Chapter | *5*

Comparison of different treatment strategies in in vitro fertilisation: Methodological considerations

Eijkemans, M.J., Heijnen, E.M., de Klerk, C., Habbema, J.D., Fauser, B.C.

Human Reproduction Vol 21 No. 2 pp. 344-351, 2006

Introduction

The public health challenge for IVF today is to increase availability and acceptability and reduce adverse effects without compromising effectiveness. This paper will address the methodological issues in designing a trial to test a less complex protocol against a common version of the standard current protocol.

In vitro fertilisation (IVF) has been the treatment of choice in severe tubal infertility. For most other indications, IVF is applied as a last therapy after the failure of other treatment modalities. The high costs of the treatment, the burden of the ovarian stimulation for the patient and the complications (136), most notably the high chance of a multiple pregnancy and the associated costs, have prohibited the widespread use of IVF as a first line treatment option (117,111). However, the recent introduction of gonadotropin-releasing hormone (GnRH) antagonists has opened novel possibilities for milder stimulation protocols, which are better tolerated by the patient and less costly than the conventional stimulation regimens (26,112). Moreover, there is a growing awareness that the high rate of multiple pregnancies may be greatly reduced by a restricted, single embryo transfer policy (6,40,172,173,43,50). In theory, these developments hold a promise for the future by reducing complications for both mother and child.

Single compared to dual embryo transfer has reduced success rates per fresh embryo transfer cycle, which can only be overcome by establishing a high-quality frozen-thawed embryo program (43). The pregnancy rates per cycle following GnRH antagonist co-treatment have been shown to be slightly, but significantly, inferior to those of the classical GnRH agonist long protocol (26). Nevertheless, the mild stimulation approach might have advantages when evaluated over an entire (multiple cycle) treatment strategy, since the amount of time needed to complete a single IVF cycle is less and the costs of stimulation are reduced (26,112). More cycles could on average be performed in the same period of time for the same amount of money. Due to the better tolerability for patients, dropout rates may be reduced, so that the number of patients reaching pregnancy within a given period of time could very well be higher compared to the conventional ovarian stimulation approach, with similar costs per pregnancy (163). Hence, a mild ovarian stimulation protocol with GnRH antagonist co-treatment could offer a means to compensate for reduced pregnancy chances when single embryo transfer is considered. Applying such an approach, pregnancy rates will be reduced when evaluated per cycle (46,37), but not for a given treatment period, which is more relevant to the patient. The importance of defining success of infertility therapies as live birth per treatment started instead of per cycle has been stressed recently (105). The time has come to seriously reconsider the definition of successful IVF (6), and design future studies accordingly.

We designed a randomised controlled trial to investigate whether IVF using mild ovarian stimulation combined with single embryo transfer is not inferior in clinical effec-

tiveness, more patient friendly and more efficient in cost-effectiveness compared with conventional treatment. In this paper, the design of the study is presented and discussed in detail.

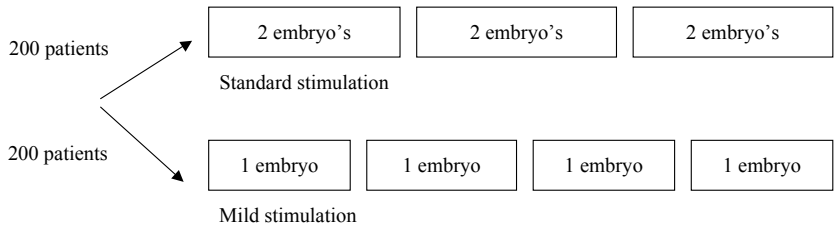
Methodological Considerations

The study is designed as a 2-arm randomised controlled non-inferiority effectiveness trial. The treatment strategies are mild ovarian stimulation with GnRH antagonist co-treatment along with the transfer of a single embryo versus 'standard' ovarian stimulation combined with pituitary down-regulation through the administration of a GnRH agonist long protocol, and transfer of two embryos. In brief, patients with a regular indication for IVF (with or without the addition of intra-cytoplasmic sperm injection (ICSI)), female age < 38 years, normal menstrual cycle (interval between periods 25-35 days) and without severe obesity or underweight (Body mass index 18-28 kg/m^2) were eligible for the study. Two academic medical centres (Rotterdam and Utrecht) participated in the study. Patient data are collected on standard patient-record forms. Patients will be followed-up for a maximum of 12 months treatment plus resulting pregnancy, until 6 weeks post-term. Analysis will be performed according to the intention-to-treat principle. The primary outcome measures are: (1) pregnancy within one year after randomisation leading to term live birth, (2) total costs per couple and child up to 6 weeks after expected delivery (3) overall patient discomfort within one year of randomisation. In the next sections, we will describe the background of the study and motivate the choices that were made in the design of the study.

