patients required to show a statistically significantly higher pregnancy rate increases substantially for each 1% absolute decrease of the estimated treatment difference. Assume the comparison of two binomial proportions involves a two-sided significance level \( \alpha = 0.05 \), and a \( \beta \) level = 20% (power of 80%) and 20% control pregnancy rates. Approximately 300 patients are needed per treatment group to detect a 10% difference between the treatments, 500 patients for an 8% difference and 1000 patients for a 5% difference (Dupont and Plummer, 1990).

Statistical significance versus clinical relevance

Large trials typically involve numerous comparisons among secondary outcomes, some of which may be statistically significant (\( P < 0.05 \)) by conventional standards. The clinical relevance of such statistically significant comparisons among secondary outcomes should be interpreted with caution. In an assisted reproductive technology study involving \( \sim 600 \) patients (with \( \alpha = 0.05 \) and \( \beta = 0.20 \)), apparently significant differences between treatment groups will be common (Armitage, 1989). If these differences were on the order of one oocyte retrieved, 12% fertilization rate, 0.5 mm in endometrial thickness, and 0.5 days in gonadotrophin treatment duration, they could be statistically significant, but probably would have no meaningful clinical importance.

In contrast, statistically significant differences in the rate of adverse events are very unlikely, because even large trials are not powered to evaluate uncommon events (CIOMS Working Group III & V, 1999). The lack of power, however, means that finding no statistically significant difference is not necessarily reassuring. For example, a trial involving 600–700 patients would only be able to detect a statistically significant difference in the incidence of OHSS of \( \sim 4\%--5\% \) (if the baseline incidence is on the order of 2–3%), when clinically it would be interesting to document a difference of \( \sim 1\%--2\% \). Interpretation of statistical significance and non-significance for secondary endpoints should thus consider the clinical importance.

Methodological issues in the analysis

The analysis plan

The analysis plan for a RCT is based on the hypothesis under investigation. The significance level for rejecting the null hypothesis is usually set at 5%. The treatment effect may be expressed as two proportions (the absolute difference), the odds ratio, the relative risk, or the relative hazard.

When the allocation groups are balanced, the analyst sets up the two-by-two table of treatment group by outcome with the statistical comparison being a \( \chi^2 \)-test with one degree of freedom. When the allocation groups are not in balance the analysis plan may call for a logistic regression which takes the potential confounding variables into account. However, if the random allocation was stratified, the primary analysis should be adjusted accordingly, by including these stratification variables in the model as covariates.

With modern drug developments, large differences between drug effects are uncommon and trials frequently fail to find superiority of one drug over another. If the null hypothesis cannot be rejected, a further analysis can be undertaken to evaluate non-inferiority according to guidelines established by regulatory authorities if a non-inferiority limit has been pre-specified and the quality of the trial is acceptable (Committee for Proprietary Medicines, 2001). Non-inferiority can be established if the non-inferiority limit has been defined \( a \) priori and the two-sided 95% confidence interval of the treatment differences lies above that limit as previously described. For example, if treatment A is less costly or has fewer side-effects than treatment B, then it might be accepted as non-inferior to treatment B if it retained 90% of the effectiveness of treatment B.

The analysis groups

Intention-to-treat analysis

The analysis of the primary outcome of a RCT should be by intention to treat (Sackett and Gent, 1979). This strategy is typical in effectiveness trials, where the treatment is evaluated within the variability of the real-life clinical world. In typical clinical settings, the decision about treatment may be altered before the treatment actually takes place, because another event (such as pregnancy) might intervene, or the patient may change her mind for whatever reason, or clinical...
findings may dictate a different choice. Trial results should correspond to these normal clinical realities, because future patients who receive advice based on the trial findings will be at a similar stage in the decision stream. Because intention-to-treat analysis keeps patients in their allocation group even if they have used the comparison treatment, there is a loss of contrast between the groups and on average the experimental treatment is less likely to prevail; a conservative condition. Regulatory agencies have recognized this conservatism and now generally require an intention-to-treat analysis as the primary analysis of benefit even in efficacy trials (Committee for Proprietary Medicines, 2001; Piaggio and Pinol, 2001).

As noted in previous sections, randomization in both effectiveness and efficacy trials should occur as late as possible before the start of the manoeuvre in order to minimize post-randomization losses (Hughes, 2003). In efficacy assisted reproductive technology trials, the groups to be analysed for the primary endpoint are the patients randomized. This strategy legitimately excludes from the analysis patients who for example are lost before randomization through failure to be adequately down-regulated, spontaneous pregnancy or other reasons for early withdrawal prior to starting the treatment tested. This approach is optimal as pre-treatment losses are not attributable to a difference between the treatments being tested, and having to retain these patients in the denominators for the analysis would reduce further the contrast between the treatment groups.

Other analysis groups

The primary benefit analysis should be by intention to treat, but secondary analyses may be based on other group definitions. For efficacy trials, especially early in the stage of drug development, a per protocol analysis involving only those patients in each group who were compliant will help to determine the underlying value of the new intervention. For all trials, valid information about adverse effects requires an analysis according to actual treatment received, regardless of whether the patient was compliant. However, this may introduce bias as it is not strictly a comparison of the randomized groups.

Summary

A successful outcome of assisted reproductive treatment is the delivery of a healthy baby, and therefore there is an increased regulatory requirement that products used for the treatment of fertility should show efficacy in terms of pregnancy or live births. Ongoing pregnancy rate should be the recommended primary endpoint for efficacy trials as it is the best predictor of live birth, but is less affected by influences that are external to the assisted reproduction cycle intervention. When efficacy has been proven, there will remain a need for effectiveness trials that incorporate the variability of typical clinical settings and these trials would have as primary outcome live birth, with or without the qualifiers singleton and term gestation.

Continuing efforts are needed to identify methodological pitfalls in the design of efficacy trials in assisted reproductive technology. Areas for improvement include stricter protocols, use of identical fertility concomitant medications and doses, reduction of variability arising from different centre protocols for stimulation goals, dose adjustments, timing of triggering final follicular maturation, and luteal support, as well as different procedures and policies for embryo transfer and freezing. Within embryology laboratories there should be consistency in handling procedures and in the timing of the assessments of embryo quality if these endpoints are to be included in the trial protocol. Scientific evidence derived from optimal trial designs should guide decisions on treatment strategies.

References


Committee for Proprietary Medicines (2001) Points to consider on switching from optimal trial designs should guide decisions on treatment strategies.

Design of efficacy trials in assisted reproductive technologies.


Submitted on November 7, 2004; resubmitted on January 27, 2005; accepted on January 28, 2005