Statistical issues in the design and analysis of expertise-based randomized clinical trials

S. D. Walter^{1,*†}, A. S. Ismaila¹, and P. J. Devereaux^{1,2} for the SPRINT Study Investigators

¹ Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ont., Canada L8N 3Z5

² Department of Medicine, McMaster University, Hamilton, Ont., Canada L8N 3Z5

SUMMARY

In order to avoid certain difficulties with the conventional randomized clinical trial design, the expertise-based design has been proposed as an alternative. In the expertise-based design, patients are randomized to clinicians (e.g. surgeons), who then treat all their patients with their preferred intervention. This design recognizes individual clinical preferences and so may reduce the rates of procedural crossovers compared with the conventional design. It may also facilitate recruitment of clinicians, because they are always allowed to deliver their therapy of choice, a feature that may also be attractive to patients.

The expertise-based design avoids the possibility of so-called differential expertise bias. If a standard treatment is generally more familiar to clinicians than a new experimental treatment, then in the conventional design, more patients randomized to the standard treatment will have an expert clinician, compared with patients randomized to the experimental treatment. If expertise affects the study outcome, then a biased comparison of the treatment groups will occur.

We examined the relative efficiency of estimating the treatment effect in the expertise-based and conventional designs. We recognize that expected patient outcomes may be better in the expertise-based design, which in turn may modify the estimated treatment effect. In particular, a larger treatment effect in the expertise-based design can sometimes offset a higher standard error arising from the confounding of clinician effects with treatments.

These concepts are illustrated with data taken from a randomized trial of two alternative surgical techniques for tibial fractures.

Keywords

randomized clinical trial; expertise; statistical efficiency; bias

*Correspondence to: S. D. Walter, Department of Clinical Epidemiology and Biostatistics, McMaster University, HSC-2C16, Hamilton, Ont., Canada L8N 3Z5.

†walter@mcmaster.ca

INTRODUCTION

In the conventional randomized controlled trial (RCT), patients are randomized to receive an intervention A or B, and participating clinicians administer those interventions based on the randomization. In other words, the same clinicians will sometimes administer intervention A and sometimes intervention B depending on the randomization sequence. While this is a well accepted design for areas such as pharmaceutical interventions, it is less acceptable in other areas of medicine. For example, in surgery, surgeons are often less willing to accept the possibility of having to use both of the two alternative types of surgical intervention under comparison. Instead, they often tend to prefer one intervention and may feel more comfortable using it, as opposed to the other intervention [1].

Another example of this kind is education-based health interventions. When specialized knowledge or skill is required to introduce a particular educational program, it may be difficult for health care providers to contemplate delivering both interventions in the context of the same study. Furthermore, even if the conventional design is attempted in this context, there is the possibility of inadvertent contamination of one intervention by the other, because it may be difficult for health care providers to clearly separate the educational components of each intervention in their minds, and thus they may be unable to administer a ‘pure’ intervention as defined by the randomized assignment for a given patient.

These difficulties have led to the introduction of the expertise-based (EB) design [2, 3]. In this design, health care providers administer only one of the treatment interventions, A or B—whichever they are most familiar and comfortable with—and then they use it for *all* their patients. The treatment assignment for a given patient is still randomized, but now each patient is allocated to a health care provider associated with the A or B group, according to the randomization. This is shown schematically in Figure 1. A variant of this design is possible, in which clinicians may declare that they have expertise in *both* A and B. In this situation, we would have a hybrid between the conventional and the EB randomized designs.

While early experience suggests that surgeons and other clinicians are often very much in favour of using the EB design, little is known about its relative merits compared with the conventional design. In this paper, we will review some of the potential advantages and disadvantages of the EB design, and then discuss its statistical efficiency relative to the conventional RCT. The ideas will be illustrated by data from a large randomized trial of tibial fractures.

METHODS

Advantages of the EB design

The EB design takes advantage of the fact that most clinicians will tend to favour one treatment over another. This may be because they have had more exposure to that treatment during or since their training, or because they believe that it is more effective. The conventional RCT design ignores these clinical preferences, and puts participating clinicians in the position of having to administer both interventions under comparison, including the one in which they have less confidence or expertise.

In the context of the EB design, because clinicians are restricted to administer only the intervention that they prefer, patient crossovers from one comparison treatment to the other are less likely. In the conventional design, clinician investment may make them inclined to deliver other beneficial aspects of care more aggressively to patients receiving their preferred intervention. One can anticipate, in the EB design, that all clinicians will be aggressive in optimal management of their patients [4–6].

