Large trials or meta-analysis? That is not the question

José Villar* MD
Gilda Piaggio PhD

UNDP/UNFPA/WHO/World Bank, Special Programme of Research, Development and Research Training in Human Reproduction, Department of Reproductive Health and Research, World Health Organization, Geneva, Switzerland

Selection and evaluation of evidence to identify the most effective treatment modalities is a difficult process. Randomized controlled trials are well accepted as the least biased means of evaluating medical, surgical, screening or preventive manoeuvres. The most commonly used statistical strategy to pool results from trials identified during a systematic review is the meta-analysis. Heterogeneity of trial results should be evaluated, interactions and specific populations identified for the planning of the new large trial.

Key words: meta-analysis; large trials; agreement between sources of evidence.

Selection and evaluation of evidence to identify the most effective treatment modalities is a difficult process, which is the first step for practitioners of evidence-based medicine. The tools and techniques available for practising evidence-based medicine have been discussed in detail elsewhere in this issue. In short, randomized controlled trials are well accepted as the least biased means of evaluating medical, surgical, screening or preventive manoeuvres. The best strategy to summarize such available evidence is to conduct a systematic review based on a formal protocol. The most commonly used statistical strategy to pool results from trials identified during a systematic review is the meta-analysis.

It is in the context of this process that the relative value of the results from a large randomized trial compared with those from a meta-analysis, usually of small trials, is presented. Which is the best estimate of treatment effect? Is a new trial needed considering the results of the meta-analysis?

These two estimates are complementary strategies available to help practitioners to make the best evidence-based clinical decisions according to the condition of their patients.

* Corresponding author.
E-mail address: villarj@who.ch (J. Villar).
EVALUATION OF AVAILABLE EVIDENCE

When confronted with the question of which is the best treatment or preventive strategy for a given condition, one should evaluate all the available, least biased data in the context of a systematic review. Such systematic reviews are formal projects that are not always feasible for individual practitioners.

Systematic reviews involve the collection of evidence on a clearly formulated question using systematic and explicit methods to identify, select and critically appraise relevant primary research, and the extraction and analysis of data from studies included in the review.\(^5\) The findings of such systematic reviews might be no trials at all, a series of small trials or a combination of small trials with one or, in some instances, two or three large trials. Heterogeneity of results among these trials is often present.\(^6\) Statistical methods for pooling results, such as meta-analysis, may or may not be used. Meta-analysis is a quantitative approach for systematically assessing the results of previous research in order to arrive at conclusions about the body of research\(^7\) in question, e.g. a treatment effect.

If there are no trials, or only a few, and the health question is very relevant, a randomized controlled trial is considered. We stress here the relevance of the health or medical problem; many reviews could be needed, but it is crucial to set priorities due to research resources allocation. In modern medicine, such new trials tend to be large because of the nature of the question, the relatively low rates of diseases and the expected moderate effect of emerging new treatments relative to existing ones. It is very important that these systematic reviews are conducted by people with good knowledge and experience of the relevant medical topic. That is the only way in which the possible clinical factors affecting results, subgroup analysis and their interpretation can be identified and scientifically evaluated. These are also the people who need to be motivated in order to update the reviews when new evidence becomes available.

HETEROGENEITY OF TRIALS' RESULTS

Researchers conducting systematic reviews and meta-analyses are confronted by heterogeneous trial results in many reviews. Statistical heterogeneity is reported to be present in as many as 25% of all randomized trials, using a very conservative criteria of \(P < 0.10\), from a test of heterogeneity in trials evaluating obstetrics and perinatology treatments, and in as many as 36% in other disciplines.\(^5\) To most practitioners, heterogeneity is something to be expected, even desired.

The evaluation of heterogeneity should be conducted formally in all systematic reviews including detailed tabulation and graphical displays. Variables that can account for this heterogeneity and which are candidates for stratified analysis should always be identified in the protocols of systematic reviews. This is expected to reduce difficulties in interpreting analyses based on ‘post hoc’ stratification, and can only be done by experts on the topic addressed in the trial.

Forcing the results of a systematic review with heterogeneity into a single summary estimator is not recommended. Such pooling will be analogous to pooling results from several strata in a randomized trial if effect modification of a baseline variable is detected.

