Interpretation of clinical trial results: a committee opinion

Practice Committee of the American Society for Reproductive Medicine

American Society for Reproductive Medicine, Birmingham, Alabama

This document provides guidance, background, and tips on how to recognize quality trials and focuses on evaluating the validity, importance, and relevance of clinical trial results. This document replaces the document of the same name, last published in 2008 (Fertil Steril® 2008;90:S114–20). (Fertil Steril® 2020;113:295–304. ©2019 by American Society for Reproductive Medicine.)

Discuss: You can discuss this article with its authors and other readers at **https://www.fertstertdialog.com/users/16110-fertility-and-sterility/posts/56006-29283**

Evidence from clinical trials is fundamental to ethical medical practice. Along with patient preferences, circumstances, and clinical experience, evidence 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)? Third, could the results apply to individual patients encountered in daily practice (relevance)? This document provides background and tips on how to recognize trials of quality and focus 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. The CONSORT (Consolidated Standards of Reporting Trials) guidelines were initially developed in the mid-1990s, and refined in 2010, to provide guidance for authors in an effort to improve the reporting of study results (1). Adherence to the CONSORT checklist provides authors with a comprehensive framework to improve the clarity and transparency of reporting study methodology, results, and conclusions (Table 2). The checklist can also serve as a guide for readers to assess the quality of reporting. Guidelines for efficient study interpretation have been published elsewhere (2, 3).

In this summary, the elements of critical appraisal have been organized to first address study validity, then

Received November 7, 2019; accepted November 7, 2019.
Correspondence: Practice Committee, American Society for Reproductive Medicine, 1209 Montgomery Highway, Birmingham, Alabama 35216 (E-mail: asrm@asrm.org).

Fertility and Sterility® Vol. 113, No. 2, February 2020 0015-0282/$36.00

https://doi.org/10.1016/j.fertnstert.2019.11.009

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 PCOS, is clomiphene citrate or letrozole more effective in achieving live birth? (4) 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:* Polycystic ovary syndrome is one of the most common causes of female infertility and affects 5%-10% of reproductive aged women. Clomiphene citrate has been used for decades as first line ovulation induction therapy. However, limitations of therapy include poor efficacy, high multiple pregnancy rate, and undesirable side effect profile. Previous studies of treatment have been limited by insufficient power and usage of surrogate endpoints including ovulation or hormone levels. This study sought to compare the safety and efficacy of clomiphene citrate compared to letrozole in achieving live birth, the most meaningful outcome in infertility studies, in women with PCOS (5).

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, maldistribution 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:* A total of 750 patients with polycystic ovary syndrome were randomized to treatment. A total of 158 women dropped out or were excluded from further analysis; 85/376 (22.6%) in clomiphene group and 73/374 (19.5%) in letrozole group, $P = .30$. This suggests

TABLE 1

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

Filter	Questions
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. Interpretation of clinical trials results. Fertil Steril 2019.

TABLE 2

CONSORT 2010 checklist of information to include when reporting a randomized trial.