Even if patients are randomized immediately before the clinical intervention—such as just before surgery—new clinical information about the patient may come to light during the surgery itself, and that may necessitate a change of plan. However, such procedural crossovers are presumably more likely if surgeons have to sometimes use a technique that is less familiar to them, and with which they have less confidence, such as occurs half the time in the conventional design. In contrast, in the EB design, surgeons are more likely to adhere to the original planned intervention, particularly if they believe that the technique they are using is in fact the better of the two interventions being compared.

The EB design has some ethical advantages over the conventional design. Patients involved in an EB study can be assured that they will receive an intervention from a clinician who has expertise in and prefers that particular procedure—something that cannot be offered to patients in the conventional RCT design. Such assurance to the patient at the stage of acquiring informed consent may increase their participation rates. Furthermore, recruitment of participating clinicians may also be easier if they can be told that they need only deliver their therapy of choice.

Because of their tendency to favour one of the comparison interventions A or B, clinicians will be likely to develop differential skill and experience with respect to those interventions. If one intervention is used more often than the other in routine clinical practice, then the possibility of so-called expertise bias exists if the conventional trial design is adopted. For example, if surgical intervention A is more commonly used, and there are more surgeons in the community who are familiar with it, then patients who are randomized to A will be more likely to have a surgeon with experience in his or her assigned operative technique, compared with patients who are assigned to technique B. This puts B at a disadvantage—leading to the so-called differential expertise bias [2, 3, 7].

As an example, in preparing to execute a randomized clinical trial of two surgical techniques for the repair of tibial fractures, we surveyed a group of surgeons who were likely to take part [1]. The survey showed that surgical technique A (which involved the use of a reamed intramedullary nail) had been used much more often than the alternative technique B (which involved an unreamed nail). Furthermore, the survey revealed that a large majority of surgeons believed that the reamed procedure was superior. These results imply that in a conventional design many surgeons would often be asked to carry out a procedure which they believed was inferior.

One possible solution to the problem of limited expertise is to require clinicians to have administered each intervention a minimum number of times before beginning their participation in the randomized trial. Unfortunately, in some situations a very large number

of procedures may be required before skill is optimized. For example, it has been shown that as many as 250 laparoscopic inguinal hernia surgeries are required before hernia recurrence rates are significantly reduced [8]. In other words, for many trials it must be accepted that clinicians are still learning even as the study progresses.

Disadvantages of EB design

There are some important statistical implications of using the EB design compared with the conventional design, particularly with respect to its efficiency. By definition, in the EB design, clinicians are nested within intervention groups, which typically will lead to a higher standard error for the estimated treatment effect than a randomized design in which clinicians are crossed with treatments, and approximately or even exactly balanced across intervention arms. In addition, the treatment effect size in the EB design may be different from that in the conventional design, because surgeons who prefer A may be more skilled than those who prefer B (or vice versa). The effects of differential clinician skill will enhance the magnitude of the treatment effect (compared with that in the conventional design) if the surgeons delivering the superior treatment are those with superior skill. On the other hand, if the more skilled surgeons are delivering the inferior treatment, the effect of clinical expertise may be greater for the inferior treatment. Depending on how expertise affects patient outcomes in a specific study, there may or may not be an enhancement of the treatment effect, which might counteract the higher standard error of the treatment effect in the ED design.

In examining the statistical efficiency of the EB and conventional trial designs, we will take account of individual clinician effects. This is clearly desirable in the context of the EB design, where the clinician effect is nested within treatment, and hence clinician effects are potentially confounded with treatment. For the sake of comparability, clinician effects should also be taken into account in the analysis of conventional designs, even though crossed clinician effects would avoid confounding. However, adjustment for clinician effects in the conventional design will potentially decrease the standard error of the treatment effect, and hence increase power in its test of statistical significance. We note in passing that incorporation of clinician effects in the analysis of conventional trials is something that has rarely been done in practice. (An exception might be certain pharmaceutical trials, where effects of treatment centres are fitted, and which may correspond to clinicians.)

In addition to the statistical issues discussed above, there are other practical issues of feasibility and interpretation of the EB design, which we will discuss later.

MODEL DEVELOPMENT

For the conventional RCT design, we suppose there are k surgeons, and for simplicity we will assume a regular and balanced allocation of patients such that each surgeon treats m patients in each arm of the study. The total number of patients in the study is thus $N = 2km$. Let Y_{ij} be the response variable for patients j treated by surgeon i ($i = 1, 2, \dots, k$) in treatment group t . Our model for the expected response is

$$Y_{tij} = \mu + \alpha_t + \beta_{ti} + \varepsilon_{tij} \quad (1)$$

Here μ is the overall mean, α_t is the effect of being on treatment t , β_{ti} is the effect associated with surgeon i in treatment group t , and ε is the error. We take the treatment effect as fixed, surgeon as a random effect with mean zero and variance σ_s^2 , and the errors as independently and distributed with mean zero and variance σ^2 . The analysis of variance (ANOVA) table for the conventional design is listed in Table I. Note again that published analyses of conventional trials typically neglect to take account of the surgeon (or, in general, clinician) effect. The observations can be thought of as having a clustered data structure [9–11].