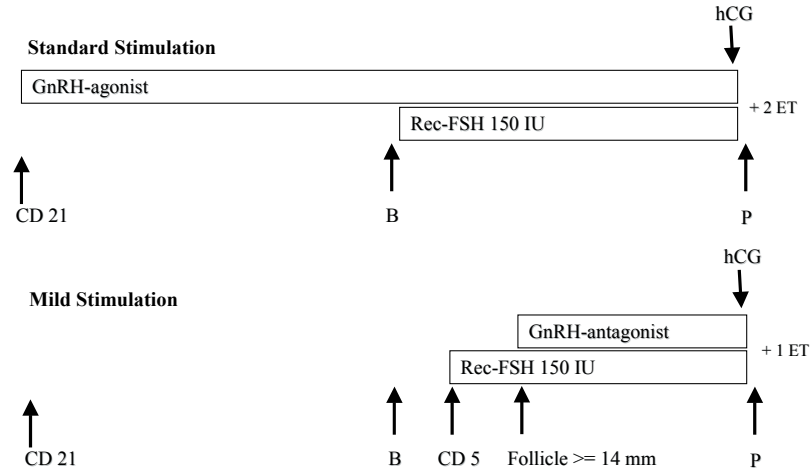
Treatment protocols

The two treatment protocols were executed in a standardised fashion, as depicted in Figure 1. In the standard, GnRH agonist long protocol, two-embryo transfer (ET) arm, standard ovarian stimulation is performed. After approximately 2 weeks GnRH agonist subcutaneous (s.c.) daily, starting during the mid-luteal phase of the pre-treatment cycle (leuproline, 0.2 mg/day; or triptoreline, 0.1 mg/day, depending on the clinic), ovarian stimulation is started with a starting dose varying between patients from 112.5 to 150 IU/day recombinant FSH (recFSH) s.c.. The recFSH dose can be adjusted in subsequent cycles if needed. Human chorionic gonadotropin (hCG) 10,000 IU s.c. is administered for the induction of final oocyte maturation, when the largest follicle reaches at least 18 mm in diameter and at least 1 additional follicle > 15 mm is observed (112). Oocyte retrieval and fertilization are performed according to standard procedures, as described previously (174,175). A maximum of 2 (best quality) embryos is transferred (176). Luteal phase supplementation by progesterone, 600 mg/day, intravaginally is started at the

a) The standard stimulation, 2 ET and the mild stimulation, 1ET arms



b) Standard and mild stimulation protocol per cycle



CD 21:day 21 of the preceding cycle
B: day of bleeding
CD 5: day 5 of the cycle
P: day of follicle puncture for oocyte retrieval

Figure 1. Schematic overview of the study design.

evening of oocyte pick-up and continued until 12 days thereafter. In case good quality excess embryos are available they are cryopreserved and transferred in the subsequent unstimulated cycle, according to standard procedures (177). The maximum number of IVF cycles is 3.

In the mild, GnRH antagonist co-treatment, single ET arm, mild ovarian stimulation is performed by a fixed starting dose of 150 IU recFSH s.c. per day, initiated on cycle day 5. GnRH antagonist (ganirelix, 0,25 mg/dag; or cetrorelix, depending on the clinic) is

administered s.c. if at least 1 follicle ≥ 14 mm is observed (112). The starting day or dose can be adjusted in subsequent cycles. Similar criteria apply for hCG, for oocyte retrieval and fertilization procedures as in the standard group. Only the best quality embryo is transferred. Standard luteal phase support, and criteria to cryopreserve embryos will be applied as in the standard arm. The maximum number of mild IVF cycles is 4.