Section/topic	Item no.	Item checklist
Title and abstract	1a	Identification as a randomized trial in the title
	1b	Structured summary of trial design, methods, results, and conclusions (for specific guidance see CONSORT for abstracts [21,31])
Introduction		
Background and objectives	2a	Scientific background and explanation of rationale
	2b	Specific objectives or hypotheses
Methods		
Trial design	3a	Description of trial design (such as parallel, factorial) including allocation ratio
	3b	Important changes to methods after trial commencement (such as eligibility criteria), with reasons
Participants	4a	Eligibility criteria for participants
	4b	Settings and locations where the data were collected
Interventions	5	The interventions for each group with sufficient details to allow replication, including how and when they were actually administered
Outcomes	6a	Completely defined pre-specified primary and secondary outcome measures, including how and when they were assessed
	6b	Any changes to trial outcomes after the trial commenced, with reasons
Sample size	7a	How sample size was determined
	7b	When applicable, explanation of any interim analyses and stopping guidelines
Randomization		
Sequence generation	8a	Method used to generate the random allocation sequence
	8b	Type of randomization; details of any restriction (such as blocking and block size)
Allocation concealment mechanism	9	Mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken to conceal the sequence until interventions were assigned
Implementation	10	Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions
Blinding	11a	If done, who was blinded after assignment to interventions (for example, participants, care providers, those assessing outcomes) and how
	11b	If relevant, description of the similarity of interventions
Statistical methods	12a	Statistical methods used to compare groups for primary and secondary outcomes
	12b	Methods for additional analyses, such as subgroup analyses and adjusted analyses
Results		
Participant flow (a diagram is strongly recommended)	13a	For each group, the numbers of participants who were randomly assigned, received intended treatment, and were analyzed for the primary outcome
	13b	For each group, losses and exclusions after randomization, together with reasons
Recruitment	14a	Dates defining the periods of recruitment and follow-up
	14b	Why the trial ended or was stopped
Baseline data	15	A table showing baseline demographic and clinical characteristics for each group
Numbers analyzed	16	For each group, number of participants (denominator) included in each analysis and whether the analysis was by original assigned groups
Outcomes and estimation	17a	For each primary and secondary outcome, results for each group, and the estimated effect size and its precision (such as 95% confidence interval)
	17b	For binary outcomes, presentation of both absolute and relative effect sizes is recommended
Ancillary analyses	18	Results of any other analyses performed, including subgroup analyses and adjusted analyses, distinguishing pre-specified from exploratory
Harms	19	All important harms or unintended effects in each group (for specific guidance see CONSORT for harms [28])
Discussion		
Limitations	20	Trial limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses
Generalizability	21	Generalizability (external validity, applicability) of the trial findings
Interpretation	22	Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence
Other information		
Registration	23	Registration number and name of trial registry
Protocol	24	Where the full trial protocol can be accessed, if available
Funding	25	Sources of funding and other support (such as supply of drugs), role of funders

ASRM. Interpretation of clinical trials results. Fertil Steril 2019.

> that the results of the study were not biased by differences in withdrawal between treatment groups (4).

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 (6).

> *Example:* Subjects were randomized using a 1:1 treatment ratio using stratified randomization with permuted blocking via web-based secured randomization service (4).

Were subjects and assessors blinded to intervention and was a placebo used? Where decisions about treatment are made by caregivers 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. For this reason, blinding and the use of placebo are both positive features of a trial.

> *Example:* The study was a double-blinded, multicenter randomized trial. The primary outcome was live birth during the treatment period, defined as delivery of any viable infant (4). Live birth is the most relevant and meaningful primary outcome in an infertility trial and previous randomized studies of letrozole were limited due to small sample size and inconsistent study design.

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:* A total of 750 patients with polycystic ovary syndrome were randomized to treatment. Study participants were followed for up to five treatment cycles and were followed with visits to determine ovulation and pregnancy and this was followed by tracking of pregnancy outcomes. A total of 158 women dropped out or were excluded from further analysis; 85/376 (22.6%) in the clomiphene group and 73/374 (19.5%) in the letrozole group, $P=.30$. The authors acknowledge that drop-out rate was higher than expected in this study but the rates of drop out were similar in each group (4). This suggests that the results of the study were not biased by differences in withdrawal between treatment groups.

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 are 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:* A total of 750 patients were randomly assigned to clomiphene citrate or letrozole in 1:1 permuted block of two, four, or six for up to 5 treatment cycles. The last enrolled patient finished study medication in July 2012 and the last birth was reported in February 2013. There were no significant differences in drop out or exclusion rate and no significant differences in reason for withdrawal. Patients were included in analyses, as assigned. No crossovers were reported (Figure 1) (4).

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:* Live-birth rate was the primary outcome assessed. This is the most relevant and meaningful outcome in an infertility study and was a major strength of the study design.

