NEW DEBATE

Life table (survival) analysis to generate cumulative pregnancy rates in assisted reproduction: are we overestimating our success rates?

Salim Daya

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

Corresponding address: Department of Obstetrics and Gynaecology, McMaster University, 1200 Main Street West, Hamilton, Ontario, Canada L8N 3Z5. E-mail: dayas@mcmaster.ca

The variability in the numbers of treatment cycles couples may undertake with assisted reproductive technology (ART) and the length of time they may have to wait between successive cycles of treatment make the evaluation of treatment efficacy and prognosis complicated. The cumulative pregnancy rate using the life table method of analysis is being used more frequently to estimate the effectiveness of treatment. Although this approach is valid in some areas of infertility research, its use in ART is not appropriate, because the factors necessary for the analysis (particularly the scale for measuring the passage of time and lack of informative censoring) are not satisfied. Consequently, an overestimation of the effect of treatment is produced that may lead to biased decision making. Although there is no easy solution to this problem, several options for summarizing the outcome data are offered: pregnancy rate per cycle, time-limited analysis using proportions, conservative cycle-based cumulative pregnancy rate and real-time-based cumulative pregnancy rate. In this manner, more realistic information can be generated to counsel patients, evaluate the efficacy of treatments, compare rates among centres and guide the formulation of policies for infertility management and resource allocation.

Key words: cumulative pregnancy rate/informative censoring/life table method of analysis/survival analysis/treatment efficacy

Introduction

The growing numbers of options becoming available to treat couples with infertility make it increasingly important for accurate outcome data to be generated and collected, so that inferences can be made about prognosis and efficacy of treatment. The widespread acceptance that assisted reproductive technology (ART) provides effective treatments for infertility has produced a shift in reporting outcome rates from a per-patient to a per-cycle basis. The variability in the numbers of treatment cycles a couple may undertake and the length of time couples may have to wait between successive cycles of treatment make the evaluation of treatment efficacy and prognosis more complicated. Also, given that the most desired outcome of interest in infertility management is pregnancy, it is surprising that there still is much variability in its definition and no consensus on which definition has the most clinical relevance. In the literature, the pregnancy outcomes that have been reported include biochemical, clinical and ongoing pregnancy rates, with increasing attention now being directed to live birth (especially singleton live birth) as the outcome of most value (Daya, 2003). Because each cycle of treatment with ART is relatively short in duration, and women often undergo several cycles of such treatment, the cumulative pregnancy rate using life table analysis is frequently employed to estimate the effectiveness of treatment and guide clinical decision-making.

Here the methodology of the life table analysis will be reviewed and the pitfalls in its use for reporting and comparing outcomes with ART will be discussed.

Life table method of analysis

In some clinical studies evaluating the prognosis of a disease or the outcome of an intervention, subjects with the disease or having received the intervention must be followed for a period of time until the outcome is observed. Ideally, all subjects should be enrolled simultaneously, but, for practical reasons, subjects usually are enrolled sequentially as the study progresses. Consequently, at the conclusion of the study, subjects have been followed for varying lengths of time and three end-points have been encountered: some subjects have dropped out of the study and become lost to follow-up, some have experienced the outcome being evaluated, and the rest have not yet reached this outcome by the time the study is closed.

The data can be summarized in a variety of ways (Norman and Streiner, 2000). Assuming that the outcome of interest in a study is death, one can calculate the mean survival time by
using the data from only those who died. Thus, mean survival time is calculated by dividing the time to outcome by the number of subjects experiencing the outcome. Although this calculation is simple, it may not provide a reliable estimate of survival time because the data from subjects who have not died by the time the study ends, or those who are lost to follow-up, are not included in the analysis. Subjects remaining outcome free at the end of the study are labelled as 'censored'. Including censored subjects when analysing the outcome data, by using the length of time they were in the study, would provide additional information regarding the mean survival time. However, the result is likely to underestimate the true survival time, since many of the censored subjects will have survived beyond the time of study closure.

Another method is to calculate the survival rate up to a specified period of time, e.g. 1-year survival, 5-year survival, and so on. The survival rate is calculated by dividing the number of subjects surviving to a specific time point by the total number of subjects. This approach reduces, but does not eliminate, the effect of censoring.