Background ovarian stimulation

In the standard long-protocol ovarian stimulation, the pituitary-ovarian axis is suppressed through the administration of a GnRH agonist. Subsequently, "high dose" gonadotropins are needed over a long period of time to let the FSH levels rise above the threshold for ovarian stimulation, and the FSH 'window' is widened for an extended recruitment of follicles. A heterogeneous cohort of follicles is recruited in this way.

In mild ovarian stimulation, natural recruitment of follicles is achieved by the inter-cycle FSH rise (178) and exogenous FSH is administered only during the mid-follicular phase, allowing more than one follicle to gain dominance (112). This mode of stimulation interferes less with natural follicle selection and results in a lower number of aneuploid embryos, as shown recently (179).

Trial design

Effectiveness versus efficacy

The current trial is an effectiveness trial, aimed at answering the question: will the treatment strategy under consideration achieve the desired benefits in everyday routine practice. This type of trial is also referred to as a management trial (180) and should be distinguished from an efficacy or explanatory trial, which answers the question: can a treatment work under ideal circumstances (181,182). In an effectiveness trial, inclusion criteria and clinical protocols should resemble everyday reality. We used broad inclusion criteria and different pharmaceutical products, according to the daily routine in the two participating centres. The multi-centre design in itself leads to results that are more relevant to daily practice and less idealized than a highly controlled single centre trial.

2 versus 4 arms

By combining the choice between two ovarian stimulation strategies with the choice between single and dual ET, 4 different combinations are possible, at least in theory. The current study compares only two arms: mild ovarian stimulation and GnRH antagonist co-treatment combined with single ET versus standard stimulation and GnRH agonist co-treatment combined with dual ET. The reason for this choice is both pragmatic (the

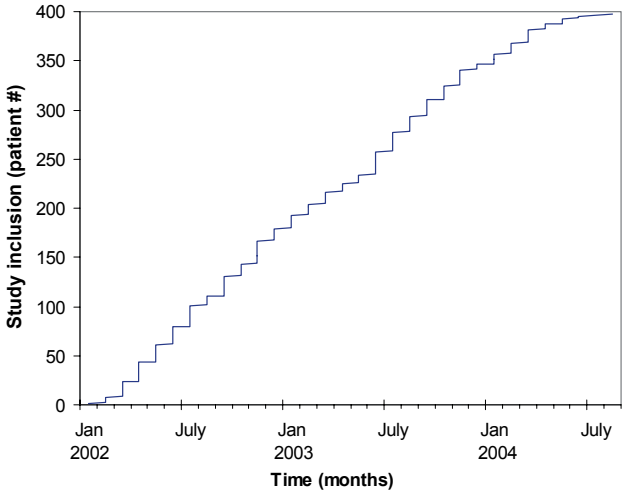


Figure 2. Accrual rate of the trial: Cumulative number of patients included in the study against calendar time.

statistical power of a four arm trial would be much less, given the number of participants that could feasibly be recruited) as well as conceptual (the current comparison is between the standard 'gold standard' treatment strategy in Northern Europe at the time of design of the study (183) and a new, potentially more patient -and child- friendly integrated approach). The possibility to perform more cycles in the same period of time (because of better patient tolerance) renders mild stimulation a suitable combination with single embryo transfer. More cycles means additional pregnancy chances, which can compensate for the reduction in live birth rate per cycle due to the use of GnRH antagonist co-treatment along with the transfer of a single embryo. The acceptance of the proposed treatment strategies is illustrated by the timely accrual of patients into the study as depicted in Figure 2.

A maximum of three fresh IVF cycles was chosen in the standard arm, for practical reasons: it is the number of cycles traditionally covered by insurance in the Netherlands. In the new treatment strategy, one extra cycle was allowed to let patients realize the potential of more cycles in the same amount of time. The cumulative number of cycles completed by the first 200 patients included is depicted in Figure 3.

The other two alternatives have a priori disadvantages: mild stimulation with dual ET might give more pregnancies over time, but does not reduce the twin pregnancy rate. Standard stimulation with single ET does not diminish the physical and psychological burden of the standard stimulation regime. Lower pregnancy rates have been observed (46,37) following the transfer of fresh embryos only, and similar when cryo transfer is also considered (43). A cryo policy is also applied in the current study.