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 nonoccurrence of an event such as live birth or disease, four simple expressions are frequently used:

- Relative risk (RR)—the ratio of the probability of success with experimental treatment over the probability with the control treatment;
- Risk difference (RD)—the absolute difference between the probability of success with experimental treatment and the probability of success with the control treatment;
- Number needed to treat (NNT)—the number of subjects that must be treated to achieve one more outcome with intervention than control;
- Odds ratio (OR)—the ratio of the odds of success with experimental treatment over the odds with the control treatment. It is a measure of the probability of success over the probability of failure.

For an event that occurs in 6 of 10 individuals; the rate or probability is 6/10; the odds, however, are 6/4 (p/1-p). 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 with retrospective case-control studies because the odds ratio in case-control studies approximate the risk ratio. However, in prospective studies, for the odds ratio to approximate the risk ratio, the rare disease assumption must be met in which the outcome of interest occurs in less than 10% of the study population. The treatment effect presented depends on study question, study design, and the findings the authors are trying to emphasize.

> *Example:* In the PPCOS II trial, the authors presented the ratio of the cumulative incidence of live birth as the primary outcome measure. Patients were followed for up to 5 treatment cycles to allow sufficient time to achieve live birth. As shown in Figure 2, the group of women who received letrozole had more live births than the group of women who received clomiphene (103/374; 27.5% vs. 72/376; 19.1%, P=.007). The number of live births in patients who took letrozole was 103 out of 374 patients (0.28) and following clomiphene was 72 out of 306 patients (0.19). The risk ratio is the ratio of these cumulative incidence rates per person. Women who took letrozole were found to have 1.44 times the rate of live birth as women who took clomiphene over 5 treatment cycles (4).

Using the PPCOS II trial data, we can calculate relative risk and odds ratios.

2 × 2 tables provide a template for calculating relative risk and odds ratios.

	Exposed (letrozole)	Control (clomiphene)	Total, n
Live birth	103 (A)	72 (B)	175
No live birth	271 (C)	304 (D)	575
Total	374	376	

$$RR = (A / A + C) / (B / B + D)$$

$$(103 / 103 + 271) / (72 / 72 + 304) = 1.44$$

$$OR = AD/BC$$

$$(103 * 304) / (72 * 271) = 1.60$$

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 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 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.