For the EB design, we suppose there are k' surgeons in each treatment group, and that they each see m' patients. Here we have that the total number of patients in the study is $2k'm'$. The ANOVA for the EB design is listed in Table II, with clinicians being a nested factor within treatment groups. Note that because clinicians are no longer crossed with treatments, the expected mean square for treatment in the EB design now includes a component involving σ_s^2 .

We denote the effect of treatment A relative to treatment B by $\theta = \alpha_1 - \alpha_2$. For both the conventional and EB designs, the estimated treatment effect is $\hat{\theta} = \bar{Y}_{1..} - \bar{Y}_{2..}$, being the empirical difference in mean responses between the study arms. While the estimates of treatment effect are the same, their precisions are different. From first principles, we may derive that the variances of $\hat{\theta}$ are

$$\text{var}(\hat{\theta}_C) = \frac{2\sigma^2}{m'k'}$$

and

$$\text{var}(\hat{\theta}_{EB}) = \frac{2}{mk} [\sigma^2 + m\sigma_s^2]$$

for the conventional and EB designs, respectively. For comparability we will now suppose that the total number of patients is the same in the two designs, so that $N = 2k'm' = 2km$. Accordingly, the relative efficiency (RE, assessed by the ratio of standard errors of the estimated treatment effects) of the conventional vs the EB design is given by

$$\begin{aligned} RE &= \text{var}(\hat{\theta}_{EB}) / \text{var}(\hat{\theta}_C) = \frac{m'k'}{mk} [1 + m\sigma_s^2 / \sigma^2] \\ &= 1 + m\sigma_s^2 / \sigma^2 \end{aligned}$$

Thus, as long as $\sigma_s^2 \neq 0$, the EB design is less efficient than the conventional RCT, for a given total number of patients in the study. The RE of the conventional design increases with the number of patients seen per surgeon. Efficiency of the EB design can therefore be maintained by using more surgeons with as few patients each as possible, but that strategy is not necessarily very practical.

For some studies, it may be possible to characterize the level of expertise of each clinician on one or both treatments. For instance, in our motivating example of the randomized trial of tibial fractures, we were able to ask about the experience of the surgeons participating at the start of the study, in terms of the number of times they had used reamed and unreamed operative procedures during specialty training, and since completion of their training. We were then able to group those surgeons into levels of experience with either or both interventions. Although we suppose that experience with the operative procedure in use is the most pertinent for a given patient, it is possible that experience with *other* similar procedures (and the comparison procedure in particular) may also tend to improve patient outcomes. Therefore, to avoid loss of generality, we will characterize a surgeon's expertise according to his/her experience with each of the study interventions.

In model (1), $\theta = \alpha_1 - \alpha_2$ should be regarded as the *inherent* effect of treatment, exclusive of any enhanced (or diminished) effect associated with expertise. In the EB design, we anticipate potentially enhanced benefits with *both* treatments, because of the greater expertise levels of the participating clinicians, who here exclusively use techniques with which they are most familiar. In contrast, in the conventional RCT, clinicians are crossed with treatments, and hence they are using the less familiar technique for half the time. The net impact in the EB design is an anticipation of a different estimate of the treatment effect, compared with the estimate from the conventional design. If a larger treatment effect occurs in the EB design that may be sufficient to offset the design's lower efficiency as demonstrated earlier.

If estimates of the expertise effects (β) are available, we can evaluate the expected treatment effect with any given distribution of clinical expertise in the study comparison groups. Let $\{\pi_{tm}; m = 1, \dots, M\}$ represent the probability distribution of clinicians over expertise categories m in treatment groups t ; thus $\sum_m \pi_{tm} = 1$, with the summation being over all expertise levels m . The associated expected benefit B_t in treatment group t , derived from clinical expertise, is thus

$$E(B_t) = \sum_m \pi_{tm} \beta_{tm}$$

The values of B_t can be combined with the inherent treatment effect θ , to obtain an expected treatment efficacy, i.e. including the possibility of differential expertise effects, as $\theta + (B_1 - B_2)$.