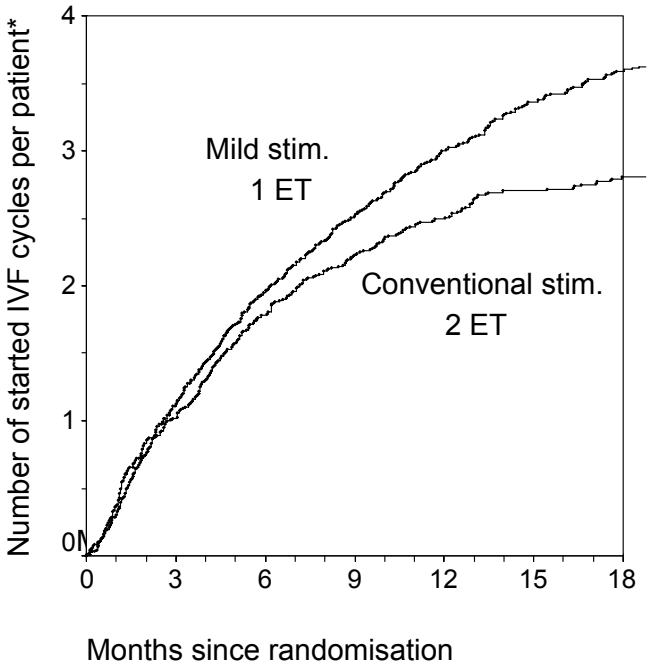


Figure 3. Cumulative number of started IVF cycles per patient against time since randomisation, separately for the standard stimulation + 2 ET and mild stimulation + 1 ET group. Couples who became pregnant are censored: the curve represents the theoretical number of cycles in case no one would become pregnant.

Non-inferiority versus equivalence: one-sided versus two-sided testing

The study is a non-inferiority trial. A non-inferiority trial is appropriate when a new intervention has fewer adverse effects and/or lower costs, and one might accept a little less than the benefit of the standard intervention to gain this advantage in adverse effects or costs. It is well established that the overall costs of pregnancy as well as the complications are greatly reduced by single ET, due to the elimination of twin pregnancies (117,35,184,56,55). If we are able to demonstrate that the mild stimulation/single ET strategy is not worse in clinical outcome compared with the standard strategy, the reduction in multiple pregnancies with their associated higher complications and costs will become decisive in favour of the new strategy. Even when the new strategy would be less effective, the reduction in costs may still make it the more efficient option. Therefore, the focus in the statistical comparison will be to establish that the mild stimulation, single ET strategy is not inferior, within a predefined margin, to the long protocol, dual ET strategy, i.e. a one sided hypothesis.

We calculated the required sample size for the study on a non-inferiority criterion derived from cost-effectiveness considerations. We used the total costs of one IVF treatment cycle of 1,500 Euro from Goverde et al (109), and data regarding costs of

pregnancy, separately for singletons and for twins from Wolner-Hanssen et al (55), 5,300 and 46,000 Euro respectively, including costs of delivery, neonatal care and disability. Furthermore, we chose 45% as the total live birth rate in the standard IVF arm (with a maximum of 3 cycles), of whom 30% are twins, based on annual reports of Utrecht and Rotterdam IVF data, which is compatible with other published Dutch data (185,162). The expected costs per live birth would then be 26,000 Euro. We assumed that the mild stimulation, 1 ET strategy (with a maximum of 4 cycles) could have a lower cumulative live birth rate but also lower costs, due to the absence of twin pregnancies. We tested a range of differences (from -5% to -15%) in live birth rate between the new and the standard strategy and calculated at each specified difference the costs per extra live birth of the standard strategy compared to the experimental strategy. This cost-effectiveness ratio varied from 90,000 Euro (at a difference of -5%) to 25,000 Euro (at -15% difference). At a difference of -12.5%, cost were 35,000 Euro. At this latter figure we (rather arbitrarily, and only for the calculation of sample size) considered the standard strategy no longer acceptable. Therefore, we used a difference in live birth rate between the experimental and the standard strategy of -12.5% as the critical threshold for non-inferiority.

