Interpretation of clinical trial results

The Practice Committee of the American Society for Reproductive Medicine

American Society for Reproductive Medicine, Birmingham, Alabama

This educational bulletin provides background and tips on how to recognize quality trials and then focuses on evaluating the validity, importance, and relevance of clinical trial results. (Fertil Steril® 2006;86(Suppl 4): S161–7. ©2006 by American Society for Reproductive Medicine.)

Evidence from clinical trials is fundamental to ethical medical practice. Along with patient preferences, circumstances, and clinical experience, it is central to effective clinical decision-making. Applying evidence to clinical questions requires filtering in the form of three questions: first, do the trial results reflect true effects of intervention, rather than artifactual ones (validity)? Second, do the results suggest that the intervention is clinically useful (importance)? And third, could the results apply to individual patients encountered in daily practice (relevance)? This educational bulletin provides background and tips on how to recognize quality trials and then focuses on evaluating the validity, importance, and relevance of clinical trial results (Table 1).

BACKGROUND

Chance, Bias, and Treatment Effect

There are three reasons why an intervention may appear to be effective: chance, an accidental event; bias, a systematic deviation from the truth caused by extraneous factors other than the intervention; and truth, a real treatment effect. Chance must always be considered when interpreting trial results and is explored in this document's section on appropriate statistical interpretation. Bias may enter studies of all types but is least likely to be present in well-designed and executed clinical trials. Finally, although results from a valid study may be statistically significant, they may not translate into a clinically important benefit. A true effect may be too small or unimportant to help an individual patient.

Clinical Trials

Clinical trials are experimental studies that compare a specific intervention with an alternative intervention, placebo or no treatment, with measurement of specific outcomes. Random allocation to intervention or control groups is a key step in trial design. Random allocation is designed to balance the distribution of prognostic factors between the groups. Prognostic factors that are linked to the outcome but independent of intervention may confound the study results if they are unevenly distributed between groups. In subfertility, female age and duration of subfertility are typical prognostic factors and potential confounders; examples in a menopause trial include severity of symptoms and time since menopause. A major strength of random allocation is its potential to distribute known and unknown confounders evenly between intervention and control groups. This balance is essential when the outcome of interest occurs independently of treatment, which is common with subfertility and menopausal symptoms.

Maximizing the Value of Time Spent Appraising Studies

Although clinical trials provide the most valid evidence for addressing therapeutic questions, their relevance and quality vary. Guidelines for efficient study interpretation have been published elsewhere (1, 2). Here, the elements of critical appraisal have been organized to first address study validity, then clinical importance, and finally, relevance to your practice (Table 1). It is logical to filter in this sequence because a trial that is of insufficient quality to meet validity criteria may be bypassed without an assessment of importance or clinical relevance. Validity can be assessed from a perusal of the methods (and sometimes the methods section of the abstract) without reading the entire paper, thus making the most of the limited and valuable reading time available to clinicians.

Does the research question specify the population, intervention, and outcomes? Good trials provide a succinct and clear statement of the research question which is paramount to interpreting the results. Subject characteristics, such as stage of disease, gender, age, and ethnicity must be defined before extrapolating from the trial to individual patients or populations. The dose and mode of administration of the intervention determines whether it is relevant to clinical practice. The choice of outcomes or endpoints should be clearly stated. A published clinical study will be used to illustrate this and other key points of this discussion.

> Example: Among infertile women with minimal and mild endometriosis, does laparoscopic resection or ablation, compared with diagnostic laparoscopy alone, increase the prob-

Educational Bulletin
Reviewed June 2006.
Received November 26, 2003; revised and accepted November 26, 2003.
No reprints will be available.
Correspondence to: Practice Committee, American Society for Reproductive Medicine, 1209 Montgomery Highway, Birmingham, Alabama 35216.

doi:10.1016/j.fertnstert.2006.08.017

TABLE 1

Questions to help interpret study results using three filters: study validity, clinical importance, and clinical relevance.

Filter I: Are the Study Methods Valid?
1. Was the assignment of patients randomized?
2. Was the randomization list concealed?
3. Was follow-up sufficiently long and complete?
4. Were all patients analyzed in the groups to which they were allocated?

