original article

Baseline comparisons and covariate fishing: Bad statistical habits we should have broken yesterday

Stefan L. K. Gruijters
Maastricht University

Introduction

Checks on baseline differences in randomized controlled trials (RCTs) are often done using null-hypothesis significance tests (NHSTs). In a quick scan of recent publications in the journal Psychology and Health, from 2015 to most recent (accessed 4-2-2016), I noticed that it is common for RCTs to include results of NHSTs on baseline variables, and this tendency seems pervasive throughout the literature. In itself, the enterprise of establishing baseline similarity across conditions is a worthwhile venture, since empirical conclusions based on non-comparable samples would hamper progress in the field. The ability of RCTs to provide unbiased estimates of causal relationships between variables is crucial for scientific progress. Poor tests of theory, misguided follow-up research, misapplication of theory in practice, and waste of research funds: all hang in the balance.That being said, the use of NHSTs to establish the degree of baseline similarity is inappropriate, potentially misleading (Altman & Doré, 1990; De Boer, Waterlander, Kuijper, Steenhuis, & Twisk, 2015; Roberts & Torgerson, 1999; Senn, 1994), and, simply, logically incoherent.

NHSTs on baseline variables are often done under the guise of 'establishing whether randomization was successful', or 'identifying potential confounds and covariates to control for in further analyses'. Despite a large number of authors (e.g. Altman & Doré, 1990; Austin, Manca, Zwarenstein, Juurlink, & Stanbrook, 2010; De Boer et al., 2015; Roberts & Torgerson, 1999; Senn, 1994) who have argued against the use NHSTs to compare baseline differences in RCTs, or as basis for covariate selection, the habit appears hard to eradicate. De Boer et al. (2015) speculate that a we-do-as-others-do tendency, perhaps a form of Bandurian learning, might underlie the persistence of researchers, reviewers, and editors to report and request such tests.

In what follows, I discuss several issues related to this practice, including 1) whether the use of NHSTs as a method of checking randomization procedures is appropriate, and 2) whether selection of covariates is feasible on this basis. The arguments described here are not new or complex, but worth repeating given the persistent habit to involve NHSTs in baseline comparisons. Alternatives and suggestions for improvement on both of the above points will be briefly discussed.

Testing for baseline differences

The CONSORT statement (Moher et al., 2010), to which many medical and epidemiological journals adhere, explicitly states that NHSTs should not be used to test for baseline differences. Instead, descriptive information about baseline data across conditions, combined with proper description of randomization procedures should be given. This statement does not simply follow arbitrary convention, but is rooted in the logic that NHSTs can only result in type I – errors (falsely rejecting a true null hypothesis). To illustrate this, consider first that in a random assignment procedure the samples are by definition drawn from the same population, since all variables have the same

expected means and distributions across samples. For a given variable, both sample means are thus estimators of the same population parameter. Of course, in a single randomization, sample estimates can show large fluctuations on specific baseline variables; depending on the population standard deviation (σ), and size (n) of the samples.

Accordingly, researchers often proceed testing these observed baseline difference for significance, and this is where practice runs into a logical caveat. There is nothing against calculating descriptive statistics to check baseline similarity across groups, including an appropriate test–statistic and a corresponding probability (p). The p-value then tells us something about how likely a given baseline difference is, given that we are randomizing individuals from the same population into samples of size n[1]. However, to involve this p-value in a null-hypothesis test against a rejection criterion (a significance level of .05, say) - a move towards inferential statistics - is logically incoherent.

Consider a two-group t-test comparison on a scale level baseline variable. To easily spot the error, the tested null-hypothesis on baseline similarity (i.e. h0: $\mu 1=\mu 2$), can also be phrased as: "Both samples come from the same population", which – as described above- we already know is the case given random assignment. Thus, when researchers decide to reject the null-hypothesis of baseline similarity in a RCT (given $p < .05$), they are in effect implying that samples drawn from the same population are not from the same population[2]. Because this is a logical contradiction, the only conclusion that follows from a significance conclusion on baseline dissimilarity is that a Type-I error has been made. Indeed, it seems quite bizarre to examine the evidence against a null-hypothesis that a priori we know to be true in RCTs.