The number of patients should be at least 200 per arm (400 in total) to assure with 80% power that the *lower bound of the 95% one-sided confidence interval* around the difference in live birth rate between the experimental and the standard group will not fall below –12.5%, in case there is no difference in reality. The use of a one-sided alpha is allowed in this case since we have a non-inferiority trial (186). Normally, one-sided confidence intervals are disdained because they prohibit testing a treatment-effect in the direction opposite to anticipation. Here, the opposite direction would be that the new strategy is really inferior. However, it would be of no concern that the new strategy were so inferior that the difference was *statistically* significant: as long as the difference remains -with 95% confidence- within the predefined non-inferiority margin, it is not *clinically* relevant.

Randomisation

Block-randomisation, stratified by clinic, was applied to achieve balance between the two groups within each centre. Randomisation was performed by sealed envelopes available at a central location in both centres. Envelopes were opened by the treating physician at the IVF-intake. As appropriate for an effectiveness trial, the analysis will be according to the intention-to-treat principle, meaning that all patients will be analysed in the group they were randomised to, whether they received the allocated treatment or not. This also applies to patients who cross over to the other treatment group. Again, this is in line with the spirit of an effectiveness trial, since in everyday practice patients may also display a preference for another treatment modality than the one they started with.

Numerator: cumulative live birth as end-point

We defined as primary outcome a pregnancy leading to a term live birth. Term live birth is defined as live birth after a normal gestational length of 37 to 42 weeks. The debate is ongoing whether twins should be regarded as a success (6) or as a complete medical failure. From the clinical perspective, a term twin birth without complications is definitely a success. However, the increased rate of complicated deliveries, preterm births, and low birth weight (all giving rise to increased chances for perinatal morbidity or mortality) associated with twin pregnancies, have led to the opinion that medical intervention in infertility should preferably aim at establishing a singleton pregnancy (6). Our choice of term live birth as primary outcome was made to give a fair advantage to healthy twin births, instead of counting all twins as failure. In this way the increased chance of complications of twins will be expressed in the higher rate of preterm deliveries and discounted proportionally in the outcome.

Denominator: per treatment period versus per cycle

For an effectiveness trial, the natural focus is not on the (technical) results per cycle, but rather on the overall result that a patient may expect over a given treatment period (105). Therefore we have chosen an analysis per treatment period, which will allow the treatment strategy that is best tolerated by the patients and requires the least amount of time per cycle, to realize more treatment cycles -thus more 'chance exposure'- than the other treatment strategy. Dropouts who do not wish to receive any more treatment will be assumed to have a zero chance of the outcome, i.e. a pessimistic assumption (162). In this way we establish a statistical penalty for dropout due to intolerability of the treatment. The time period of analysis will start from the moment of randomisation, to avoid post-randomisation selective dropout.

Health economics considerations

The economic evaluation of the study uses the societal perspective, which is central to health economics as it explicitly considers the question of how to get the most benefit from the scarce resources available to a society (187). It implies that not only medical costs, i.e. costs made within the health care sector, should be included, but also non-medical costs, when relevant. For both medical and non-medical costs, we consider direct costs, defined as directly related to the health care problem (infertility) and treatment (IVF) under consideration as well as indirect costs, which are made after the treatment period.

The costs of the two IVF strategies at hand can be distinguished into two stages:

(1) the costs of IVF treatment itself, starting with the first IVF cycle and ending with the outcome of the last IVF-cycle within a given time period (being pregnant, no pregnancy or drop out);
(2) the costs of antenatal, peripartal and post partum care in women who have become pregnant after IVF treatment.

Since the applied embryo transfer policy during treatment will affect costs during pregnancy, the cost analysis should include all costs from the start of the first IVF cycle up to and including the costs of post partum care. Post partum costs will be counted until 6 weeks post term, since the term period (40 weeks gestation) is the only time horizon that is uniformly applicable to all patients. Costs are measured as the product of health care resource use ('volumes') and cost per unit estimates ('prices').