Filter II: Are the Study Results Clinically Important?
1. Was the outcome of sufficient importance to recommend treatment to patients?
2. Was the treatment effect large enough to be clinically relevant?
3. Was the treatment effect precise?
4. Are the conclusions based on the question posed and are the results obtained?

Filter III: Are the Results Relevant to Your Practice?
1. Is the study population similar to the patients in your own practice?
2. Is the intervention reproducible and feasible in your own clinical setting?
3. What are your patient's personal risks and potential benefits from the therapy?
4. What alternative treatments are available?

ASRM Practice Committee. Interpretation of clinical trial results. Fertil Steril 2004.

ability of live birth (3)? The cited report should clearly define the population, the intervention, and the primary outcome.

Is the question clinically important and unanswered? Good trials address questions that are important enough to involve human subjects, where the value of medical or other alternatives remains in doubt. Papers that are worth reading should also provide evidence that the question has not already been answered through a systematic literature review.

> Example: Endometriosis is diagnosed in up to 68% of women investigated for infertility. Laparoscopic ablation or resection is widely practiced. A systematic review and meta-analysis identified only five cohort studies and one pseudo-randomized trial demonstrating questionable benefit from ablation and no significant benefit from alternative medical treatments (4).

FILTER I: ARE THE STUDY METHODS VALID?

Once it is determined that a study has a reasonable chance of addressing the clinical question, it is time to look closely at the quality of the methods to decide whether the results are valid.

1. Was the assignment of patients randomized?

Random allocation is the cornerstone of a clinical trial. Unless this process is truly impartial, mal-distribution of important confounders between groups may occur. Open random number tables or pseudo-random methods such as chart or social insurance number are insecure and should not be trusted. The most secure methods blind the investigators to group assignment. Two further questions about the balance between groups after randomization are relevant to the overall validity of a trial.

Was randomization effective? Randomization does not *guarantee* a balanced distribution of confounders. The number of subjects and the distribution of important prognostic factors should be similar between the groups. This information may be in the methods but frequently is presented in the first results table. Significant imbalance may reflect insecure randomization or the play of chance. Both should be considered when assessing results.

Were interventions other than the one(s) under study evenly distributed between groups? Co-intervention, the planned or unplanned exposure of subjects to a potentially effective maneuver other than the intervention under study, happens even in carefully executed trials. Reporting such exposures allows the reader to decide if results may be biased by uneven distribution of these post-randomization confounders.

> Example: In the Endocan trial, 16 women in each group reported at least one co-intervention during follow-up. Of the six women having IVF or ovulation induction, five were in the laparascopy only (control) group (3).

Thus, the laparoscopy-only subjects (controls) appear to have received the same or perhaps more intensive co-intervention than the ablation group. This suggests that the effect of ablation has not been exaggerated and may have been diluted by the co-interventions that occurred during followup.

2. Was the randomization list concealed?

Unless it is impossible for recruitment personnel to know which allocation is coming up next, conscious or unconscious steering of patients may introduce imbalance between the groups. The order of allocation must be concealed in addition to ensuring that patients, clinicians, and outcome assessors are blinded, because allocation concealment cannot always be achieved simply by blinding. Third-party randomization by phone or pharmacy is the most secure option. Numbered, opaque sealed envelopes are less expensive and reasonably tamper-proof.

The importance of designs that conceal the order of allocation was illustrated by a systematic review of 250 trials. Those which did not describe the method of concealment, or employed an insecure method, reported treatment effects that were 33% and 41% higher, respectively, than studies reporting secure allocation methods (5).

Example: Randomization to ablation or no ablation of visible endometriosis was done during surgery, after laparoscopic staging of disease, by phone call to a centralized randomization service (3).

Were subjects and assessors blinded to intervention and was a placebo used? Where decisions about treatment are made by care givers and decisions about outcomes involve judgment, blinding is essential to prevent conscious and unconscious bias. Subfertility trials, particularly surgical ones, are rarely blinded. However, even objective outcomes such as pregnancy may be influenced by knowledge of exposure, and for this reason, blinding and the use of placebo are both positive features of a trial.