The above argument, of course, hinges on the notion that randomization used in a particular study was in fact truly random (i.e. the study is in fact a RCT). To determine this, researchers should consider whether the procedure used to randomize resulted in a given person drawn from the population to have equal probability of being assigned to each group. For example, a simple randomization procedure in which a set of random numbers is generated using computer software can impossibly be biased – i.e. given a proper algorithm underlying the number generator. To inform reviewers and readers about whether 'randomization was successful', researchers should thus refer to the method of randomization instead of supplying NHSTs.

There are instances where a randomization procedure does not guarantee that the samples in a study reflect the same baseline population originally randomized into an RCT. In this sense randomization is a necessary, but not sufficient reason to assume an unbiased estimate of an experimental effect. For example, individuals with specific characteristics might be more prone to drop out in one of the conditions (due to the condition). Such missing data resulting from non-random drop out complicates matters further, and might require additional steps in order to ensure an unbiased estimate of an experimental effect (Groenwold, Donders, Roes, Harrell, & Moons, 2012).

There are other (randomization-related) circumstances where researchers would need to control for baseline differences. In the case of non-randomized pre-test / post-test designs, though, the question is whether this should be done using analysis of covariance (ANCOVA) or analysis of variance (ANOVA) using change scores (see for discussions, Van Breukelen, 2006, 2013). In RCTs, the crucial point is that randomization issues – and potential bias - can be anticipated by scrutinizing the RCT methodologically (the potential for

1 I leave aside, though, whether probability information of finding specific sample differences in a single randomization is actually informative.
2 This violates a fundamental doctrine of logic: something either is, or is not, and if something is, it cannot be that it is not.

selection bias, missing not at random, and the chosen randomization method) and not statistically using NHSTs.

Covariate selection in randomized controlled trials

Although NHSTs on baseline variables are meaningless in RCTs, this does not imply that chance differences on baseline variables cannot influence the estimate of an experimental effect. In randomized studies, baseline variables (such as demographics, trait measures, and pre-measures of outcome variables) are very often included using ANCOVA models, which applies a linear adjustment to the experimental effect, correcting for between-groups differences on the covariate. As discussed, the decision to include a covariate should not be made based on NHSTs. Moreover, this decision should also not be based on probability values of group differences on a potential covariate. Small p-values for baseline differences do not imply that a particular covariate is worth including a model. Instead, the size of the association between covariate and outcome (in terms of coefficient r, or other standardized indices of effect size) are more clear indicators of a covariate's potential contribution.

It is worth noting that the inclusion of covariates in RCTs (due to randomization) rarely alters conclusions about the size of an experimental effect in the population, i.e. adjusts for confounding. However, adjustment for covariates might affect conclusions about the significance of an experimental effect, due to the resulting increase of statistical power by reduction of error variance. The value of ANCOVA models in RCTs, then, lies in the potential of covariates to decrease the error variance in the outcome variable, not so much in decreasing bias of the estimate of an experimental effect (Van Breukelen & Van Dijk, 2007). This notion is of importance in deciding to include covariates in the analyses, since an imbalance of a covariate across conditions, i.e. the association of a condition variable (X) and covariate (C), is less relevant for the power to detect an experimental effect than the strength of the relationship between the covariate (C) and the outcome variable (Y). The correlation between a covariate and outcome is, thus, a more relevant criterion for inclusion in a model than the existence of baseline differences on the covariate.

When covariates are selected on the basis of substantial influence (as opposed to significance) on an outcome variable, it is unlikely that researchers run into these in an exploratory fashion. Instead, such variables are included in the study protocol in the first place because of the literature suggesting their relevance. In this sense, covariate selection should always be confirmatory, and be included in the study protocol and analysis regardless of any baseline differences (Senn, 1994). Inclusion of covariates on the basis of sample information, indicating baseline dissimilarity across conditions, or an unexpected effect on outcome (Y) is at best statistically suspect, and is an unwarranted form of covariate fishing. This habit might lead meaningful covariates (with an hypothesized, perhaps replicated effect on the outcome variable) to decrease in their potential contributions to the model. In addition, the inclusion of such 'fished' covariates leads to a loss of parsimony, and meaningless corrections to the estimated experimental effect. In sum, two suggestions for improvement of RCT analysis arise from the above discussion:

1) There is no scientific justification for using NHSTs as a tool to establish baseline comparability; researchers should stop doing it, and reviewers and editors should stop asking for it.