EXAMPLE

The motivating example for this paper comes from a randomized trial intended to compare the effectiveness of two alternative surgical techniques for tibial fractures. Specifically, patients were randomized between reamed or unreamed intramedullary nails, and the study as a whole involved approximately 1350 patients who were operated on by approximately 140 surgeons in Canada, the US and the Netherlands. This study, one of the largest orthopaedic trials ever undertaken, will be referred to here by its acronym—SPRINT—Study to Prospectively evaluate Reamed Intramedullary Nails in Tibial fractures [12].

Before the trial began, the reamed nailing technique was used much more by most surgeons. In the year prior to participating in SPRINT, less than 10 per cent had not performed a reamed procedure, whereas the corresponding figure for unreamed nails was 35 per cent. Over 30 per cent of surgeons reported having done at least 20 reamed operations in the previous year, compared with approximately 15 per cent who had done at least 20 unreamed operations. Of the responding surgeons, 87 per cent thought that the reamed procedure was superior, and a majority of respondents were moderately or extremely confident in that opinion [1]. A further survey carried out among participating surgeons when approximately 900 patients had been randomized showed no important change in these beliefs. These differences in opinions and beliefs between reamed and unreamed nails suggest the strong possibility of differential expertise bias, because patients randomized to the reamed procedure had a much higher chance of being operated on by a surgeon who had more experience in this technique and who probably believed it to be the preferred method.

Although SPRINT was carried out with a conventional RCT design, information available on the expertise of a subset of participating surgeons allows us to project the expected outcomes if an ED design had been used instead, and thereby to estimate the RE of the two alternative designs. In order to make this comparison consistent, we will limit our analysis to patient outcomes associated with the subset of 76 surgeons where expertise with reamed and unreamed nailing procedures was known. For the purpose of this example, a surgeon will be defined as an expert in one or both techniques if the total number of operations for tibial fractures of all grades performed during and after orthopaedic training was greater than the median for the reamed technique (equal to 77).

There were 596 patients associated with the 76 surgeons whose expertise was known, corresponding to an average of 3.92 patients per surgeon in each treatment arm. We will consider the expected outcomes under two potential ED designs, selected to represent two plausible scenarios under which the EB design might have been used. In the first, to be denoted EB-I, the study is administered in such a way that the same total number of patients per surgeon ($2 \times 3.92 = 7.84$) is seen, but by definition each surgeon only uses the one technique in which he is defined to be an expert. Accordingly, in EB-I there would only be half the number of surgeons ($\frac{76}{2} = 38$) compared with the number that was actually used in the conventional design of SPRINT. In the second ED design, denoted as EB-II, we maintain the same number of surgeons (76) as was actually used in SPRINT, but where each surgeon sees only half the number of patients (3.92) compared with what actually occurred in SPRINT.

For our example, we will use the physical symptoms component of the SF-36 scale as our outcome variable [13, 14]. We first carried out an ANOVA in the spirit of Table I, recognizing that surgeons were inevitably not perfectly balanced between treatment arms. We estimated the variance components associated with surgeon and error, together with the estimated treatment effect between randomized groups. The estimated treatment effect was 0.64, the surgeon variance σ_s^2 was 3.91, and the error variance σ^2 was 90.57. While σ_s^2 is considerably smaller than the error variance, it is nevertheless quite large in comparison with the overall treatment effect, indicating that the surgeon may be an important determinant of a patient's outcome.

Using these results in the corresponding estimated variances for the treatment effect in the two competing designs, we obtained $\text{var}(\hat{\alpha}) = 0.61$ for the conventional RCT design and $\text{var}(\hat{\alpha}) = 0.81$ for the ED design EB-I. These figures correspond to a RE of 1.33 of the conventional design compared with EB-I. If the same calculation is made for EB-II, the RE of the conventional design is 1.16. As expected, the RE for EB-II is somewhat better than for EB-I because of the smaller number of patients per surgeon, thus limiting the variance inflation factor associated with the clustered allocation of patients to surgeons.

The EB design necessarily has lower statistical efficiency (in terms of the standard error of the estimated treatment effect) than the conventional design. However, this lower efficiency must be considered in comparison with a potentially different treatment effect when expertise is taken into account. Depending on the relative magnitudes of the expertise effects associated with each of the two alternative interventions, the treatment effect may be enhanced or diminished compared with the expected result from the conventional design. Accordingly, the EB design may be more or less powerful than the conventional design.