The life table method makes use of the data from all subjects and is preferred because it provides more reliable estimates of survival than the methods described so far. For each subject, there is a defined point event, often called failure, occurring after a length of time called the failure time (Cutler and Ederer, 1958; Cox and Oakes, 1984). Failure can occur only once in any subject. The life table method was originally devised for analysing light bulb failure times and was then applied clinically for analysing death rates (as the failure event), so that failure rates resulting from cancer could be ascertained and survival estimates could be generated (Berkson and Gage, 1950). This method (usually known as survival analysis) was a significant improvement over the approach of using gross death rates, because it incorporated the actual rates of death observed at various time points following diagnosis or commencement of treatment. Also, this method does not require subjects to enter the study simultaneously and can make use of the data from subjects who have dropped out of the study or are lost to follow-up (Cutler and Ederer, 1958; Cox and Oakes, 1984; Katayama, 1975). The most important aspect of the life table method is the incorporation in the survival analysis of the duration of the time taken to reach the outcome event.

There are two approaches for generating the life table: the actuarial approach and the Kaplan–Meier approach (Berkson and Gage, 1950; Kaplan and Meier, 1958). Although both methods are similar, there are some important differences. In the actuarial method, the study period is divided into equal and arbitrary intervals (e.g. yearly intervals). The probability of death (i.e. the hazard) is calculated by dividing the number of subjects who died in that interval by the number at risk of death at the beginning of that interval. With the Kaplan–Meier approach, the exact time of death is used in the calculation rather than placing death within some arbitrary interval. The cumulative probability of survival (known as the survival function) is calculated over the time period of the study. With the actuarial approach, the survival function is calculated at fixed times (depending on the interval), whereas with the Kaplan–Meier approach it is calculated only when an outcome occurs. Consequently, the survival curve derived from the actuarial approach changes only at the end of an interval, whereas with the Kaplan–Meier approach it changes each time an outcome has occurred. Finally, subjects lost to follow-up or censored require a correction factor in the actuarial approach by assuming that these events occurred in the middle of the interval. Such correction is not required in the Kaplan–Meier approach. With numerous events and shorter intervals, the methods converge.

The life table method was extended to studies in infertility >30 years ago in an effort to standardize methods for evaluating treatment efficacy and avoid the imprecision of crude pregnancy rates (Lamb and Cruz, 1972). An adaptation was required for use in infertility so that pregnancy could be selected as the failure outcome, and the corresponding time to pregnancy could serve as the failure time. Unlike traditional survival curves which show declining survival with time, the survival curves for infertility generate the cumulative pregnancy rate, which shows an increase with time.

Requirements for survival analysis
Survival analysis requires that a number of factors be in place to enable valid calculations of failure times. The requirements will be reviewed as they pertain to infertility research and illustrative examples will be discussed, so that areas of departure from these criteria can be highlighted.

Precise definition of time origin
The time origin (i.e. the starting point) of the study should be defined explicitly for each subject. In randomized trials evaluating therapy, the date of randomization is an appropriate starting time. For prognosis studies, the date of commencement of treatment or diagnosis of disease is an appropriate time origin. For simple follow-up studies, the starting time should be the date of registration in the clinic. In any study of infertility, it is important to enrol subjects as soon as the eligibility criteria have been satisfied. An exception to this rule occurs when randomization is used, in which case the randomization process should be delayed until the intervention is about to commence, to minimize post-randomization interference.

If one is comparing laparoscopic tuboplasty with abdominal tuboplasty in a randomized trial, the randomization should be performed just prior to booking the surgery. The booking to surgery interval should be kept relatively short to avoid subjects dropping out for reasons such as spontaneous pregnancy, change of mind or geographic relocation, all of which could affect the overall conclusion of the study. In a prognosis study of tuboplasty, the day of surgery should be the time origin.

If one is comparing the efficacy of one versus two consecutive inseminations per cycle with donor sperm (up to a maximum of six cycles of treatment) in couples with male factor infertility in a randomized trial, the randomization should be undertaken at the commencement of the first cycle.
and the resulting allocation be maintained for all subsequent cycles. The day of randomization would serve as the point of origin for the follow-up analysis. In a prognosis study of donor insemination, the day of the first insemination could be identified as the time origin.