The costs of IVF treatment are distinguished into direct medical costs in the hospital and outside the hospital, as well as non-medical direct costs. Direct medical costs in the hospital consist of scheduled and unscheduled outpatient visits, number of IVF cycles, personnel time per cycle, use of GnRH analogues and rec-FSH, costs of ultrasound and hormonal monitoring, the embryo transfer procedure and costs associated with complications. Outside hospital costs consist of GP visits, while indirect non-medical costs include travel and time costs and absence from work/sick leave due to treatment or complications. Cost volumes in the treatment stage are recorded with case record forms (CRFs), hospital-based management and budgetary information systems, patient questionnaires and literature. Prices of hospital-based care are estimated as 'true' economic costs (including fixed costs and overhead), as variable costs, and in terms of reimbursement fees. Out of hospital care is priced with reference values for the Netherlands (188). To describe the variability in costs between the two centers, resource use and critical cost parameters are documented for each participating center separately.

The costs of pregnancy and obstetric care can be distinguished into direct medical costs in the hospital (secondary obstetric care) and direct medical costs outside the hospital (e.g. primary obstetric care, GP care, etc.). The pregnant patient will receive questionnaires covering three months periods of their pregnancy, regarding the out of hospital costs. The last questionnaire covers the period around the calculated term date, until 6 weeks thereafter. This means that the neonatal costs are covered for a 6-week period post term. For preterm births, the postnatal period that we consider will therefore be extended resulting in higher costs, as is customary in studies on neonatal care (189).

The incidence of disabilities is markedly increased in multiple pregnancies, and the associated long-term costs might be included in a cost analysis (190). In our study we will add the costs related to long-term health consequences in a scenario analysis, i.e. we will repeat the calculations, with projected costs of life-long disability added to the cost analysis.

Psychological Considerations

Since many decades, outcome measures of medical interventions have not been restricted to rates on survival, mortality, morbidity, and – in reproductive medicine – pregnancies, but have involved other life aspects as well. Many of these are subsumed under the denominator of 'quality of life'. Quality of life measures encompass: (1) global measures of patient satisfaction, (2) multi-dimensional measures of health status (which often include social, psychological and physical dimensions), (3) disease-specific measures that chart problems associated with a specific illness, and finally (4) domain-specific measures that focus on a specific psychological outcome, such as anxiety or depression. Case reports have shown that IVF treatment is sometimes accompanied by intense moments of stress and emotional instability. Aside from being caused by physical stimuli, this emotional instability can also be attributed to the fact that patients swing between hope for a successful pregnancy and fear of failure. When choosing psychological outcomes to be included in an IVF effect study, it therefore seems essential to register negative emotions and moods, rather than assessing psychopathology.

Most psychological effect studies that have been carried out in a medical setting involved patients with a chronic disease. Often, retrospective questionnaires that cover a relatively long period of time are applied in these studies, since short term psychological changes are less relevant in the context of chronic illness. In case of episodic diseases or treatments (e.g. migraine and its medication), diary measures are used to monitor the day-to-day mood fluctuations that may accompany the different stages of the disease and the treatment. While the use of diary measures may reduce recollection-bias (van den Brink *et al.*, 2001), compliance to retrospective questionnaires may be better, as keeping a diary might be a burden to patients. In small studies, interviews are sometimes conducted to explore patients' reactions more thoroughly. Given the complexity of IVF treatment, a combination of retrospective questionnaires and diary measures would be optimal for recording both its long-term and short-term psychological effects.

Many previous studies examining the psychological consequences of IVF treatment have used depression and anxiety as their main outcome variables. These outcomes are usually measured at a few specific moments during IVF treatment (often before or after a treatment cycle) with retrospective questionnaires, like the Spielberger's State and Trait Anxiety Inventory (STAI) and Beck's Depression Inventory (BDI). Other outcomes that are frequently measured with retrospective questionnaires in psychological IVF studies are marital adjustment and self-esteem. Aside from these general adjustment measures, some studies have used infertility-specific stress measures. The Fertility Problem Inventory (FPI) for example, measures five domains of stress that are specific to infertility: social concern, sexual concern, relationship concern, need for parenthood and rejection of childfree lifestyle. Infertility-specific stress measures are believed to be more sensitive

to patient responses to infertility and its treatment than general stress measures. The use of standardized diaries to measure psychological variables is not widespread in the IVF field, with the exception of the Daily Record Keeping Chart (191). This questionnaire has been developed to assess daily emotional, physical and social reactions to infertility treatment.