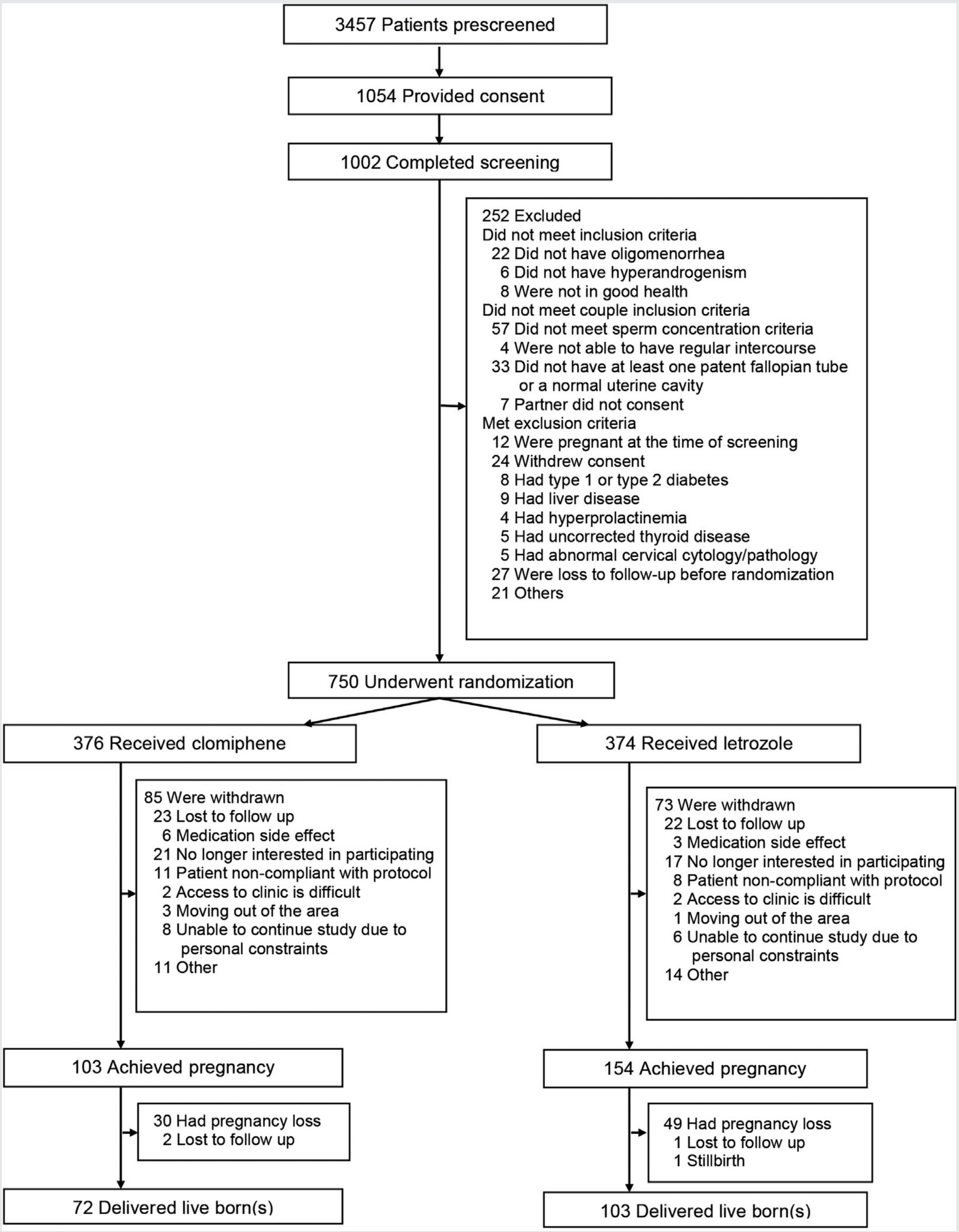
RD = 27.5-19.1 = 8.4% (letrozole live-birth rate –
clomiphene live-birth rate)

$$NNT = 1 / 0.084 = 11.9\ (n = 12)$$

> Example: The absolute effect of treatment must be calculated: the difference in live-birth rate between groups shows in the calculation above to be 8.4% (Table 2). In order to express this figure as a whole number, the reciprocal of 0.084 can be used to give a number needed to treat as shown above. Rounding upward, approximately 12 women must be treated with letrozole to achieve one additional live birth (4).

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%

FIGURE 1


3457 Patients prescreened
1054 Provided consent
1002 Completed screening
252 Excluded
Did not meet inclusion criteria
22 Did not have oligomenorrhea
6 Did not have hyperandrogenism
8 Were not in good health
Did not meet couple inclusion criteria
57 Did not meet sperm concentration criteria
4 Were not able to have regular intercourse
33 Did not have at least one patent fallopian tube or a normal uterine cavity
7 Partner did not consent
Met exclusion criteria
12 Were pregnant at the time of screening
24 Withdrew consent
8 Had type 1 or type 2 diabetes
9 Had liver disease
4 Had hyperprolactinemia
5 Had uncorrected thyroid disease
5 Had abnormal cervical cytology/pathology
27 Were loss to follow-up before randomization
21 Others
750 Underwent randomization
376 Received clomiphene
85 Were withdrawn
23 Lost to follow up
6 Medication side effect
21 No longer interested in participating
11 Patient non-compliant with protocol
2 Access to clinic is difficult
3 Moving out of the area
8 Unable to continue study due to personal constraints
11 Other
103 Achieved pregnancy
30 Had pregnancy loss
2 Lost to follow up
72 Delivered live born(s)
374 Received letrozole
73 Were withdrawn
22 Lost to follow up
3 Medication side effect
17 No longer interested in participating
8 Patient non-compliant with protocol
2 Access to clinic is difficult
1 Moving out of the area
6 Unable to continue study due to personal constraints
14 Other
154 Achieved pregnancy
49 Had pregnancy loss
1 Lost to follow up
1 Stillbirth
103 Delivered live born(s)