In trials of ART, decisions about the starting time are problematic. For example, if one is comparing one gonadotrophin with another, the day that ovarian desensitization starts should be counted as the time origin. Unfortunately, although subjects may have been randomized at the beginning of the cycle, some drop out for various reasons and at different points in the cycle before embryo transfer is performed. Investigators incorrectly discard the data from these dropouts, depending on when in the cycle they ceased to remain in the trial. This approach introduces bias in the trial, because of post-randomization withdrawals. Hence the time origin for ART trial needs to be established clearly and in a manner that avoids bias so that the results are valid. Ideally, the time origin should be the day of randomization for efficacy studies, and the day of registration into the programme for prognosis studies.

**Comparability of subjects**

In randomized trials of therapy, all subjects should be as comparable as possible at their time origin to ensure homogeneity of the sample of participants in the trial. This requirement provides assurance that the results are not biased by important differences in the characteristics of the subjects. Comparability of subjects depends on strict adherence to the inclusion and exclusion criteria established for the trial. Subgrouping of subjects is useful when there are variations in important prognostic variables, e.g. severity of disease, female age, duration of infertility, and so on. With ART, an often overlooked issue relates to the subject’s previous experience with treatment. Ideally, when evaluating therapeutic efficacy, especially when treatment is undertaken on a per cycle basis, subjects enrolled should be receiving treatment for the very first time (Daya, 2003). This approach of using ‘ART-naïve’ subjects to evaluate treatment reduces any bias that may result from the effect of the treatment, or the subject’s experience of treatment, in a previous cycle. It is possible that subjects who failed to conceive in a previous cycle belong to a different prognostic category from those undergoing treatment for the first time. Only when the research question calls for evaluation in subjects with previous failures (e.g. poor responders) should women having had previous treatment cycles be enrolled in efficacy trials. Alternatively, the grouping of subjects may be within strata based on the number of previous cycles. This stratification is important because subjects who have failed in one, two or three previous cycles may have different chances of becoming pregnant in a subsequent cycle of treatment.

It is also important to decide *a priori* how many cycles of treatment constitute ‘complete therapy’. In this manner, subjects failing to conceive in the first cycle are expected to go on to the next and subsequent cycles of treatment within this pre-determined definition of complete therapy.

**Scale for measuring the passage of time**

The scale for measuring the passage of time needs to be established clearly. In infertility research, it is usually real time, because that has direct meaning for the couple embarking on treatment. For example, the results from the tuboplasty trial would offer useful information to women contemplating surgery, in terms of the probability of achieving pregnancy after a defined period of time. With such standardization of the measurement of time, two individuals treated in the same way, assuming that all other factors remain the same, should be in a similar state after a lapse of equal time.

Unfortunately, for ART, no emphasis is placed on this time scale criterion. The conventional approach is to use a ‘cycle of treatment’ as the scale for measuring the passage of time. Thus, two subjects, each having had three cycles of treatment, are considered similar for purposes of comparison and data analysis. This approach is valid in situations in which cycles of treatment are based on consecutive menstrual cycles (e.g. consecutive cycles of ovulation induction or intra-uterine insemination). However, although ART cycles are embarked upon consecutively, they are separated by a passage of time that varies from one subject to the next (Figure 1). Thus, one woman may undergo three cycles of treatment with natural cycle ART in three consecutive calendar months, whereas another woman may do so over a 1 year period. This lack of comparability of the time scale is even more readily appreciated with stimulated cycle ART in which one woman may undergo three consecutive cycles of treatment in <1 year, whereas another woman may do so over a period of 2 or 3 years. In this scenario, it is meaningless to compare live birth rates between the two groups.

The issue becomes even more complicated when one is comparing two different treatments measured on different

**Figure 1.** Cumulative ongoing clinical pregnancy rates after three cycles of ART treatment carried out over different durations of time. The curves shown are for three consecutive cycles of treatment in three different scenarios, each with a different time scale. In all three scenarios, the first cycle is performed in the first month. Subsequent cycles are undertaken monthly (square), within 12 months (circle) and within 3 years (diamond). The cumulative ongoing pregnancy rate after three cycles in all three scenarios is the same at 30%.
time scales. For example, when one is trying to evaluate the relative efficacies of tuboplasty and ART treatment for tubal infertility, the pregnancy rates are based on real time for tuboplasty, but are cycle based for ART treatment. Unless these rates can be transformed to a similar unit, the outcomes cannot be compared (Daya, 1995).

