DRAFT

Cochrane Effective Practice and Organisation of Care Group
Methods Paper
Issues Related to Baseline Measures of Performance

Introduction

When conducting systematic reviews of studies that are included in EPOC's scope, reviewers will note that some of the studies will include a baseline measure of performance while others will not. The purpose of these guidelines is to help reviewers to make decisions about interpreting the quality and the analysis of studies with and without baseline measures.

Rationale for baseline measures

Baseline measures are useful because they provide an estimate of the magnitude of a problem. Low performance scores prior to the intervention may indicate that performance is poor and there is a lot of room for improvement. On the other hand, high performance scores may indicate that there is little room for improvement (ceiling effect). In a randomised or quasi-randomised study, comparing baseline data in the experimental and control groups can provide some reassurance about the adequacy of the allocation process.

Baseline imbalance

Randomisation of allocation units (practice, provider, patient) to experimental or control conditions seeks to minimise selection bias in a study. In many studies of behaviour change, the number of allocation units in a study may be small. This is particularly the case in studies when the unit that is randomised is a cluster such as a practice. When there are small numbers in a study, randomisation may be inadequate to balance the groups on important variables such as baseline performance. Therefore, it is possible that any difference between groups at baseline could be potentially greater than the difference between groups post intervention. Imbalance at baseline may indicate less than adequate study design.

Example of baseline imbalance

(From Anderson 1994)

In this study, hospitals were assigned 'by lot' to one of three groups: CME (continuing medical education); QA (CME plus quality assurance); or control (aware of study). The targeted behaviour was the use of prophylaxis for venous thromboembolism for hospitalised patients. The data were analysed at the level of the patient (percentage of patients who received prophylaxis). Table 1 indicates the data as reported in the paper. Table 2 provides a suggestion of how data could be summarised in a review.

Table 1 From Anderson 1994 Percentage of patients receiving prophylaxis

	Baseline	Post	P
CME	21%	49%	<.001
QA	27%	55%	<.001
Control	40%	51%	<.001

This study illustrates many of the problems encountered when completing a review. Firstly, the unit of analysis is incorrect: the unit of randomisation was the hospital but the unit of analysis was the patient. This means that correlation within a hospital has not been considered in the analysis, resulting in overly narrow confidence intervals or spuriously low p values. Secondly, a comparison between or among groups has not been done but rather, a comparison within a group before and after an intervention. Therefore, the experimental and control interventions cannot be compared. Thirdly, the groups are not balanced at baseline. The authors indicate this in the paper, but the analysis does not take this into consideration. Examination of baseline performance shows the control group performing significantly better than either experimental group. If baseline performance were ignored, then it would appear as if there is little difference among the groups post-intervention (-4% relative difference indicating that the CME group got worse). The difference between groups can be seen in Table 2 by reporting the percentage change. This shows the difference between the groups standardised by the control difference. Since there are no estimates of variation for any of the groups, the difference cannot be tested. Reviewers should report the baseline, post scores, absolute difference, and standardised difference. Table 2 provides some suggestions about how these data can be summarised in a review.

Table 2 Example from Anderson 1994. Percentage of patients receiving prophylaxis

	Baseline	Post	Baseline-Post Difference	(%) Change*
CME	21%	49%	28%	155%
QA	27%	55%	28%	155%
Control	40%	51%	11%	-

* % change= (baseline-post score)/(control baseline-post score)

Adjusting for baseline performance

It may be reasonable to expect that there will be variation in a sample, with some individuals performing well and others performing below that which is desired. It is possible that baseline performance may be a good predictor of post intervention performance (ref). Variation in individual performance should be taken into consideration in the analysis. In a primary study, the original investigators can include baseline performance as a co-variate, thereby adjusting the post intervention scores. Alternatively, investigators may calculate two change scores, one for each group and then compare the magnitude of the change score for the two groups. For example, McConnell and colleagues (1982) designed a study to reduce the use of tetracycline for upper respiratory tract infection. At baseline, the experimental group prescribed an average of 12.6 prescriptions over 6 months while the control group prescribed 7.5 prescriptions over the same time period. After the intervention, the experimental group average was 1.8 and the control was 3.2. The authors used an analysis of co-variance to show that the reduction in the mean number of prescriptions in the experimental group was significantly greater than that of the control group (F=4.34,P<0.05).

