Review

Pro/con clinical debate: It is acceptable to stop large multicentre randomized controlled trials at interim analysis for futility

David A Schoenfeld[1] and Maureen O Meade[2]

[1]Professor of Medicine, Harvard Medical School, Professor, Department of Biostatistics, Harvard School of Public Health, Massachusetts General Hospital, Boston, Massachusetts, USA
[2]Associate Professor, Department of Medicine and Department of Clinical Epidemiology & Biostatistics, McMaster University, Hamilton, Ontario, Canada

Corresponding author: David A Schoenfeld, dschoenfeld@partners.org

Published online: 9 December 2004
This article is online at http://ccforum.com/content/9/1/34

Critical Care 2005, **9**:34-36 (DOI 10.1186/cc3013)

Abstract

A few recent, large, well-publicized trials in critical care medicine have been stopped for futility. In the critical care setting, stopping for futility means that independent review committees have elected to stop the trial early – based on predetermined rules – since the likelihood of finding a treatment effect is low. For bedside clinicians the idea of futility in a clinical trial can be confusing. In the present article, two experts in the conduct of clinical trials debate the role of futility-stopping rules.

Keywords clinical research, futility, interim analysis, randomized controlled trials, stopping rules

The scenario

You are a clinician in an intensive care unit and you have recently heard that some very large trials have been stopped at interim analysis for futility. Although you have not yet seen the results, this cessation concerns you because you were anxiously awaiting the results of these trials since you felt they were very relevant clinical questions that would impact on your treatment decisions. Your concern is based on the fact that you are uncertain whether clinical trials should ever be stopped for futility.

Pro: Futility stopping can speed up the development of effective treatments

David A Schoenfeld

A futility-stopping rule for a clinical trial is a plan in which the results of a clinical trial are periodically reviewed and the clinical trial is stopped if the treatment difference is smaller than some predetermined value. The idea is to stop trials that would not have shown statistical significance had they gone on to completion. A futility-stopping rule can drastically reduce the time and money spent on clinical trials, and can more rapidly find effective treatments. In the present paper I describe the available methods used for futility stopping. I then quantify the advantages of futility stopping in a drug development programme. Finally, I will discuss some of the problems of futility stopping and how clinicians should interpret trials that stop early for futility.

There are two methods of futility stopping. The method that was used in early trials was based on the principal of stochastic curtailment [1]. A review committee would analyse the results of a trial and calculate the probability that the trial will give a significant result if it is completed. If this probability was small, say less than 25%, then the trial would be stopped. This probability calculation depends on an assumption about the actual success rates of the treatments.

PEEP = positive end-expiratory pressure.

The safest assumption is to use the original difference that was used to calculate the sample size.

The second method is to use asymmetric stopping boundaries [2,3]. The futility boundary can be based on how quickly you want to stop the trial if the treatment is ineffective. The faster you stop for an ineffective treatment, however, the larger the sample size needs to be to detect a difference if it is there.

Suppose we use a futility-stopping rule based on the work of Demets and Ware [3]. Using this rule to detect a 10% difference in mortality (from 30% to 20% mortality) will require a maximum sample size of 830 patients, rather than the 800 patients required without this rule. If the drug is ineffective, however, it is likely that the trial will be stopped early. A more important number is the expected sample size, which is the average sample size if the trial was repeated over and over again. This expected sample size is 480 patients, a saving of 320 patients over the 800 patient sample size that we would have without futility stopping.

The greatest disadvantage of a futility-stopping rule is that it is much more difficult to interpret a negative study. If the study was stopped for futility then the confidence bounds will be much wider than they would have been had the study continued. Furthermore, the estimated treatment difference is biased downward. There are ways to compensate for this but they are computationally difficult [4]. This bias is particularly troublesome when these estimates are used in a meta-analysis. Clinicians should be careful not to overinterpret the negative evidence from such a trial. Futility-stopping rules should not be used when large segments of the community already believe that a treatment is effective.

Con: the hazards of stopping for futility

Maureen O Meade

Large trials in critical care assume a broad mandate. Objectives typically include determining effects on survival and other important outcomes, estimating complication rates, and identifying predictors of response. The over-riding goal is to advance clinical care. With this assurance, study patients assume risk, clinicians devote their energy, and sponsors invest financially.

'Futility' implies little hope of achieving study objectives with the planned sample size. Ironically, stopping for futility undermines each of the aforementioned objectives. To highlight these issues, I will discuss the ALVEOLI trial that compared higher positive end-expiratory pressure (PEEP) with lower PEEP in the management of acute respiratory distress syndrome [5]. My choice of example implies no adverse criticism of the investigators, who conducted their study with the highest scientific and ethical standards.