The problem is compounded further when the additional contribution from cryopreserved embryos is considered. The constituents of one cycle of treatment will vary from one woman to another, because although the cycle usually involves a fresh embryo transfer procedure, it may also include frozen embryo transfer cycles in women who have surplus embryos cryopreserved. The length of time it takes to complete one cycle of treatment will depend on the number of embryos in cryopreservation and the number of menstrual cycles required to transfer all these embryos. This variability in the duration of a treatment cycle emphasizes the need for standardizing the reporting process by imposing a reasonable time limit within which all cryopreserved embryos should be transferred if no pregnancy has occurred. A 6 month limit, beginning from the onset of treatment with the fresh transfer cycle, would seem appropriate for this purpose.

It is well recognized that fecundity declines with increasing female age. Consequently, unless the time scale of measurement is the same amongst individuals (or can be made similar by valid assumptions), the approach of using the ‘cycle of treatment’ as a surrogate for time is inappropriate for use in survival analysis, because of the variability in the length of real time between cycles of treatment both within and among subjects.

**Clear meaning of failure**

Survival analysis requires that the outcome be dichotomous and well defined. The meaning of failure must be defined explicitly, to enable the consistent calculation of failure times and clear interpretation of the results. In infertility research, the outcome that identifies failure is usually pregnancy. The reporting of ‘failure’ with ART includes some or all of the following: biochemical pregnancy, clinical pregnancy, ongoing pregnancy, live birth and singleton live birth. However, there is much variability and no consensus on the definition of pregnancy that is of most clinical relevance (Daya, 2003).

The relatively uncommon but potentially problematic issue of additional pregnancies arising from cryopreserved cycles requires some discussion. For example, if a woman conceives in her first cycle in which fresh embryos were transferred and then returns 1 or 2 years later and conceives using frozen embryos, she is still in her first cycle of treatment, but now has two pregnancies that are attributed to this cycle. Although this situation is relatively uncommon, it requires a consistent approach to its management. It is not unusual to see reports in which the first pregnancy, having ended as a miscarriage, ectopic pregnancy or any other pathological condition, is replaced by a subsequent successful pregnancy resulting from a frozen embryo transfer. For the purposes of evaluating outcomes from treatment, these additional pregnancies, regardless of the outcome of the first pregnancy, should not be included in the analyses; only the first pregnancy can be attributed to the cycle of treatment.

Until agreement is reached on an approach to reporting outcomes that is consistent, it will be difficult to compare the results of survival analyses across studies of ART.

**Avoidance of informative censoring**

The assumption in survival analysis is that subjects who drop out and become lost to follow-up do so for reasons that have nothing to do with the outcome. In infertility research, subjects may drop out for a variety of reasons unrelated to the outcome. For example, after tuboplasty, reasons for loss of follow-up are relocation to another city or country, marital breakdown, family tragedy, and so on. These reasons are unlikely to influence the outcome because the probability of pregnancy in these dropouts is likely to have been the same as if they had remained in the study. Thus, the assumption that the loss to follow-up is not related to the outcome is satisfied for subjects who drop out for the reasons mentioned.

The data from such subjects who did not exhibit failure (i.e. did not become pregnant) up to the point of dropping out are useful and can be entered into the survival analysis up to that time without bias. The subjects are then excluded from the study and are not considered in the analysis at subsequent time intervals. Similarly, with ART, subjects may initially start treatment and then drop out for reasons such as excessive travelling distance to the clinic, requirement to take time off from work making it difficult to complete the treatment schedule, and cost of repeated cycles of treatment.

If the reasons for the loss of follow-up are related to the outcome, then the estimation of the survival function will be seriously affected. For example, couples who have undergone one cycle of ART treatment may be advised not to pursue further treatment, because of poor prognostic factors such as poor cycle performance, raised basal level of FSH, poor ovarian stimulation, and so on (Olive, 1986). This type of dropping out behaviour, known as informative censoring, will introduce bias into the standard methods used for survival analysis (Clark et al., 2003).