2)Covariate selection should be made solely a priori and based on importance of association – implying that covariates should be specified in advance in the study protocol and listed in papers' methods sections.

These recommendations endorse those by previous authors (e.g. Austin, Manca, Zwarenstein, Juurlink, & Stanbrook, 2010; De Boer et al., 2015; Roberts & Torgerson, 1999; Senn, 1994). In following these recommendations, researchers can increase the statistical power to detect an experimental effect in RCTs, and in a non-optimal world of NHSTs this could potentially change a dichotomous significance conclusion. But, the size of experimental effects can (and should) be interpreted independently from any covariates in RCTs, using effect size indices, foremost those that are insensitive to error variance magnitude (e.g. eta-squared, though not a partial eta-squared[3]). For properly powered RCTs, interpretation of such effect size indices is not affected by the inclusion of a priori selected covariates (i.e. not beyond inconsequential changes in the point estimate and corresponding confidence intervals). Therefore, when discussion of results shifts the focus to such indices instead of significance, this arguably renders the use of ANCOVA models in RCTs of little value altogether.

Acknowledgements

For their valuable input and comments, I am thankful to Gjalt-Jorn Peters (Open University), Ilja Croijmans (Radboud University), Bram Fleuren and Niek Zonnebeld (Maastricht University).

3 I'm not making the argument that a non-partial eta-squared is statistically superior to the partial variant. Instead, given that partial eta-squared is more sensitive to the inclusion of covariates (because it is based on the condition versus error sum of squares ratio; an error term that decreases with every covariate added), this measure might be more rewarding to the practice of covariate fishing than the regular eta-squared.

References

Altman, D. G., & Doré, C. J. (1990). Randomisation and baseline comparisons in clinical trials. The Lancet, 335(8682), 149-153. doi:10.1016/0140-6736(90)90014-V

Austin, P. C., Manca, A., Zwarenstein, M., Juurlink, D. N., & Stanbrook, M. B. (2010). A substantial and confusing variation exists in handling of baseline covariates in randomized controlled trials: a review of trials published in leading medical journals. Journal of clinical epidemiology, 63(2), 142-153.

De Boer, M. R., Waterlander, W. E., Kuijper, L. D., Steenhuis, I. H., & Twisk, J. W. (2015). Testing for baseline differences in randomized controlled trials: an unhealthy research behavior that is hard to eradicate. International Journal of Behavioral Nutrition and Physical Activity, 12(1), 1-8. doi:10.1186/s12966-015-0162-z

Groenwold, R. H. H., Donders, A. R. T., Roes, K. C. B., Harrell, F. E., & Moons, K. G. M. (2012). Dealing With Missing Outcome Data in Randomized Trials and Observational Studies. American Journal of Epidemiology, 175(3), 210-217. doi:10.1093/aje/kwr302

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P. J., . . . Altman, D. G. (2010). CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. Journal of clinical epidemiology, 63(8), e1-e37. doi:10.1016/j.jclinepi.2010.03.004

Roberts, C., & Torgerson, D. J. (1999). Baseline imbalance in randomised controlled trials. BMJ : British Medical Journal, 319(7203), 185-185.

Senn, S. (1994). Testing for baseline balance in clinical trials. Statistics in Medicine, 13(17), 1715-1726. doi:10.1002/sim.4780131703

Van Breukelen, G. J. P. (2006). ANCOVA versus change from baseline: More power in randomized studies, more bias in nonrandomized studies. Journal of clinical epidemiology, 59(9), 920-925. doi:10.1016/j.jclinepi.2006.02.007

Van Breukelen, G. J. P., & Van Dijk, K. R. A. (2007). Use of covariates in randomized controlled trials. Journal of the International Neuropsychological Society, 13(5), 903-904.

Van Breukelen, G. J. P. (2013). ANCOVA Versus CHANGE From Baseline in Nonrandomized Studies: The Difference. Multivariate Behavioral Research, 48(6), 895-922. doi:10.1080/00273171.2013.831743

Stefan L. K. Gruijters
Department of Work and Social Psychology, Maastricht University
Stefan.Gruijters@maastrichtuniversity.nl