Example: Although on-going pregnancy at 20 weeks is a well-defined, unambiguous primary outcome, patients in this trial were blinded to the intervention (3). One reason for doing so was to reduce the likelihood of laparoscopy-only patients requesting co-intervention treatments.

3. Was follow-up sufficiently long and complete?

Loss to follow-up of more than 20% of subjects is likely to seriously undermine the validity of results, less than 5% loss is reassuring. For rates in between, it may be helpful to consider how study findings would vary if all lost subjects had either conceived or all had failed to conceive. This "sensitivity analysis" tests the robustness or reliability of findings. If similar proportions of subjects are lost from intervention and control groups, the effects of loss to follow-up are more likely to be balanced.

Example: Nine women withdrew from the laparoscopic surgery group and 12 from the control group of this trial, after randomization. Eighteen of these consented to be called after the routine follow-up time of nine months. None were pregnant or had delivered (3). The numbers lost were evenly balanced and outcomes were actually checked for most withdrawn subjects, suggesting no significant impact on the final results.

Appropriate duration of follow-up is a further important but subjective issue. In studies of postcancer therapy survival, several years may be desirable. Anything less may fail to detect clinically important differences. Conversely, nine months of untreated observation may be unacceptably long for subfertile couples, and the chance that an intervention such as ablation may influence conception after 12 months is small.

4. Were all patients analyzed in the groups to which they were allocated?

An important issue is whether all subjects randomized to intervention or control are included in an "intention to treat" analysis. Subjects who do not complete treatment and may therefore have a suboptimal response and those who switch to the alternate treatment are kept in their allocated group for analysis. In subfertility trials, subjects who have spontaneous pregnancies after randomization but before the intervention would be analyzed with the group to which they were allocated. An intention to treat analysis resembles clinical practice where patients frequently decide to stop or switch treatments. Therefore the results of an intention to treat analysis are relevant to patients having their initial discussion about treatment when their treatment and follow-up is uncertain. If a study fails to include all randomized subjects in this way, it is likely to overestimate the size of the effect of the intervention.

Example: Three women randomized to laparoscopy alone, and four to laparoscopic surgery, did not meet the inclusion characteristics and were excluded from this trial. All others, including those with co-intervention and dropout, were included in analyses, as assigned. No crossovers were reported.

While there is no argument about including all eligible patients in an intention to treat analysis, an argument can be made that inadvertently ineligible subjects also should be included, because failure to ensure eligibility for treatment probably occurs with similar frequency in clinical practice.

FILTER II: ARE THE STUDY RESULTS CLINICALLY IMPORTANT?

Having established that the quality of the study design is sufficiently good to ensure that the results are valid, the next step is to look critically at the results and determine whether they are important enough to matter in clinical practice. In other words, would patients be interested in hearing about this outcome, and is the effect large enough to make a difference in their clinical management?

1. Was the outcome of sufficient importance to recommend treatment to patients?

Clinicians should make their own judgments about the clinical relevance of surrogate outcomes: for example, oocyte number, implantation rate, and positive pregnancy test are not clinically important outcomes in most circumstances. Such surrogate outcomes are often used incorrectly to increase study power and efficiency of follow-up.

In subfertility trials, live birth is the generally accepted primary endpoint. Secondary outcomes, such as multiple pregnancy and neonatal morbidity rates, should also be reported, since they are essential elements of effectiveness.

Example: In the Endocan study, the primary outcome was pregnancy conceived during the follow-up period that carried beyond 20 weeks gestation, a very close approximation to live birth and therefore a clinically useful surrogate endpoint (3).

2. Was the treatment effect large enough to be clinically relevant?

A short summary of treatment effects would be useful before tackling this question. In assessing the occurrence or nonoc-

currence of an event such as pregnancy or disease, four simple expressions are frequently used:

1. the relative risk (RR) is the ratio of the probability of success with experimental treatment over the probability with the control treatment;
2. the risk difference (RD) — the absolute difference between the probability of success with experimental treatment and the probability of success with the control treatment;
3. the number needed to treat (NNT) — the number of subjects that must be treated to achieve one more outcome with intervention than control;
4. the odds ratio (OR) is the ratio of the odds of success with experimental treatment over the odds with the control treatment.