Figure 2 shows the expected results in SPRINT patients according to whether the surgeon was defined as an expert in reamed, unreamed or both types of surgery. We see that the patients whose surgeons were expert in both techniques had the best expected outcomes, followed by patients for surgeons who were expert only in the non-reamed technique, patients whose surgeons were expert in the reamed technique, with the patients whose surgeons were expert in neither technique having the poorest expected outcomes. As shown in Table III, relative to the mean outcome for patients in the last group (surgeons expert in neither technique), the estimated expertise effects were 2.83 (surgeons expert in both techniques), 1.78 (surgeons expert in reamed only), and 2.24 (surgeons expert in unreamed only).

In this example we have defined surgeon expertise in terms of experience with both alternative surgical techniques, even though only one of them is directly relevant to a particular patient. The rationale here is that surgical experience with tibial fractures in general may improve patient outcomes. As it turns out in this example, the surgeons who are expert in both techniques did indeed turn out to have the best outcomes. This suggests some 'transferability' of the expertise effect from one technique to the other.

We may now use these estimates of the expertise effects to project the expected outcome of an EB-based study to compare the same two interventions, if the EB design had been used in

place of the conventional design. For illustrative purposes, we will suppose that surgeon expertise is uniformly distributed between the four categories defined by surgeon expertise on A or B alone, both A and B, or neither. In the EB design, patients randomized to the reamed nail (intervention A) would, under these assumptions, be equally likely to have a surgeon who is expert in A only or expert in both A and B. (By definition, patients randomized to A cannot be allocated to a surgeon who is not expert in A). Accordingly, as shown in Table III, the expected change in outcomes for patients in this group will be an average of the expertise effects associated with these two groups of surgeons, i.e.

$\frac{1}{2} \times (2.83 + 1.78) = 2.31$. Correspondingly, patients randomized to B in the EB design will have equal chances of having a surgeon expert in B only or both A and B. This leads to an adjustment in the expected outcome for this group of $\frac{1}{2} \times (2.83 + 2.24) = 2.54$.

The net effect of these expertise adjustments between the two randomized intervention groups is $2.31 - 2.54 = -0.23$. When added to the original treatment effect of 0.64 seen in the conventional design, we would therefore anticipate a treatment effect in the EB design of $0.64 - 0.23 = 0.41$. Hence, in this example the expected treatment effect in the EB design is actually smaller than the corresponding effect in the conventional design. This is because the expertise effect associated with surgeons who are expert in B only (2.24) is larger than the corresponding expertise effect for surgeons who are expert in A only (1.78), while at the same time the conventional treatment effect estimate favoured intervention A. The expertise advantage associated with surgeons who are expert in both techniques does not contribute to the difference in expected outcomes between randomized intervention groups, because such surgeons are equally likely to be used in either group.

We may now integrate the findings on the anticipated treatment effects and their associated standard errors in order to make an overall comparison of the ED designs with the conventional design. The main results are summarized in Table IV. As noted earlier, the anticipated treatment effect will be smaller in the EB design for this example compared with the conventional design (0.41 vs 0.64). Additionally, the standard error in the EB design is larger than in the conventional design, particularly for EB-I, where the number of patients per surgeon is maintained at the level as observed in the original SPRINT study. In combination, the overall power of the alternative designs may be summarized by the ratio of the anticipated treatment effect to its standard error. Relative to the conventional design, we see that the EB designs both have an effect/SE ratio of approximately 50 per cent, suggesting a substantial loss of power in this particular example. Of course, in other examples the EB design might turn out to be *more* powerful than the conventional design.

DISCUSSION

As we have seen, the EB design presents some potential advantages over the conventional design. However, these advantages must be contrasted with potential disadvantages and questions of feasibility of executing the EB design. One of the major feasibility issues in implementing the EB design is that patients will be randomized to a clinician in the A group or the B group, and so two clinicians must be on hand to administer the randomized assignment to a given patient. This is obviously more difficult for acute care interventions such as emergency surgery, but it is less of an issue for elective surgery. However, even for

elective procedures, patients may be unwilling to travel to different facilities to receive their intervention, so the EB design is perhaps most feasible in the context of group practices or large hospitals where teams of clinicians are available, having a mixture of expertise. Even if multiple clinicians are available in the same institution, one still has the practical difficulty that the patient will have consulted one clinician at an early stage, but then may be randomized to an intervention requiring the expertise of a different clinician later on. Explanation of this complication to the patients may be difficult and a potential disincentive to participate in the randomized trial, compared with the patient who deals with only one clinician in the conventional design.

Quite apart from the direct impact of expertise on patient outcomes in the way we have discussed, there are other potential advantages of the EB design that may enhance its estimated treatment benefit. First, the EB design recognizes clinician preferences, and by definition involves them in treating patients exclusively with a technique with which they are familiar. In turn, this is likely to lead to less frequent procedural crossovers, compared with the situation, where for 50 per cent of patients clinicians will be using the technique with which they are less familiar, less expert, and therefore possibly less comfortable. Additionally, differential co-interventions may be reduced when surgeons are all using the procedure in which they are invested.