In summary, issues of heterogeneity are closely tied to the study designs and the populations in the trials included (both settings and eligibility criteria).
The generalizability of the trials’ results and the clinical relevance of observed discrepancies have to be taken into account with respect to pooling. Possible interactions and identification of specific populations in whom the treatment may be more effective will need to be assessed in a new large trial.

WHY DO WE NEED LARGE TRIALS?

When implemented correctly, randomization is the best way to prevent bias introduced by subjective allocation of participants to the treatment modalities being evaluated (selection bias), and to control for any known or unknown factor that may influence outcomes other than the treatment under study. In modern obstetrics, even in poor populations, the best that can realistically be expected for most new treatments or modified regimens is a moderate effect on severe morbidity outcomes and mortality. Therefore, we need to design studies that can discriminate reliably between treatments with moderate differences in effects but which still have important clinical or programmatic value. These studies must guarantee protection against selection bias by proper randomization, and provide sufficient statistical power. This often requires large numbers of participants to obtain a sufficient number of cases or events. Large trials or trials of adequate power to detect clinically relevant differences satisfy these criteria.

If the treatment effect is very large, there may be no need for randomized trials at all. For example, trials were not necessary when penicillin was first introduced (although trials were necessary for streptomycin). However, for most, less dramatic, effective interventions aimed at most of today’s priority diseases, it is unrealistic to expect such large effects on mortality or severe morbidity.

Some treatments do have large effects on other less substantive outcomes or on intermediate mechanisms or markers of the pathophysiology of the condition: antihypertensive drugs readily lower blood pressure during pregnancy, fetal growth can be monitored by ultrasound, blood glucose can be controlled by diet or treatments, and tumour growth can be controlled temporarily by radiotherapy or chemotherapy. Although the effects on these intermediate outcomes are large, effects on mortality or severe morbidity are usually more modest. For example, mild-to-moderate pregnancy-induced hypertension could be treated with antihypertensive drugs. There is, of course, good evidence that these drugs reduce blood pressure, but evidence from a recent systematic review failed to confirm their beneficial effects on pre-eclampsia, preterm birth or perinatal death, and some drugs even increase intra-uterine growth restriction.

Reducing the potential for bias is equally important when evaluating complex healthcare interventions. For this, a large number of randomized patients is the key issue. For example, a large body of observational evidence, collected during the 1970s and 1980s, showed a strong inverse relationship between the number of antenatal care visits and the risk of having a low-birthweight baby or a perinatal death. Subsequent large randomized trials conducted in both developed and developing countries showed that the risk of low birth weight and perinatal mortality is similar for antenatal care programmes with fewer visits but focused on effective interventions compared with programmes with the traditional number of visits.

To recruit large numbers of people over a reasonable period of time at an affordable cost, large trials need to be simple. Complexity is a barrier to recruitment, interferes with clinical practice, encourages participants to leave the study early, and restricts
generalizability of the results. If assessment of eligibility is complex or is based on criteria not widely used in clinical practice, many patients will be excluded and the results will only apply to the relatively small group of patients recruited. For rapid recruitment of large numbers of participants, eligibility criteria and the procedure for trial entry must be simple so that they can be adapted and integrated into existing clinical practice.

For example, screening for inclusion in the World Health Organization (WHO) Misoprostol Third Stage of Labour Trial\textsuperscript{11}, which recruited over 18,000 women over less than 3 years, included only four questions and there was no test or complex clinical examination. The WHO Calcium Supplementation During Pregnancy Trial, just completed, screened close to 15,000 women within 2 years and recruited 8300 of them to the trial.\textsuperscript{12} The Collaborative Eclampsia Trial\textsuperscript{13} recruited 1687 women with eclampsia; this is by far the largest trial on this topic. A criterion for selection of centres in this trial was the high prevalence of eclampsia, but centre recruitment was facilitated by making women's clinical care easier than in normal practice.

Furthermore, delivery of the intervention should be feasible within the existing health services. Data collection must be simple and based on information that is likely to be readily available in routine clinical records. Information should only be collected if it is specified in the protocol. It should be clinically relevant, so no effort is wasted in collecting information that is unlikely to be used. An additional advantage is that, in large trials, if randomization is conducted correctly and the treatment allocation is concealed adequately, the baseline characteristics of treatment groups will be well balanced at trial entry, so only data for important prognostic variables need to be collected. It is usually preferable to collect 10 times less data on 10 times more patients.\textsuperscript{8} Again, we conclude that large trials can provide the answer to the question in a relatively short time.