For an event that occurs in six of ten individuals; the rate or probability is 6/10; the odds, however, are 6/4. Odds ratios are easier to calculate but more difficult to interpret because odds are seldom used in clinical practice, where risks or rates are more intuitive. The odds ratio is mainly useful in rare conditions, where its value approaches that of the relative risk. The formulas are outlined in Table 2.

Example: The number of ongoing pregnancies following laparoscopic surgery was 50 out of a total of 172 subjects (0.29) and following control, 29 out of 169 subjects (0.17) in this study. Relative risk (in this case, it is relative benefit) is the ratio of these two rates: 0.29/0.17 = 1.7 (Table 2).

The figures for relative risk and odds ratio are similar in this example:

$$\text{Relative risk (RR): } (50 \times 169)/(29 \times 172) = 1.7 \text{ (95\% CI } 1.\mathit{1}, 2.5)$$

$$\text{Odds ratio (OR): } (50 \times 140)/(29 \times 122) = 2.0 \text{ (95\% CI } 1.2, \mathit{3.3}) \text{ (Table 2)}$$

The measure of effect that makes the most sense in clinical practice is the RD, because it is a natural description of the difference between outcomes and has a straightforward interpretation. Also, RD is the clinically important difference

TABLE 2

The effect of laparoscopic ablation of minimal and mild endometriosis compared with diagnostic laparoscopy alone, on pregnancy >20 weeks gestation in women with endometriosis associated subfertility (3).

	Outcome measure: pregnancy >20 weeks		
Allocation group	**Yes**	**No**	**Total**
Experimental group, laparoscopic ablation	50 [a]	122 [b]	172 [a+b]
Control group, laparoscopy only	29 [c]	140 [d]	169 [c+d]
Total	79 [a+c]	262 [b+d]	341 [a+b+c+d]
Control group event rate (CER)	c/(c + d)	= 29/169 = 0.17	
Experimental group event rate (EER)	a/(a + b)	= 50/172 = 0.29	
Relative risk (RR)	EER/CER	= 0.29/0.17 = 1.71 (95% CI, 1.13, 2.54)	
Odds of event in control group	c/d	= 29/140 = 0.21	
Odds of event in experimental group	a/b	= 50/122 = 0.41	
Odds ratio (OR) (= relative odds)	(a/b)/(c/d) = ad/bc	= 1.98 (95% CI, 1.18, 3.32)	
Relative risk difference (RRD)	EER – CER/CER	= 0.29 – 0.17/0.17 = 0.71 (71%)	
Risk difference (RD)	EER – CER	= 0.29 – 0.17 = 0.12 (95% CI, 0.03, 0.21)	
Number needed to treat (NNT)	1/risk difference	= 1/0.12 = 9 (95% CI, 5, 33)	
Also: NNT	1/(RRD × CER)	= 9	

Formulas for 95% CI are at http://bmj.com/collections/statsbk/index.shtml.

ASRM Practice Committee. Interpretation of clinical trial results. Fertil Steril 2004.

that would be used to calculate sample size in the planning stage of the majority of clinical trials. More importantly, the inverse of the RD is the number needed to treat (NNT), an estimate of how many persons would need to receive the experimental intervention before there would be one more or less event, as compared with the controls. The NNT is usually expressed according to a unit of time during which the treatment is given or effective. Absolute benefit and number needed to treat are crucial to patients choosing treatments because relative risk or benefit may be quite misleading.

Example: The suggestion that pregnancy is about two-fold more likely following ablation than laparoscopy alone seems encouraging. However, these statistics provide little insight into an individual patient's real chance of pregnancy. The absolute effect of treatment must be calculated: the difference in conception rate between groups, in this case $0.29 - 0.17 = 0.12$ (Table 2). In other words, each laparoscopic ablation procedure is responsible for an additional 0.12 of a pregnancy. In order to express this figure as a whole number, the reciprocal of 0.12 is used to give a number needed to treat of $1/0.12 = 8.3$. Rounding upward, approximately 9 women must undergo ablation to achieve one additional pregnancy.