It is well known that crossovers are likely to attenuate the difference between randomized treatment groups, leading to a loss of power in the intention to treat analysis, which is the preferred analysis to avoid possible selection biases associated with alternative analyses such as the per protocol or as treated in [15–17]. Even a relatively small proportion of patients crossing over to the non-randomized intervention can lead to a substantial reduction in the expected difference in outcomes between randomized groups.

An additional advantage of the EB design is that it may lead to improved recruitment of clinicians and patients. Clinicians may be more willing to participate in a study if they know that they will always be using their preferred technique. Patients may be more willing to enrol in a randomized trial if they can be assured that their treating clinician will be an expert in the technique to which they will be randomized. Both of these design features are thought to be stronger from the ethical perspective.

For reasons of generalizability, we would recommend that the EB design be used primarily in situations where moderate or large numbers of surgeons and centres can be involved. In contrast, EB studies with very small numbers of surgeons and centres would lead to results whose generalizability would remain uncertain. However, the reality is that most surgeons tend to select a specific approach to manage a given clinical problem, and therefore the conventional design is unlikely to overcome the obvious problems of generalizability of a small EB trial. For instance, consider the smallest possible case involving only two surgeons. In such a situation, it is probable either that both surgeons would have greater expertise in one of the interventions, thus biasing the trial toward that intervention, or that one surgeon has expertise in one technique and the other surgeon has expertise in the other technique. In this situation, it is unlikely that a conventional design would provide a more valid result than

the EB design. Similar concerns would affect studies with more but only very limited numbers of surgeons.

In our example of the SPRINT study, the EB design was disadvantaged both with respect to a larger standard error and a smaller treatment effect, compared with the conventional design. However, the latter is not always true, and in other situations the EB design might yield a larger treatment effect, in particular when the expertise advantage is associated more strongly with the intervention that has inherently better outcomes. In such a case, there would be a trade-off between the loss of statistical efficiency through a larger standard error with an enhanced estimated treatment effect, and it is therefore possible that the EB design might on balance be more powerful than the conventional design.

In our example, we calculated the expected treatment effect in the EB design by combining the expected advantages associated with surgeon expertise with the inherent effects of the treatment themselves. As such, these estimates might be thought of in the sense of treatment efficacy, because each treatment would be used in relatively ideal circumstances, with an expert surgeon. It is interesting to note that while the relative advantage of the reamed nail compared with the unreamed nail was eroded when expertise was taken into account, nevertheless all of the expertise effects were positive. In other words, the expected outcomes for patients receiving either intervention would be expected to improve when surgery is carried out by an expert in the relevant technique. Thus, while the EB design would show a smaller overall advantage of the reamed nail, it could be argued that the EB design is preferred from the ethical viewpoint, because expected outcomes for patients in both groups would be improved compared with the conventional design.

As noted earlier, the adjustment for expertise effects may produce an increase or decrease in the unadjusted estimated treatment effect. In the scenario where the expertise adjustment results in an increased treatment effect, the efficacy interpretation of study results would lead one to recommend additional training in the superior intervention, so that more clinicians may acquire the expertise necessary to use it.

A further scenario is where the adjustment for expertise causes the advantage of the unadjusted estimate of benefit for treatment A to be completely negated (or even exceeded). This type of result would indicate that additional training to gain expertise in B could potentially improve patient outcomes to the same level as those for patients with treatment A (even when delivered by clinicians expert in A), or even to exceed outcomes for patients on A. This would amount to a qualitative reversal in the interpretation of the analyses adjusted or unadjusted for expertise.

A counter-argument to making the adjustment for expertise would be that treatment effects should be thought of in terms of effectiveness, as opposed to efficacy, and should therefore remain unadjusted for expertise effects. This position is reasonable in situations where either treatment A or B could be recommended for particular individual patients, depending on other considerations related to their clinical circumstances, or because of patient values and preferences. Thus, there may be good reasons why clinicians might decide to administer one or the other of two alternative interventions, even though they might not be equally expert in

both. Here, the treatment estimate, being unadjusted for expertise effects, would represent the benefit anticipated in routine clinical practice, and as administered by experts or non-experts, as the case may be.