In addition, there are special situations in which large trials are needed, for example, to demonstrate equivalence between treatments. This is the case when a new treatment is claimed to have equal effectiveness to that of a standard treatment but with fewer or less severe side-effects, an easier mode of administration, or better cost-effectiveness. Demonstrating clinical equivalence within a margin often requires a larger sample size than for superiority or effectiveness trials. Results suggesting equivalence in small trials are not evidence of the absence of clinically important differences, which might turn out to be significant when sufficient numbers of participants are accumulated in large trials or meta-analyses.

A second case of a need for large trials is given by cluster randomized trials to evaluate interventions that are implemented at group level rather than individual level. There is a growing interest in the use of these cluster randomization or community trials. For most trials, the unit of randomization is the individual person being allocated to a specific intervention or to a control or placebo. Other units (clusters) of randomization, such as clinics, hospitals, physicians or families, can be used, and these are particularly attractive for evaluation of health services. Cluster randomization trials have the advantages of reducing ‘contamination’ of the interventions between groups, they can increase participation, and they allow for better administrative and logistic organization in implementing the intervention.\textsuperscript{14–16} Their disadvantage is that they usually need to enrol more subjects than a comparable trial based on individual randomization. If there is a high degree of homogeneity within each cluster, such as that observed in families or medical practices, the number of clusters needed is larger although the cluster size could be small. It is more effective to include a large number of small clusters (i.e. clinics) in the trial than a small number of very large clusters (i.e. cities) (Tables 1–3).
**Table 1.** Why do we need large simple trials?

<table>
<thead>
<tr>
<th>Reasons</th>
</tr>
</thead>
<tbody>
<tr>
<td>To reduce, as much as possible, variation due to random errors</td>
</tr>
<tr>
<td>To have the power to assess moderate effects</td>
</tr>
<tr>
<td>To have the power to assess effects on rare conditions</td>
</tr>
<tr>
<td>To have the power to assess effects in clinically important subgroups</td>
</tr>
<tr>
<td>To have the power to demonstrate clinical equivalence</td>
</tr>
<tr>
<td>To make (possible) use of a cluster randomization design</td>
</tr>
<tr>
<td>To make results applicable to a wide range of people and settings</td>
</tr>
</tbody>
</table>


**Table 2.** Some advantages of simple pragmatic trials.

<table>
<thead>
<tr>
<th>Advantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feasible to recruit large numbers of subjects</td>
</tr>
<tr>
<td>Simple eligibility criteria</td>
</tr>
<tr>
<td>Simple trial entry procedure</td>
</tr>
<tr>
<td>Conducted within the existing health services</td>
</tr>
<tr>
<td>Intervention feasible without additional staff or technology</td>
</tr>
<tr>
<td>Data collection based on information likely to be available in routine records</td>
</tr>
<tr>
<td>Considerably less expensive than more complex studies</td>
</tr>
<tr>
<td>Minimal additional work for already busy clinicians</td>
</tr>
<tr>
<td>Encourages participants to stay in the trial</td>
</tr>
<tr>
<td>More complete and better quality data</td>
</tr>
<tr>
<td>Simpler data management</td>
</tr>
<tr>
<td>Results are relevant to clinical practice in a wide range of settings</td>
</tr>
</tbody>
</table>


**Table 3.** Factors related to disagreements between meta-analyses and large trials.

<table>
<thead>
<tr>
<th>Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factors related to trials</td>
</tr>
<tr>
<td>Protocol differences</td>
</tr>
<tr>
<td>Risk of disease level in control group</td>
</tr>
<tr>
<td>Other population characteristics</td>
</tr>
<tr>
<td>Treatment compliance in small trials vs pragmatic trials</td>
</tr>
<tr>
<td>Factors related to meta-analyses</td>
</tr>
<tr>
<td>Publication bias</td>
</tr>
<tr>
<td>Evaluation of trials’ quality</td>
</tr>
<tr>
<td>Statistical analysis</td>
</tr>
<tr>
<td>Heterogeneity in trials’ results (qualitative and quantitative)</td>
</tr>
<tr>
<td>Factors related to the comparison between a meta-analysis and large trials</td>
</tr>
<tr>
<td>Parameter used (difference between RRs, log-odds ratios or similar summary statistic, kappa, positive predictive value, correlation coefficient)</td>
</tr>
<tr>
<td>Definition of largest trial as gold standard</td>
</tr>
<tr>
<td>Random vs fixed effect model in the meta-analysis</td>
</tr>
<tr>
<td>The bias of the ‘large trial never done’</td>
</tr>
</tbody>
</table>
META-ANALYSIS WITHIN SYSTEMATIC REVIEWS