There have been many studies undertaken to try to understand dropout behaviour among subjects undergoing infertility treatment with ART. The issue of poor prognosis was ascertained in one study, in which the characteristics of the first ART cycle among the dropouts were compared with those amongst the subjects who continued for another cycle of treatment (Roest et al., 1998). No differences were observed in female age, fertilization rate, and occurrence of poor oocyte and embryo yield. Similarly, no differences were observed in the second ART cycle among those who dropped out, compared with those who continued for a third cycle of treatment. These observations led the investigators to conclude that there was no selective dropping out of couples with a poor treatment prognosis.

In contrast, in a study conducted in a setting in which up to three cycles of ART are funded by the government, in 25% of women not achieving a live birth, a poor prognosis was identified as the cause of discontinuing treatment despite the fact that ART treatment was offered free of charge.
(Olivius et al., 2004). In another study, the likelihood of pregnancy in a second cycle of ART was studied in relation to the performance in the first cycle. It was observed that, in women who did not conceive or had their cycle cancelled, the likelihood of pregnancy was significantly lower in the subsequent cycle compared with those who conceived in their first cycle and returned for a second cycle of ART treatment (Croucher et al., 1998).

Another aspect of informative censoring relates to subjects dropping out for psychological reasons. In a prospective cohort study undertaken to gain more insight into the psychological aspects of dropping out from ART treatment, it was observed that pre-treatment psychological factors were related to the likelihood of subjects dropping out from further treatment (Smeenk et al., 2004). The treatment-independent pregnancy rates among these dropouts were 26% after the first cycle and 13% after the second cycle of ART treatment. These observations suggest that the prognosis for pregnancy in subsequent cycles is not constant. This study also demonstrated that couples who were denied further ART treatment on the recommendation of their physician were significantly more distressed than those who chose either to discontinue or to continue further treatment (Smeenk et al., 2004). Several studies have shown a relationship between distress and poor IVF outcome (Domar, 2004).

Collectively, these observations emphasize the fact that the reasons for dropping out from subsequent treatments are often related to the outcome, and the prognosis in subsequent cycles is influenced by previous performance. Thus, informative censoring is a serious and significant problem for survival analysis in infertility research, especially when evaluating ART treatment.

**Dropouts who return for evaluation**

In traditional follow-up studies, those who drop out provide data up to the point they were last seen and then are excluded from the study and are not considered in the analysis of subsequent time intervals. In infertility research, a unique problem is presented by the reappearance of those who had dropped out and subsequently became pregnant (Doody, 1993). These women often return to their physician to have their pregnancy confirmed, ensure that it is progressing normally, and receive a plan for ongoing care. Sometimes the original physician becomes aware of the pregnancy when contacted by the obstetrician for information on the couple’s infertility treatment. A non-pregnant dropout subject would not have such follow-up. For the purposes of survival analysis, the returning dropout subject behaves as if she had never left the study. Computer simulation exercises have demonstrated that the cumulative pregnancy rates become significantly inflated from such returning dropouts (Doody, 1993). The bias is more apparent with lower fecundity rates, lower cure rates, higher dropout rates and higher returning pregnant dropout rates. Given the relatively low cure rates with ART, there is no lower limit for returning pregnant rates that would offer a tolerable level of bias. Among women receiving ART treatment, the experience of having undergone intensive monitoring during their cycles predisposes them to return to the clinic for careful monitoring of the pregnancy. Consequently, the use of survival analysis in deriving summary estimates of outcome will result in the effect of treatment perceived as being better because of the inappropriate contribution to the success rates by returning dropouts.

**There is no secular trend**

The advantage of the life table approach is that subjects can be enrolled over time, because the analysis is constructed in a manner that adjusts the starting time for all participants to a common time origin. The assumption with the analysis is that there has been no major change over the duration of the study that would affect who gets into the study, what is done to them and what factors influence the outcome (Norman and Streiner, 2000). If changes occur over this time period, then subjects recruited towards the latter part of the study may differ systematically from those recruited earlier in the study. Such secular changes are seen when diagnostic and/or therapeutic approaches are modified. In infertility research, especially with ART therapy, the field is changing rapidly in terms of ovarian stimulation regimens, numbers of embryos transferred, selection of embryos for transfer, and so on. Secular changes of this nature introduce heterogeneity and may affect the external validity of the results.