In the present study a combination of retrospective and diary measures is used to ascertain both the long-term and the short-term effects of IVF treatment. During the first IVF treatment cycle both negative and positive affect are assessed daily with the use of the Daily Record Keeping Chart, which has shown good criterion-related and convergent validity and good internal consistency (192). Additionally, subjects are asked to fill in three retrospective questionnaires several times during the first treatment cycle: After randomisation (baseline), on the first day of ovarian stimulation (to assess the effects of pituitary down-regulation) and after embryo transfer. This last moment is considered to be the most stressful stage of IVF treatment by many patients (193). The retrospective questionnaires are also used to measure possible psychological effects during subsequent IVF cycles. To gain insight in possible side effects related to IVF treatment, self-reported physical discomfort is measured with the somatic subscale of the Hopkins Symptom Checklist (194). The Dutch version of the Hopkins Symptom Checklist has shown adequate to good test-retest reliability, internal consistency and validity (195). Additionally, subjective sleep quality is measured with the Subjective Sleep Quality Scale, a Dutch questionnaire (196), which consists of ten items on various aspects of sleep. This scale has shown good reliability and homogeneity. Finally, stress is assessed with the Hospital Anxiety and Depression Scale (197), which have been developed as a screening tool to detect anxiety and depression in medical patients. The Dutch version of the HADS (198) has shown good test-retest reliability, homogeneity and internal consistency in previous studies.

Discussion

In the current paper we describe the design of a study attempting to answer the question whether the use of a mild ovarian stimulation protocol (using GnRH antagonist co-treatment) combined with single embryo transfer is not inferior to a standard stimulation protocol (using GnRH agonist co-treatment) with dual ET, while resulting in reduced patient discomfort and lower overall costs per pregnancy.

Success of IVF treatment has for long been focussed towards technical aspects of the treatment: The number of follicles harvested, the fertilization or implantation rate. The only outcome of interest to the patient, and therefore the one that should be of interest to the doctor, is whether the procedure will lead to the desired result, a healthy baby

(106,84,105). All other outcome measures are no more than surrogate for this endpoint. Treatments should be evaluated against this outcome measure. A point of ongoing discussion is how to define "healthy". Certainly, pre-term and higher order multiple births are outcomes that should be avoided if possible, but increased perinatal morbidity is also reported following twin pregnancies (6). Should a distinction between twins versus higher order multiples be made or should only a singleton, term delivery be regarded as a success? The current study uses a term live birth as primary clinical outcome measure, which implies that adverse effects of multiple pregnancies will be reflected in a higher rate of pre-term births.

In the field of infertility treatment, the chances of success come in discrete, biologically defined, portions of time, i.e. the menstrual cycle of the woman. Because of the ease of analysis and the simplicity of the cycle concept, the focus in the literature on treatment results has been almost entirely on results per cycle, particularly in IVF. An improvement seems the reporting of cumulative pregnancy rates per patient over multiple cycles (105). However, like in other medical fields, the interest of the patient will be how long it will take until the desired outcome is reached. Obviously, duration of treatment is also related to costs. Cumulative rates over a number of cycles are not very informative if it remains unknown how long it will take to finish the treatment. Thus, the concept to assess success rates per given time interval should be considered. In our study we hypothesized that the mild stimulation method may lead to a shorter duration of a single treatment cycle and therefore the possibility to do more cycles in the same amount of time compared to the standard method.

However, success rates –regardless of how this is defined- still should not be the only outcome used when comparing treatment options. The costs associated with the treatments, the patient discomfort, side effects and complications (mainly ovarian hyperstimulation syndrome and multiple pregnancies as mentioned earlier) should also be part of the equation. In the current study we measure all these aspect in order to give an integrated evaluation of the tested two treatment strategies. In case one treatment strategy is comparable to the other as far as success is concerned, but with a reduced complication rate, and better in the psychological and cost dimensions, it is clearly preferable. In other cases, the costs and patient stress and discomfort will be related to the success rate in a cost-effectiveness analysis. The preferability will then depend on how high the extra costs and psychological burden of the most successful treatment strategy are per extra pregnancy. The design of this study allows assessing all these aspects and obtaining a complete evaluation of two treatment strategies.