A properly conducted randomized controlled trial with adequate statistical power will provide an unbiased precise estimate of the treatment effect under evaluation. The most common measures of the effect of a treatment are the relative risk (RR), the odds ratio and the risk difference.

An alternative to conducting a costly large trial is to collect all available small or underpowered trials in a systematic review, and obtain a summary measure (or pooled estimate) of the effect of treatment. There is a tendency in the literature to report this single pooled estimate using meta-analytical techniques without a critical evaluation of the validity of pooling.

It is clear, however, that when there are differences in the direction of the effect among trial results (qualitative heterogeneity) and/or if there is heterogeneity in the magnitude of the effect (quantitative heterogeneity), such a single pooled estimate may be misleading. For meta-analyses without clinically significant heterogeneity including small underpowered trials, the next step after assessment of statistical heterogeneity is to evaluate whether or not pooling of trials achieves the required power to justify recommending a change in medical practice. Here we should ask, ‘can results of a systematic review (and its meta-analysis) replace the need for a large trial, which will take time, funds and human resources?’.

The issue here is the strength of the systematic review and its data. Systematic reviews’ protocols should include explicitly presented strategies to locate all available trials (even if unpublished) and describe the planned efforts to locate them. They should also identify, a priori, the range of acceptable differences among trials’ results for the clinical evaluation of heterogeneity, and identify key effect modifiers in advance of data analysis.

CAN WE EXPECT AGREEMENT BETWEEN META-ANALYSES AND LARGE TRIALS?

Meta-analyses can provide a summary measure conceptually similar to that from a large trial. The logical question is then, how do they agree? Despite the expanding use of meta-analysis and its promotion, very little empirical information was available, by 1995, on how results from a pooled estimate from a meta-analysis are confirmed or refuted by large trials. Clinicians claim that this statistical technique was introduced and promoted with limited scientific evaluation!

Confronted with this problem, we, as authors and users of systematic reviews, conducted the first formal evaluation.17 This initial paper was followed by four other papers which presented analyses of data18–21 and a review.22 Factors affecting these comparisons and the discrepancies between estimations have also been discussed.6,22,23 A related topic, the discrepancies among megatrials, has also been examined.24

As expected, comparisons between meta-analyses and large trials’ results have demonstrated large variations according to heterogeneity of individual trial’s results (clinical and statistical), methodological quality and protocol issues of included trials, the risk of the condition in the control group, publication bias, the method used to assess agreement between the meta-analysis and the large trial, and the method used to obtain the pooled estimate in the meta-analysis (i.e. fixed vs random effect model).
The range of agreement reported in papers published to date depends on the measure of agreement used, but ranges between low-to-moderate agreement beyond chance using kappa statistics, and correlation coefficients with values between 0.50 and 0.75, or about 20% disagreement if statistical significance levels of the differences in the standardized treatment effect are used.\textsuperscript{17,19,20,22}

The issue of random vs fixed effect model is of relevance as the random effect model produces a wider confidence interval of the pooled RR of the meta-analysis.\textsuperscript{5} As such, comparing meta-analyses with large trials' results using the random effect model will reduce the discrepancies between them by the very nature of the random effect model.\textsuperscript{23}

A method to assess agreement proposed in a later paper\textsuperscript{21} is based on the difference between the two results (RR or similar summary statistics). In this method, the component of the variance is used twice, based on the standard error of the difference using the random model. Therefore, the confidence interval will be wider and the conclusion of agreement will be drawn more often.