**Rates of dropout in infertility studies**

The dropout rates are surprisingly high among couples with infertility, despite their high level of motivation to pursue treatment to achieve pregnancy. The dropout rates range from 23 to 45% in infertility clinics (Gleicher et al., 1996; Redmond and Harrison, 2000), and are even higher (82%) for infertility services provided by generalists (Vanderlaan et al., 1998). High dropout rates have also been reported in ART programmes. For example, if all couples that discontinue ART treatment because of a lack of pregnancy are considered dropouts, the cumulative dropout rates in three studies ranged from 37 to 68% (Haan et al., 1991; Hershlag et al., 1991; Tan et al., 1992). High dropout rates are seen as early as the first IVF cycle (ranging from 9 to 26%), the rates increasing with each successive cycle (Land et al., 1997; De Vries et al., 1999; Olivius et al., 2002; Smeenk et al., 2004). Although it has been assumed that many couples will drop out from therapy because of the high cost of ART treatment, in countries such as The Netherlands and Sweden, where the health care system has fully funded up to three ART cycles, the experience is the same. The cumulative dropout rate after three treatment cycles was observed to be 62% in a Dutch study (Land et al., 1997). Similarly, in a Swedish study, 65% of couples that did not achieve a live birth did not take full advantage of the programme that offered three free cycles of ART (Olivius et al., 2002). These high dropout rates make it difficult for survival analyses to generate estimates of cumulative pregnancy rates that are reliable and precise. Despite such inaccuracy, the numbers of publications utilizing survival analysis to generate cumulative pregnancy rates continue to increase. The inferences being made from such papers are misleading and are likely to misdirect clinical practice.
Example

A hypothetical study will be used to illustrate the point that the high rates of censoring that occur with ART treatment will produce cumulative pregnancy rates that are inaccurate because of overestimation. Consider a study involving 1000 women aged >37 years receiving treatment with ART. The probability of ongoing clinical pregnancy is assumed to be constant at 20%. It is also assumed that, of those who do not conceive after a cycle of treatment, only 50% will return for a subsequent cycle. In the literature, this figure varies from 51 to 57%, depending on which treatment cycle in the series of the first four ART cycles is being considered (Osmanagaoglu et al., 2002). Based on these assumptions, the cumulative ongoing clinical pregnancy rate after four cycles of treatment will be 59% as shown in Figure 2. The dropouts are expected (based on the principles underlying life table analysis) to have the same probability of pregnancy as those who continue treatment. However, it is well established that the dropouts comprise two groups of subjects; one group dropping out for personal reasons, and the other group dropping out owing to medical advice or psychological factors (i.e. informative censoring). One can assume that the former group will have the same prognosis for pregnancy as the group that continues treatment. In contrast, in the informatively censored group, the probability of pregnancy is likely to be much lower and may be zero. For the purposes of this example, it is assumed that the dropouts are comprised of equal numbers of subjects in the two groups. Furthermore, the informatively censored group will be processed with a reduced pregnancy rate (ranging down from 15 to 0%) after the first cycle. This assumption is based on the evidence from the literature that the likelihood of pregnancy is lower in this group of subjects. The resulting cumulative ongoing clinical pregnancy rates for the different scenarios are shown in Figure 3, which also includes the data from the referent case shown in Figure 2. The cumulative ongoing clinical pregnancy rate is much lower than in the referent case, the rates decreasing with the likelihood of pregnancy in the informatively censored subjects. When viewed from the perspective of the worst case scenario (i.e. informatively censored subjects have a 0% ongoing clinical pregnancy rate), the data in Table I show clearly that the relative estimates of ongoing clinical pregnancy rate are much higher, depending on the respective rate chosen for the informatively censored group. Thus, the referent case (i.e. the one that is subjected to the traditional method of life table analysis) represents a relative overestimation of up to 40% in ongoing clinical pregnancy rate after four cycles.