Enrollment and outcomes of the trial (4). Reprinted with permission.

ASRM. Interpretation of clinical trials results. Fertil Steril 2019.

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).

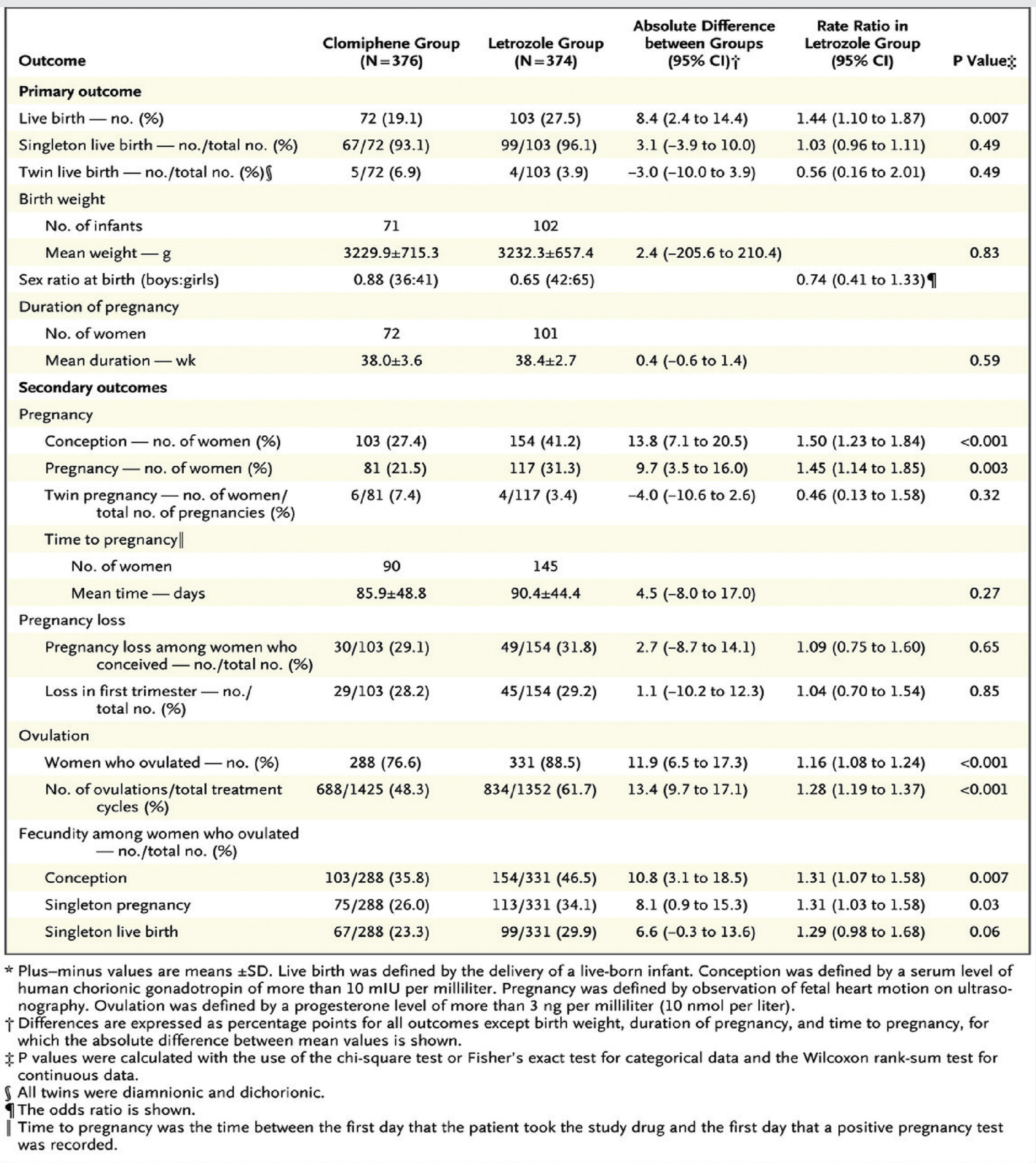
This example highlights the importance for clinicians of focusing on absolute rather than relative effects, in reading study reports and talking to patients.