There is a bias that overestimates the proportion of agreement between meta-analysis and large trials. Such bias has been called ‘the bias of the large trial never done’.\textsuperscript{23} This relates to the fact that only promising, positive meta-analysis results are followed by large trials, overstating the overall concordance between small (positive) and large trials.\textsuperscript{23}

In our earlier paper\textsuperscript{17}, we found that meta-analyses of small trials tended to demonstrate a stronger protective effect of the intervention under evaluation than the largest trial of the same treatment. Therefore, meta-analysis and the corresponding large trial might show important discrepancies of clinical relevance, reinforcing the complementary nature of these two research strategies.

We are presently confronted with the exact situation we have been discussing here. A systematic review has been published, and recently updated\textsuperscript{25,26}, with respect to the effect of calcium supplementation during pregnancy for the prevention of pre-eclampsia. This systematic review demonstrated overall statistical heterogeneity ($P=0.0007$) and suggested that a protective strong effect is only concentrated among women with low calcium intake (RR = 0.29, 95% CI 0.19–0.45, fixed effects model), with no evidence of heterogeneity of results in this subgroup analysis ($P=0.154$).

Based on this review, a large, double-blinded, randomized trial has been conducted by the WHO to evaluate these promising results. This trial has recruited over 8000 women and we are presently completing the follow-up of the last women enrolled. The question is whether or not this new, large trial, conducted in the population at highest risk, will agree with the results on the same stratum of the meta-analysis.

We simulated the scenario for this paper without having seen the data. For example, if the new trial demonstrates an identical effect in the calcium and placebo groups (100 pre-eclamptic women out of 4000 women in each group) (RR = 1.00, 95% CI 0.76–1.31), adding this new trial to the corresponding low-calcium-intake stratum will change the point estimate for the low-calcium-intake subgroup to a RR of 0.68 (95% CI 0.55–0.85, fixed effect model), still supporting a very protective effect. As expected, the test of heterogeneity of trial results now demonstrates statistically significant heterogeneity ($P=0.0005$). There is a clearly different clinical implication between the simulated results of the new large trial and the results of the meta-analysis, the former showing no effect of supplementation and the latter showing a strong protective effect. What should we recommend in this possible scenario? That is the question!

In this possible scenario, rather than pooling heterogeneous results using for example random effect models, factors responsible for heterogeneity should be
explored. If no satisfactory explanation is found, a recommendation to provide calcium supplementation could be made considering all evidence available from the systematic review.

CONCLUSIONS

Meta-analysis in the context of a systematic review is increasingly used in medical practice and literature to provide a summary measure of treatment effect. Forcing the results of a systematic review to produce a single summary estimate in the presence of heterogeneity should be avoided, and a stratified analysis should be conducted. The issue of heterogeneity of results of a systematic review is not without clinical implications as over one-third of meta-analyses could have statistical heterogeneity. An important application of meta-analysis is to summarize the available data to decide whether a new, definitive trial is needed and, if so, which is the target population and the most promising treatment regimen. This exercise includes identifying and understanding the sources of disparity among trials. Evidence from small trials summarized through a meta-analysis and a new large trial are complementary sources of evidence rather than opposing each other, and are part of a dynamic process of scientific development.

The stronger effect observed in small trials of preventive interventions implemented among healthy individuals may be related to better compliance with treatment in a small, tightly monitored, single-centre study compared with a large, pragmatic setting. Clinicians should not be surprised to see less impressive effects when implementing interventions outside experimental conditions.

In short, although there is, in general, moderate agreement between these two sources of scientific evidence, there are situations, particularly when heterogeneity is present, in which the pooled estimate provided by a meta-analysis might not be adequate. As for original research and randomized controlled trials, systematic reviews need to be rigorous, transparent (particularly about which studies are eligible for inclusion), up to date, published in detail and reviewed independently. With all these elements, we will be able to minimize bias and make our final judgement about policy, practice and research based on the best-available evidence.

---

**Practice points**

- large, randomized controlled trials provide the least biased source of evidence
- when there is heterogeneity in trials’ results, a pooled estimation should be avoided
- small trials of preventive interventions tend to overestimate the effect
- less impressive effects are often seen outside experimental conditions

---

**Research agenda**

- large, simple trials should be encouraged for any form of treatment modality
- more empirical evaluation of the properties, limitations and predictions of systematic reviews and their statistical strategies, particularly those with heterogeneity, is required urgently
REFERENCES