Stopping for futility leaves the primary research question unanswered. The ALVEOLI trial suggested higher PEEP might cause harm (relative risk of mortality, 1.11). The 95% confidence interval tells us, however, that the data are consistent with a relative risk as low as 84% and as high as 146%. Clinicians would employ PEEP very differently across this spectrum of possible 'truths'. Most critical care clinicians agree that a mortality reduction of 16%, if true, would be important.

Early stopping for futility increases the risk of imbalance in prognostic factors. The ALVEOLI trial was complicated further by baseline differences between groups with respect to two important predictors of survival: age and severity of lung injury. Adjusting for these imbalances, a valid and necessary procedure, the study results flip to support higher PEEP – continuing the trial could have clarified this issue.

Early stopping similarly jeopardizes analyses of secondary outcomes, which may be pivotal in clinical decisions when there is truly no survival effect. Data related to adverse events are limited, and subgroup analyses thwarted.

Finally, there is an opportunity cost. Does higher PEEP improve outcomes in acute respiratory distress syndrome? Unfortunately, it is unlikely that investigators will conduct this trial again, on a greater scale and with the same high quality and singularity of question.

These scientific penalties are compounded by a lack of standards for integrating statistical and judgemental criteria for early stopping. The lack of such standards increases the likelihood that stopping decisions will be idiosyncratic or self-interested. For instance, lack of standards may permit industry sponsors to characterize a trial suggesting that a therapy is harmful as a trial terminated 'for futility'.

Finally, choices about presentation may increase the risk of misinterpretation. The ALVEOLI trial stopped when '… the probability of demonstrating the superiority of the higher-PEEP strategy was less than 1% under the alternative hypothesis based on the unadjusted mortality difference'. While the authors are trying to be transparent, clinicians may not realize that the alternative hypothesis is a mortality reduction 'from 28% … to 18%' – an effect that is both very large and, perhaps, implausible. This characterization may increase the likelihood that clinicians will interpret the ALVEOLI trial demonstrating higher PEEP as ineffective; in fact, as I have noted, the results remain consistent with an important effect.

Clinical investigators have a responsibility to consider the effects of their research upon the totality of literature in their field. Stopping early for futility undermines the best of intentions.

Pro's response: who remembers lisofylline?

It is not surprising that Maureen Meade's article focused on the National Heart, Lung, and Blood Institute acute respiratory distress syndrome network ALVEOLI trial [5], which stopped early after 549 patients, rather than on the lisofylline study [6] that preceded it, which stopped early after 235 patients. Had the network not stopped the lisofylline trial early they might not have conducted the ALVEOLI trial, and we would not have had any data on treated patients with high PEEP and low PEEP – however inadequate these data are, in Maureen Meade's opinion. The fact that there may be disagreement about whether to use a futility-stopping rule for a particular trial does not negate the value of this strategy.

Con's response: Dr Meade's response

I agree with Dr Schoenfeld on many counts, and particularly that the efficient use of research resources (funding, participants and time) is paramount. In my view, the conduct of studies that cannot answer the intended question – by design, or as a result of interim decisions – represents suboptimal use of limited resources.

While there are no clear answers to this controversy, I believe that there is strong theoretical evidence that stopping for futility is often misguided, and I look forward to seeing new empirical research in this field.

Acknowledgement

Dr Meade is a Peter Lougheed Scholar of the Canadian Institutes for Health Research.

References

1. Turnbull BW: **Stochastic curtailment.** *Encycloped Stat Sci* 1997, **1:**521-523.
2. Schoenfeld DA: **A simple algorithm for designing group sequential clinical trials.** *Biometrics* 2001, **57:**972-974.
3. Demets DL, Ware JH: **Asymmetric group sequential boundaries for monitoring clinical trials.** *Biometrika* 1982, **69:**661-663.
4. Jennison C, Turnbull BW: *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton, FL: CRC Press; 2000:171–189.
5. The National Heart, Lung, and Blood Institute ARDS Clinical Trials Network: **Higher versus lower positive end-expiratory pressures in patients with the acute respiratory distress syndrome.** *N Engl J Med* 2004, **351:**327-336.
6. Anonymous: **Randomized, placebo-controlled trial of lisofylline for early treatment of acute lung injury and acute respiratory distress syndrome.** *Crit Care Med* 2002, **30:**1-6.