We feel that on balance, for situations such as those indicated through our SPRINT example, the acquisition of relevant expertise would be highly desirable. This is because the outcomes of patients may be expected to improve with either intervention, if it is delivered by an expert on the technique in question. If the treatment effect is additionally enhanced through the expertise adjustment, that argues even more strongly for additional training to acquire expertise in the superior technique. If the expertise adjustment leads to similar expected outcomes (or even a reversal of the unadjusted results) one would recommend additional training to acquire expertise in the technique that is inferior in the unadjusted analysis.

An issue for the analysis is how to represent the expertise effects. In our tibial fracture example, we characterized surgeons as being expert or not on each of the comparison surgical techniques, and both were taken into account in our model. An alternative would be to associate the expertise effect only with the particular technique in use for a given patient. Thus, for example, while a given surgeon might be an expert in both techniques, one could decide that only his experience with the particular randomized intervention at hand would be relevant. Thus, for example, experts in technique A would have no advantage over non-experts in A if their patients had been randomized to intervention B.

We defined expertise as having performed more than the median number of operations. In greater generality, we might represent expertise in more detail, for instance according to the actual number of previous operations carried out, during residency training, since completion of training, or in some period immediately before the start of the randomized trial. More detailed control of the expertise effect would be desirable in situations where it is known that a considerable amount of experience is required before no further improvements to patient outcomes might be anticipated, such as was mentioned for hernial surgery. This type of an example suggests that a run-in period or other requirement that clinicians have been involved within some minimal number of cases before the start of the study may not resolve the problem of differential expertise bias in the conventional design.

In our description of the EB design, we have focused on the notion that patients are randomized to a surgeon who has expertise in one technique or the other. We also mentioned the possibility of a hybrid design, to include surgeons who are expert in both techniques, and whose patients would therefore be randomized to interventions as usual. It is possible that surgeons may have had experience with both interventions, but they might state that they do not have a preference for either. Here the crucial question would be whether they have greater expertise in one or the other of the interventions. If they do, then the EB design is still relevant. If a surgeon has truly no preference and equal expertise in both techniques, then they could either participate in a conventional design or the hybrid design, depending on the beliefs and expertise of the other participating surgeons.

There is the possibility that some surgeons may be ‘intrinsically innovative’ and relatively willing to acquire new skills with alternative surgical techniques. In contrast, traditional

surgeons may not be as willing to change. It is clear that we would not wish to force the traditional surgeons into attempting a new technique with which they were uncomfortable, but such surgeons can nevertheless participate in an EB trial design on the conventional technique. If the trial results suggest better patient outcomes for the innovative technique, one would then need to consider the necessary training that would be required in order to permit the ‘traditional’ surgeons to incorporate the new technique into their clinical practice.

In summary, we feel that the ED design offers some attractive features, both to participating clinicians and to patients. The EB design does offer a number of potential advantages, including the minimization of differential expertise bias, reducing the impact of having unblinded interventions, reducing treatment crossovers, and potentially greater ethical strength. We recognize that disadvantages include the confounding of surgeons for treatment, and potentially limited generalizability. However, the conventional design may also have limited generalizability if clinicians are required to use interventions for which they have less experience. We primarily advocate the EB design for common surgeries (and other clinical interventions) for which it is felt that all competent clinicians could develop expertise, if the study results suggested that this was appropriate. However, more experience is needed to assess its practical feasibility, and to evaluate its statistical properties relative to the conventional randomized trial design. The framework given in this paper provides a first step in this direction.

Acknowledgments

The authors thank Prof. Gordon Guyatt of the Department of Clinical Epidemiology and Biostatistics (McMaster University) for his comments on an early draft of this paper.

References

1. Devereaux PJ, Bhandari M, Walter SD, Sprague S, Guyatt G. Participating surgeons' experience with and beliefs in the procedures evaluated in a randomized controlled trial. *Clinical Trials*. 2004; 1:225.
2. Devereaux PJ, Bhandari M, Clarke M, Montori VM, Cook DJ, Yusuf S, Sackett DL, Cina CS, Walter SD, Haynes B, Schunemann HJ, Norman GR, Guyatt GH. Need for expertise-based randomised controlled trials. *British Medical Journal*. 2005; 330(7482):88. [PubMed: 15637373]
3. Van der Linden W. Pitfalls in randomized surgical trials. *Surgery*. 1980; 87:258–262. [PubMed: 6987765]
4. DeTurris SV, Cacchione RN, Mungara A, Pecoraro A, Ferzli GS. Laparoscopic herniorrhaphy: beyond the learning curve. *Journal of the American College of Surgeons*. 2002; 194:65–73. [PubMed: 11800341]
5. Menon VS, Manson JM, Baxter JN. Laparoscopic fundoplication: learning curve and patient satisfaction. *Annals of the Royal College of Surgeons of England*. 2003; 85:10–13. [PubMed: 12585623]
6. Lobato AC, Rodriguez-Lopez J, Diethrich EB. Learning curve for endovascular abdominal aortic aneurysm repair: evaluation of a 277-patient single-center experience. *Journal of Endovascular Therapy*. 2002; 9:262–268. [PubMed: 12096938]
7. Rudicel S, Esdaile J. The randomized clinical trial in orthopaedics: obligation or option? *Journal of Bone and Joint Surgery*. 1985; 67:1284–1293.
8. Neumayer L, Giobbie-Hurder A, Jonasson O, Fitzgibbons R Jr, Dunlop D, Gibbs J, Reda D, Henderson W. Open mesh versus laparoscopic mesh repair of inguinal hernia. *New England Journal of Medicine*. 2004; 350(18):1819–1827. [PubMed: 15107485]