If one assumes that the rate of pregnancy expected in both the informatively censored group and the group which continues with treatment falls by 5% for each subsequent cycle after the first one (an assumption that is not unreasonable), then the cumulative ongoing clinical pregnancy rate generated from the data produces a more dramatic divergence from the referent case, as shown in Figure 4.

The use of the traditional life table method of analysis to generate cumulative pregnancy rates for ART cycles is misleading, because the results are overestimated. It is difficult to determine the magnitude of this overestimation, because the true event rate in the dropouts cannot be ascertained. As discussed above, it is clear that the pregnancy rate in the dropouts is not likely to be the same as in those who continue with treatment. Consequently, the use of the life table method for survival analysis to produce summary measures of outcome in ART is inappropriate.

<table>
<thead>
<tr>
<th>Expected pregnancy rate in informatively censored group (%)</th>
<th>Relative overestimation of cumulative ongoing clinical pregnancy rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>7 11 12</td>
</tr>
<tr>
<td>10</td>
<td>14 23 21</td>
</tr>
<tr>
<td>15</td>
<td>21 31 32</td>
</tr>
<tr>
<td>20</td>
<td>29 40 40</td>
</tr>
</tbody>
</table>

Overestimation is calculated relative to the control event rate assuming that the expected pregnancy rate in the informatively censored group is 0%.
cycle) as the benchmark. Although this benchmark approach
body mass index and undergoing treatment in their first ART
impression of a clinic’s performance with ART. Cumulative pregnancy rates also give an erroneous
rates with this analytical technique as it is currently used. Assumptions used when calculating cumulative pregnancy
interpreted cautiously, because of the bias inherent in the
from information generated from survival analysis should be
assumptions: both informatively censored and continuing with treatment are
20% (circle) or declining by 5% after each cycle (diamond).

Possible solutions
Policy recommendations (such as the number of cycles of
recommend to couples with infertility) made
from information generated from survival analysis should be interpreted cautiously, because of the bias inherent in the
assumptions used when calculating cumulative pregnancy rates with this analytical technique as it is currently used. Cumulative pregnancy rates also give an erroneous impression of a clinic’s performance with ART.

Is there a solution to this problem?
Unfortunately, there is no easy solution because all the
requirements for performing life table analysis cannot be satisfied when applying the methodology to assisted reproduction. However, several options should be considered.

Pregnancy rate per cycle
This approach is simple and straightforward. The pregnancy rate per cycle is calculated by dividing the number of pregnancies by the number of cycles of treatment. The outcome data are summarized as pregnancy rate per cycle (e.g. 30% ongoing clinical pregnancy rate per cycle of ART commenced). It provides a crude estimate of the likelihood of pregnancy for one cycle of treatment, but has limited use, because it does not provide any information on the reduction in rates that may occur with repeated cycles of treatment. Also, it may provide biased estimates of success rates, because of the problem of lack of independence that results when one uses data from multiple cycles of treatment from the same subject as if they were independent events.

Comparison of rates across centres becomes difficult because the case mix (i.e. relative proportions of couples with good prognosis versus poor prognosis) in each centre will vary. One solution to this problem is to perform a subgroup analysis, by calculating the pregnancy rate per cycle in the optimal cases (e.g. women <35 years with a normal body mass index and undergoing treatment in their first ART cycle) as the benchmark. Although this benchmark approach to summarizing outcome rates has value in promoting consistency in reporting, it provides only limited value in guiding decision making for the average couple.

Time-limited analysis using proportions
When comparing two different treatments or management strategies, it may be helpful to set a limit of time (e.g. 1 year) or maximal number of cycles within a specified time frame (e.g. up to three cycles, including the resulting frozen embryo transfer cycles in 1 year), and then calculate the proportion of women achieving pregnancy out of the total number who started treatment in each of the treatment comparison groups. This approach has practical value, because it provides information on the likelihood of pregnancy per woman with up to the pre-determined number of cycles of ART treatment in the time period defined. The outcome is summarized as the pregnancy rate after a pre-determined number of ART cycles in a specified period of time (e.g. a 30% ongoing pregnancy rate with up to three cycles of ART in 1 year). Statistical tests can be applied to the resulting proportions so that inferences on treatment effectiveness can be derived.