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: The absolute difference in live-birth rate between groups was 8.4% (95% CI 2.4, 14.4) The

FIGURE 2



Outcome	Clomiphene Group (N=376)	Letrozole Group (N=374)	Absolute Difference between Groups (95% CI)†	Rate Ratio in Letrozole Group (95% CI)	P Value‡
Primary outcome					
Live birth — no. (%)	72 (19.1)	103 (27.5)	8.4 (2.4 to 14.4)	1.44 (1.10 to 1.87)	0.007
Singleton live birth — no./total no. (%)	67/72 (93.1)	99/103 (96.1)	3.1 (−3.9 to 10.0)	1.03 (0.96 to 1.11)	0.49
Twin live birth — no./total no. (%)§	5/72 (6.9)	4/103 (3.9)	−3.0 (−10.0 to 3.9)	0.56 (0.16 to 2.01)	0.49
Birth weight					
No. of infants	71	102			
Mean weight — g	3229.9±715.3	3232.3±657.4	2.4 (−205.6 to 210.4)		0.83
Sex ratio at birth (boys:girls)	0.88 (36:41)	0.65 (42:65)		0.74 (0.41 to 1.33)¶	
Duration of pregnancy					
No. of women	72	101			
Mean duration — wk	38.0±3.6	38.4±2.7	0.4 (−0.6 to 1.4)		0.59
Secondary outcomes					
Pregnancy					
Conception — no. of women (%)	103 (27.4)	154 (41.2)	13.8 (7.1 to 20.5)	1.50 (1.23 to 1.84)	<0.001
Pregnancy — no. of women (%)	81 (21.5)	117 (31.3)	9.7 (3.5 to 16.0)	1.45 (1.14 to 1.85)	0.003
Twin pregnancy — no. of women/ total no. of pregnancies (%)	6/81 (7.4)	4/117 (3.4)	−4.0 (−10.6 to 2.6)	0.46 (0.13 to 1.58)	0.32
Time to pregnancy‖					
No. of women	90	145			
Mean time — days	85.9±48.8	90.4±44.4	4.5 (−8.0 to 17.0)		0.27
Pregnancy loss					
Pregnancy loss among women who conceived — no./total no. (%)	30/103 (29.1)	49/154 (31.8)	2.7 (−8.7 to 14.1)	1.09 (0.75 to 1.60)	0.65
Loss in first trimester — no./ total no. (%)	29/103 (28.2)	45/154 (29.2)	1.1 (−10.2 to 12.3)	1.04 (0.70 to 1.54)	0.85
Ovulation					
Women who ovulated — no. (%)	288 (76.6)	331 (88.5)	11.9 (6.5 to 17.3)	1.16 (1.08 to 1.24)	<0.001
No. of ovulations/total treatment cycles (%)	688/1425 (48.3)	834/1352 (61.7)	13.4 (9.7 to 17.1)	1.28 (1.19 to 1.37)	<0.001
Fecundity among women who ovulated — no./total no. (%)					
Conception	103/288 (35.8)	154/331 (46.5)	10.8 (3.1 to 18.5)	1.31 (1.07 to 1.58)	0.007
Singleton pregnancy	75/288 (26.0)	113/331 (34.1)	8.1 (0.9 to 15.3)	1.31 (1.03 to 1.58)	0.03
Singleton live birth	67/288 (23.3)	99/331 (29.9)	6.6 (−0.3 to 13.6)	1.29 (0.98 to 1.68)	0.06