9. Donner, A., Klar, N. Design and Analysis of Cluster Randomization Trials in Health Research. Arnold Publisher; London, U.K: 2000.
10. Murray, DM. Design and Analysis of Group Randomized Trials. Oxford University Press; New York: 1998.
11. Consort group. Consort statement: extension to cluster randomized trials. British Medical Journal. 2004; 328:702–708. [PubMed: 15031246]
12. The SPRINT Investigators. Study to prospectively evaluate reamed intramedullary nails in patients with tibial fractures (S.P.R.I.N.T.). Journal of Bone and Joint Surgery. :90A. in press.
13. Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Medical Care. 1992; 30:473–483. [PubMed: 1593914]
14. McHorney CA, Ware JE Jr, Raczek AE. The MOS 36-item short-form health survey (SF-36). II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. Medical Care. 1993; 31:247–263. [PubMed: 8450681]
15. Haynes RB, Dantes R. Patient compliance and the conduct and interpretation of therapeutic trials. Controlled Clinical Trials. 1987; 8:12–19. [PubMed: 3568692]
16. Lachin JM. Statistical considerations in the intention-to-treat principle. Controlled Clinical Trials. 2000; 21:167–189. [PubMed: 10822117]
17. Piantadosi, S. Clinical Trials: A Methodologic Perspective. Wiley; New York: 1997.

Figure 1.
Schematic diagram of expertise-based randomized trial design.

Figure 2.

Expected results in expertise-based randomized trial of surgery for tibial fractures, according to surgeon expertise.

Table 1
Analysis of variance for conventional RCT design, taking clinician (surgeon) effects into account.

Source	Df	SS	MS	E(MS)	F
Treatment	1	SST	SST	$\sigma^2 + k' m' \sum \alpha_t^2$	$\frac{MST}{MSE}$
Surgeon	$k' - 1$	SSS	$\frac{SSS}{k' - 1}$	$\sigma^2 + 2m' \sigma_S^2$	$\frac{MSS}{MSE}$
Error	$k' (2m' - 1) - 1$	SSE	$\frac{SSE}{k' (2m' - 1) - 1}$	σ^2	
Total	$2k' m' - 1$	SSTL			

Analysis of variance for expertise-based RCT design, taking clinician (surgeon) effects into account.

Table II

Source	Df	SS	MS	E(MS)	F
Treatment	1	SST	SST	$\sigma^2 + m\sigma_s^2 + k:m \sum \alpha_t^2$	$\frac{MST}{MSS}$
Surgeon	$2(k-1)$	SSS	$\frac{SSS}{2(k-1)}$	$\sigma^2 + m\sigma_s^2$	$\frac{MSS}{MSE}$
Error	$2k(m-1)$	SSE	$\frac{SSE}{2k(m-1)}$	σ^2	
Total	$2km-1$	SSTL			

Table III

Expected results of a randomized trial of surgery for tibial fractures with an expertise-based design.

		Estimates
<i>Estimated effects of expertise</i>		
<i>Expertise</i>		
A and B		2.83
A only		1.78
B only		2.24
Neither		0.00
<i>Predicted expertise effect in EB design</i>		
A-arm	$\frac{1}{2}(2.83+1.78)$	2.31
B-arm	$\frac{1}{2}(2.83+2.24)$	2.54
Net	2.31–2.54	–0.23
Effect size (assumes uniform distribution of expertise)	0.64–0.23	0.41

Table IV

Efficiency comparison of two expertise-based designs vs conventional study design, in a randomized trial of surgery for tibial fractures.

Design	Effect	SE	Effect/SE	Ratio
EB-I *	0.41	0.81	0.51	0.49
EB-II	0.41	0.71	0.58	0.55
Conventional	0.64	0.61	1.05	1.00

* See text for details of expertise-based designs I and II.