When calculating the NNT, it is customary to raise all fractions, even those below 0.5, to the next highest whole digit. The 95% confidence interval for the absolute difference between rates of pregnancy (0.12) is 0.03 to 0.21. The fact that this range does not include zero corresponds to a statistically significant effect with $P<0.05$. The reciprocals of these figures provide the number needed to treat and its confidence intervals: 9, 95% CI, 5, 33 (Table 2).

An additional attraction of the absolute measures (RD and NNT) is that they are free from the misinterpretations that accompany relative ratios (RR and OR). For example, a 35% *increase* in breast cancer risk (RR = 1.35) before age 35 among oral contraceptive users may be misinterpreted as a 35% *incidence* of breast cancer (6).

Example: In the endometriosis study, what if the numbers were 100-fold smaller? The absolute chances of pregnancy with treatment and control would be only 0.0029 and 0.0017, respectively, and the relative benefit would not change from 1.7. However, the absolute difference in benefit would be only 0.0012, and the NNT would be 833. This treatment effect is no longer of clinical interest, even though it might be statistically significant.

This example highlights the importance for clinicians of focusing on absolute rather than relative effects, in reading study reports and talking to patients.

With this background on treatment effect measurement, clinicians should ask two questions to determine whether the treatment effect was large enough to matter.

What was the size of the treatment effect? The results are not clinically important unless the effect is both statistically significant and large enough to be clinically meaningful. The effect of the intervention on the primary outcome should be sufficiently different from the effect of the alternative that the average patient would have no hesitation in making a choice.

Example: Risk difference was 12% (95% CI 3, 21) and the number needed to treat was 9 (95% CI, 5, 33) (Table 2). (Note that when you use percentage risk differences, the NNT is 100/RD.)

The unit of time in this case, where the treatment was a one-time only occurrence, is the routine follow-up time of nine months. An NNT of 9 during nine months after laparoscopy is quite satisfactory in the treatment of infertility, where live birth rates are relatively low. Since the intervention under study involves only the performance of ablation during an already planned laparoscopy and the proportion of adverse effects is small, it remains only to be seen whether the estimate is robust.

What did the investigators consider clinically important? If a trial is large enough, it may demonstrate statistically significant differences between intervention and control that are too small to have any clinical importance. Examine the methods section to see whether the authors have considered and defined a "clinically significant difference" and whether they used this difference to calculate the sample size for their study (7).

Example: In this study sample size of 330 women with endometriosis-associated infertility was calculated to be needed to detect a statistically significant difference of 15% in the primary outcome (with $\alpha = 0.05$ and $\beta = 0.20$) if the expected probability in the control group was between 15% and 30% (3). Here, the authors imply, but do not state, that they consider a 15% difference in pregnancy beyond 20 weeks to be clinically significant.

Clinicians can make their own judgment about "clinically important differences" because that is exactly how investigators arrive at the estimates for their sample size calculations. If a clinician believes that the anticipated effect size is not clinically important, even statistically significant results would not be clinically useful.

3. Was the treatment effect precise?

Statistical tests are done in order to determine whether a given result might have happened by chance. Over time the statistical test report has evolved into a yes/no answer centered on the conventional 5% probability, while 4% and 6% might be of similar importance. A more useful guide to probability is the confidence interval (usually 95%) because it shows the range of results that might be expected if the study were repeated frequently in the same setting. If the confidence interval is narrow, the study gives a more precise estimate of the true value of treatment. Better precision reduces the uncertainty that goes with applying estimates from a trial to patients, no matter how similar the patients may be to the trial subjects.

Are trial results statistically significant? A statistically significant result is simply one that has an acceptably low risk of occurring by chance and is therefore likely to have resulted from intervention. The probability that a difference is due to chance ("type I error", α) is commonly set at 1/20 or 5%. Statistical testing measures the likelihood that a type I error has occurred and expresses that likelihood as *P* values and/or confidence intervals. The confidence interval estimates the range of possible values within which the true population value would lie, typically with 95% probability. In the following example, confidence intervals for the risk differences between ablation and control groups are provided and interpreted.

> Example: The proportion of patients having a pregnancy beyond 20 weeks with laparoscopic ablation of endometriosis was 12% (95% confidence interval 3, 21) (3). Thus, the chance that the study would detect a risk difference of <3 or >21% is less than 5%. There would be no benefit if the risk difference were zero, but 0 is outside this 95% range, indicating that the result is statistically significant at a level of $P<0.05$.