* Plus–minus values are means ±SD. Live birth was defined by the delivery of a live-born infant. Conception was defined by a serum level of human chorionic gonadotropin of more than 10 mIU per milliliter. Pregnancy was defined by observation of fetal heart motion on ultrasonography. Ovulation was defined by a progesterone level of more than 3 ng per milliliter (10 nmol per liter).

† Differences are expressed as percentage points for all outcomes except birth weight, duration of pregnancy, and time to pregnancy, for which the absolute difference between mean values is shown.

‡ P values were calculated with the use of the chi-square test or Fisher's exact test for categorical data and the Wilcoxon rank-sum test for continuous data.

§ All twins were diamnionic and dichorionic.

¶ The odds ratio is shown.

‖ Time to pregnancy was the time between the first day that the patient took the study drug and the first day that a positive pregnancy test was recorded.

Outcomes with regard to live birth, ovulation, pregnancy , pregnancy loss, and fecundity. Reprinted with permission (4).

ASRM. Interpretation of clinical trials results. Fertil Steril 2019.

live-birth rate was 27.5% in women who took letrozole versus 19.1% in women who took clomiphene (4).

What did the investigators consider clinically important? If a trial is large enough, it may demonstrate statistically significant differences between intervention and control groups 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 infertility patients with PCOS, a live-birth rate of 27.5% as compared to 19.1% is clinically meaningful and an increase of 8.4% in live birth would be clinically important to patients (4).

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, a) 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 letrozole and clomiphene groups are provided and interpreted.

Example: The live-birth ratio in patients who took letrozole as compared to clomiphene was 1.44 (95% confidence interval 1.10, 1.87). Thus, the chance that the study would detect a difference of <1.10 or >1.87 is less than 5%. Another way of stating this is that there is a 95% chance that the true effect size lies between 1.10 and 1.87 (4).

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 (8).

Did the study have adequate power? The probability that by chance, a study will fail to detect a real, statistically significant difference (b), 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 a 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 18-40 years with polycystic ovary syndrome defined using modified Rotterdam criteria, not taking confounding medications, had at least one patent fallopian tube and a normal uterine cavity, a male partner with sperm concentration of at least 14 million per mL and a commitment to have regular intercourse during the study with intent of pregnancy (4).

Those outside these boundaries may respond to treatment in different ways. One evidence-based medicine book (Strauss 2005) 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? When analyzing a study and the conclusions, you must assess if the study population is generalizable to the population of patients you see in your practice to determine if the intervention would likely have the same effect. In this case, the participants included were reflective of a typical patient population with polycystic ovary syndrome and, as such, similar treatment effect would be expected.

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:* Both drugs used in the PPCOS II study are oral medications readily available for prescribers. They are both designated as pregnancy category X by the FDA although clomiphene is approved for ovulation induction. Letrozole does not have FDA approval for ovulation induction but is commonly used as off-label indication. The authors did not provide any information regarding differences in cost between drugs (5).

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. Most often, the individual reckoning will be approximate and intuitive, but sometimes an explicit calculation can be made.

> *Example:* The live-birth rate was higher with letrozole than with clomiphene and the rate of pregnancy loss, duration of pregnancy, birthweight and neonatal complications did not differ between groups. The twin pregnancy rate was lower in letrozole (3.9%) as compared to clomiphene (6.9%) although the authors acknowledge the study was underpowered to detect a between-group difference. There were four major congenital anomalies in the letrozole group and one in the clomiphene group (4). These findings should be discussed with patients to guide treatment decisions.

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?