- When including a study in an EPOC systematic review, reviewers should report baseline data and both unadjusted and adjusted scores (if available) in the results tables.
- In general, reviewers should not attempt a post-hoc adjustment for baseline imbalance unless the complete data set is available.

Reporting change scores or post scores

Sometimes, primary authors will use change scores in an attempt to correct for baseline imbalances. Although, this is less desirable than using an analysis of co-variance or similar approach, it is preferable to ignoring baseline imbalance and reporting only post-intervention scores. Reviewers should indicate in the table or text of the review that a change score was used in the analysis. The standard deviation of the change score is not the same as the standard deviation of a post score and cannot be treated as such. The appropriate test of an intervention is the difference between two groups rather than the difference within a group from pre to post intervention.

For example, Soumerai and colleagues (1998) reported the results of a randomised controlled study to assess the impact of local opinion leaders to improve the care of patients post myocardial infarction. The median change in the proportion of elderly patients receiving aspirin was +0.13 (17% increase from 0.77 at baseline) compared with a change of –0.03 (-4%) in control hospitals. The authors reported that the magnitude of the change between the experimental and control groups was statistically significant (P=0.04, one tailed) using a Wilcoxon rank sum test. In this example, both groups were similar at baseline. They also assessed the use of beta-blockers. At baseline, the mean proportion of patients receiving beta-blockers was 0.49 while the control group proportion was 0.60. Although the authors reported that the difference was not statistically significant, it is preferable to account for this difference (as the authors did by using change scores). In this example, it would not be helpful for reviewers to calculate post-intervention scores by adding the absolute difference to the baseline score as this would result in an unadjusted post score.

- If a mean change score and standard deviation is reported in a primary study, then reviewers should report these data in a review
- If the experimental and control groups are not balanced at baseline (and the units are the same before and after the intervention), it is preferable to calculate two change scores, one for each group and then compare the change score for the two groups. See example from Anderson discussed previously.
- If baseline values are similar, reviewers may calculate post intervention differences between groups.

Does a lack of baseline measurement indicate a poor quality study?

In large randomised or quasi-randomised studies, the randomisation process is likely to result in reasonably balanced baseline performance. The problem of not knowing the baseline prevalence of a problem remains but is not a reflection of study quality. One can assume that the control post-intervention performance approximates the baseline performance. The problem arises when the study sample is small and there is no baseline measurement. In this situation, it is not safe to assume the groups would have been balanced had baseline performance been measured. The following decision rules may be helpful.

- Is baseline performance measured?

If yes, is the absolute difference between the groups less than 10%?

> If yes, score the quality criterion as DONE.
>
> If no, did the analysis take into consideration the baseline imbalance (for example, analysis of co-variance or analysis by change scores between groups?

If yes, score the quality criterion as DONE.
If no, score the quality criterion as NOT DONE.

If no, then score the criterion as NOT CLEAR. Note, some studies may be large enough for reviewers to be reasonably confident that selection bias has been minimised.

Summary

- Imbalance at baseline is common in studies of behaviour change
- When including a study in an EPOC systematic review, reviewers should report baseline data and both unadjusted and adjusted scores (if available) in the results tables.
- In general, reviewers should not attempt a post-hoc adjustment for baseline imbalance unless the complete data set is available.
- If a mean change score and standard deviation is reported in a primary study, then reviewers should report these data in a review.
- If the experimental and control groups are not balanced at baseline (and the units are the same before and after the intervention), it is preferable to calculate two change scores, one for each group and then compare the change score for the two groups. See example from Anderson discussed previously.
- If baseline values are similar, reviewers may calculate post intervention differences between groups.