Conservative cycle-based cumulative pregnancy rate
If the current method of calculating the cumulative pregnancy rate using life table methodology is to be used as the summary estimate of the effect of treatment with ART, then a very conservative approach should be taken. The recommendation is to assume that all dropouts, regardless of the reason, have a 0% pregnancy rate in subsequent ART cycles. The outcome is summarized as the cumulative pregnancy rate after a specified number of cycles (e.g. a 30% ongoing clinical pregnancy rate after three cycles of ART treatment). This approach may underestimate the cumulative pregnancy rate, but will encourage clinics to be more realistic when counselling couples about prognosis and will discourage claims of treatment and clinic superiority.

Real-time-based cumulative pregnancy rate
A method of dealing with the problem of the current use of survival analysis for reporting ART outcomes is to focus on time to pregnancy as an indicator of the passage of time. This approach involves a departure from the current cycle-based method to the traditional real-time-based method. For a subject censored because of pregnancy, the failure time would be the time to pregnancy calculated as the difference between the time origin and the date of the last menstrual period for the pregnancy. It assumes that ART treatment can be undertaken over repeated cycles to maximize the success rate and includes frozen embryo transfer cycles. It requires some agreement on the average length of time (say 6 months) that a couple may require to complete one cycle of treatment (involving both fresh and frozen embryo transfer procedures). The cumulative pregnancy rate would then be calculated for any length of real time (e.g. 1-year ongoing clinical pregnancy rate). Using this method, the time to pregnancy would include one or more embryo transfer procedures within the allotted time frame. The outcome is summarized as
the cumulative pregnancy rate after a specified period of time (e.g. 30% cumulative ongoing 1-year pregnancy rate).

This approach is more useful, focuses on the couple as the unit of analysis and allows comparison with different therapeutic options in infertility management. It allows one to answer the question of the likelihood of pregnancy after undergoing ART treatment for a specified period of time. Also, it satisfies the time scale requirement of the life table method. However, it does not address the problem of informative censoring. This problem has to be resolved by assuming that all dropouts, regardless of the reason, have a 0% pregnancy rate for the duration of the remaining study period.

Summary and recommendations

Although the use of survival analysis and its cumulative plot to summarize outcome results has a valid place in some areas of infertility research (such as evaluating the efficacy of tuboplasty for tubal infertility), the current method of using it in assisted reproduction is not appropriate, because the overestimation that is produced from informative censoring provides an unrealistic impression of the effect of treatment. In addition, the factors that are necessary for survival analysis, especially the lack of informative censoring, are not satisfied. The time origin often is not precisely defined, leading to variability from one subject to the next, because no emphasis has been placed on the importance of this factor. Investigators simply have relied on the cycle of treatment to serve as an indicator of the passage of time, a function that this surrogate does very poorly. Consensus on the definition of the outcome event or failure (i.e. pregnancy) is still lacking and leads to variability in how the outcome event is reported. The inclusion of returning dropouts into the analysis exaggerates the effect of treatment. Finally, the secular trend in ART has witnessed many changes in treatment protocols, a fact that would result in subjects enrolled later in the study having a different or modified treatment compared with those enrolled earlier.

Although four alternatives have been discussed as possible solutions to the problem, the real-time-based cumulative pregnancy rate is recommended as the best option. It addresses the problem of informative censoring using a conservative approach by assuming that all dropouts have a 0% pregnancy rate. This problem can only be resolved by having access to outcome data among the dropouts (both with and without therapy). Thus, it is important to recognize that where it is reasonable, women who fail to conceive in previous cycles should be urged to undergo more treatment cycles, so that more accurate estimates of outcome after a previously failed cycle(s) of treatment can be generated. Research in this area is urgently required. In this manner, the much needed data on prognosis and treatment efficacy with repeated cycles of ART treatment can be gathered, so that appropriate analytical techniques can be used to summarize outcome measures, thereby maximizing the efficient allocation of health care resources.

Acknowledgements

The author is grateful to Dr John Collins and Joanne Gunby for reviewing this manuscript and for their helpful comments.

References


Submitted on October 13, 2004; resubmitted on February 20, 2005; accepted on February 24, 2005