> *Example:* Ovulation induction is the most effective treatment for infertile women with PCOS to achieve conception. Alternative strategies including metabolic treatments such as metformin have been investigated without evidence of benefit (9). Further studies are needed to evaluate if a subset of patients may derive greater benefit or if other metabolic agents show more promise. Lifestyle modification including weight loss has also been evaluated with evidence of an increase in unassisted conception as well as conception following clomiphene (10, 11).

SUMMARY

- Appropriate interpretation of study results involves the use of three filters:
  - I. Appraise the validity of the study.
  - II. Assess the clinical usefulness to your patients.
  - III. 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 based on 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. Although 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, considering 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 have approved this report.

This document was reviewed by ASRM members and their input was considered in the preparation of the final document. The following members of the ASRM Practice Committee participated in the development of this document. All Committee members disclosed commercial and financial relationships with manufacturers or distributors of goods or services used to treat patients. Members of the Committee who were found to have conflicts of interest based on the relationships disclosed did not participate in the discussion or development of this document.

Alan Penzias, M.D.; Kristin Bendikson, M.D.; Samantha Butts, M.D., M.S.C.E.; Tommaso Falcone, M.D.; Susan Gitlin,

Ph.D.; Clarisa Gracia, M.D., M.S.C.E; Karl Hansen, M.D., Ph.D.; Micah Hill, D.O.; William Hurd, M.D., M.P.H.; Sangita Jindal, Ph.D.; Suleena Kalra, M.D., M.S.C.E.; Jennifer Mersereau, M.D.; Randall Odem, M.D.; Robert Rebar, M.D.; Richard Reindollar, M.D.; Mitchell Rosen, M.D.; Jay Sandlow, M.D.; Peter Schlegel, M.D.; Anne Steiner, M.D., M.P.H.; Cigdem Tanrikut, M.D.; and Dale Stovall, M.D.

REFERENCES

1. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. Obstet Gynecol 2010;115:1063–70.
2. Strauss SE, Richardson WS, Glasziou P, Haynes RB. Evidence-based medicine: how to practice and teach EBM. third edition. Edinburgh: Churchill Livingstone; 2005.
3. Guyatt GH, Sackett DL, Cook DJ. Evidence-Based Medicine Working Group. Users' guides to the medical litera-ture. II. How to use an article about therapy or prevention. Are the results of the study valid? JAMA 1993;270:2598–601.
4. Legro RS, Brzyski RG, Diamond MP, et al. Letrozole versus clomiphene for infertility in the polycystic ovary syndrome. NEJM 2014;371:119–29.
5. Legro RS, Kunselman AR, Bryzski RG, Casson PR, Diamond MP, Schlaff WD, et al. The Pregnancy in Polycystic Ovary Syndrome II (PPCOS II) trial: rationale and design of a double-blind randomized trial of clomiphene citrate and letrozole for the treatment of infertility in women with polycystic ovary syndrome. Contemp Clin Trials 2012;33:470–81.
6. 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.
7. Lehr R. Sixteen S-squared over D-squared: a relation for crude sample size estimates. Statist Med 1992;11:1099–102.
8. UK National Case-Control Study Group. Oral contraceptive use and breast cancer risk in young women. Lancet 1989;1:973–82.
9. Legro RS, Barnhart HX, Schlaff WD, Carr BR, Diamond MP, Carson SA, et al. Clomiphene, metformin, or both for infertility in the polycystic ovary syndrome. NEJM 2007;356:551–66.
10. Clark AM, Ledger W, Galletly C, Tomlinson L, Blaney F, Wang X, et al. Weight loss results in significant improvement in pregnancy and ovulation rates in anovulatory obese women. Hum Reprod 1995;10:2705–12.
11. Legro RS, Dodson WC, Kris-Etherton PM, Kunselman AR, Stetter CM, Williams NI, et al. Randomized controlled trial of preconception interventions in infertile women with polycystic ovary syndrome. JCEM 2015;100:4048–58.