If no difference is detected between intervention and control, some clinicians (often those interested in carrying out a similar study) will check whether the trial was large enough to detect a clinically significant difference before dismissing the intervention as useless (7).

Did the study have adequate power? The probability that by chance, a study will fail to detect a real, statistically significant difference (β), is often set at 0.1 or 0.2. In other words, the investigators accept a 10% or 20% chance that a real treatment effect exists but will remain undetected (type II error).

Few clinicians need to take an interest in these post-hoc power estimates, but analysis programs are available on the Internet to simplify the calculations. If the power to detect a difference of the reported size were, say, less than 60%, then additional adequately powered studies are needed to answer the clinical question.

4. Are the conclusions based on the question posed and the results obtained?

Once study validity, clinical importance, and statistical significance have been evaluated, it is time to weigh conclusions. Has the primary question been answered, and how confident are the investigators of their answer's validity? Be wary of trials that report no difference in the primary outcome but emphasize a (statistically significant) secondary endpoint. Remember that if enough comparisons are made, some will appear to be statistically significant by chance: one in 20, if α is set at 0.05. If comparisons are made between subgroups of patients after trial design and execution (*post-hoc*), chance findings that seem significant are more likely. Consider these *post-hoc,* subgroup analyses to be hypothesis generating, not hypothesis testing. They are legitimate only to the extent that they point the way to a promising new study to test the finding in an independent setting.

FILTER III: ARE THE RESULTS RELEVANT TO YOUR PRACTICE?

1. Is the study population similar to the patients in your own practice?

Enrollment in a trial is based on explicit criteria that are often narrow. These criteria must be carefully considered before extrapolating trial results to individual patients.

> Example: Age 20–39, >12 consecutive months of unprotected intercourse without pregnancy, evidence of ovulation, total motile sperm count per ejaculate >20 million, no previous surgery for endometriosis, no medical treatment for endometriosis over the prior nine months, no ovulation induction or intrauterine insemination in the prior month or other medical or surgical fertility treatment in the prior three months, no previous oophorectomy or salpingectomy, no history of pelvic infection, and no severe pelvic pain (3).

Those outside these boundaries may respond to treatment in different ways. One evidence-based medicine book suggests a different question to achieve the same consideration: is your patient (or your practice) *so different from* the study patients or practices that the study results could not apply (1)?

2. Is the intervention reproducible and feasible in your own clinical setting?

The nature and components of the intervention should be clear enough to indicate whether the intervention is feasible. Is it available locally to be purchased or acquired? Is it affordable in monetary and time costs? Is it accessible without further training? Direct and indirect costs can be forbidding limitations on the feasibility of an intervention.

> Example: Laparoscopic surgical treatment involved the destruction of all visible implants and the lysis of adhesions. Instrument choice was left to the surgeon. In the diagnostic laparoscopy group, no destruction or lysis was allowed. Operative laparoscopy increased the duration of anesthesia by a mean of 13 minutes. Minor intra- and postoperative complications were reported, but procedure costs were not.

3. What are your patient's personal benefits and potential risks from the therapy?

Individual reckoning of benefits and risks may be necessary in some cases. For example, with extensive pelvic adhesions from previous surgery, the magnitude of the adverse risk of laparoscopy would be higher; with prolonged duration of infertility, the magnitude of the benefit might be lower. Most often the individual reckoning will be approximate and intuitive, but sometimes an explicit calculation can be made.

> Example: A laparoscopy patient has seven years duration of infertility which might reduce the baseline expectation of

live birth from 17% to 8%. We fall back now on the relative benefit of endometriosis ablation, assuming that it will still be 1.7-fold higher, or 13.6% (1.7 times 8). The new risk difference would be (13.6 – 8) = 5.6% and the NNT would be 18 for this individual patient.

In the example intervention, the difference between NNTs of 8 and 18 would probably not change the small necessary clinical action of carrying out ablation of the endometriosis. Where the intervention entails a significant commitment of risk, time and money, however, the same difference might tip the balance of the clinical decision.

4. What alternative treatments are available?

After the clinician has found the study that addresses the clinical question, ensured that the results are valid and clinically important, and estimated that the results are relevant to clinical practice, one question remains: is there an alternate treatment that might be considered in place of the now-proven intervention under study? More importantly, among the alternate treatments that are available, are there any that are supported by evidence which is as valid or important as evidence supporting the intervention under study?

RESOLVING THE CLINICAL PROBLEM

In the study example, the decision about ablation took place during a laparoscopy that had been planned for other reasons. The more common clinical decision concerns whether to have a laparoscopy during the investigation of a couple's infertility. This clinical scenario is another illustration of the reasoning that is necessary to make published medical care research evidence relevant to clinical practice.

Example: Data from Endocan (3) suggest that one additional pregnancy will occur for every nine cases of ablation of mild endometriosis diagnosed at laparoscopy (95% CI 5, 33). If the incidence of endometriosis among patients undergoing laparoscopy for subfertility in your practice is 20%, the number of laparoscopies required to generate a single pregnancy is 8/0.2 = 40 (95% CI 25, 165).

When data from the other published RCT relevant to this clinical question are included, the absolute risk difference decreases to 0.27 – 0.18 = 0.09 (8). The number needed to treat increases to 1/0.09 = 12 (95% CI 6, 122). In a practice where the incidence of endometriosis is 20%, the number of laparoscopies required for an additional pregnancy rises to 12/0.2 = 60 (95% CI 30, 610).

Thus, while the treatment effect is statistically significant, it is clinically small and therefore not a compelling factor in the decision for or against laparoscopy during the investigation of subfertility.

SUMMARY

- Appropriate interpretation of study results involves the use of three filters:
  1. Appraise the validity of the study.
  2. Assess the clinical usefulness to your patients.
  3. Make a judgment about the clinical relevance of the results to your patients.
- If the methods of a study are not valid, it may be wise to move on to another report without wasting valuable time assessing importance or relevance.
- Key elements of validity include the security of the randomization process, completeness of follow-up, and an intention to treat analysis.
- The clinical importance is best evaluated on the basis of the absolute treatment effects: the risk difference and the number needed to treat.
- If the results are relevant to your practice, then cost and potential adverse effects are key issues when patients are making treatment choices.

Acknowledgments: This report was developed under the direction of the Practice Committee of the American Society for Reproductive Medicine as a service to its members and other practicing clinicians. While this document reflects appropriate management of a problem encountered in the practice of reproductive medicine, it is not intended to be the only approved standard of practice or to dictate an exclusive course of treatment. Other plans of management may be appropriate, taking into account the needs of the individual patient, available resources, and institutional or clinical practice limitations. The Practice Committee and the Board of Directors of the American Society for Reproductive Medicine approved this report in November 2003.

REFERENCES

1. Sackett DL, Richardson WS, Rosenberg W, Haynes RB. Evidence-based practice: how to practice & teach EBM. London: Churchill Livingstone, 1997.
2. Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use an article about therapy or prevention. A. Are the results of the study valid? Evidence-Based Medicine Working Group. JAMA 1993;270:2598–601.
3. Marcoux S, Maheux R, Berube S. Laparoscopic surgery in infertile women with minimal or mild endometriosis. Canadian Collaborative Group on Endometriosis. N Engl J Med 1997;337:217–22.
4. Hughes EG, Fedorkow DM, Collins JA. A quantitative overview of controlled trials in endometriosis-associated infertility. Fertil Steril 1993; 59:963–70.
5. Schulz KF, Chalmers I, Grimes DA, Altman DG. Assessing the quality of randomization from reports of controlled trials published in obstetrics and gynecology journals. JAMA 1994;272:125–8.
6. UK National Case-Control Study Group. Oral contraceptive use and breast cancer risk in young women. Lancet 1989;1:973–82.
7. Lehr R. Sixteen S-squared over D-squared: a relation for crude sample size estimates. Statist Med 1992;11:1099–102.
8. Gruppo Italiano per lo Studio dell'Endometriosi. Ablation of lesions or no treatment in minimal-mild endometriosis in infertile women: a randomized trial. Hum Reprod 1999;14:1332–4.