Statistical issues in randomised controlled trials: a narrative synthesis

Bolaji Emmanuel Egbewale*

Department of Community Medicine, Ladoke Akintola University of Technology, Ogbomoso, Nigeria

ARTICLE INFO

ABSTRACT

Randomised controlled trials (RCTs) are gold standard in the evaluation of treatment efficacy in medical investigations, only if well designed and implemented. Till date, distorted views and misapplications of statistical procedures involved in RCTs are still in practice. Hence, clarification of concepts and acceptable practices related to certain statistical issues involved in the design, conduct and reporting of randomised controlled trials is needed. This narrative synthesis aimed at providing succinct but clear information on the concepts and practices of selected statistical issues in RCTs to inform correct applications. The use of tests of significance is no longer acceptable as means to compare baseline similarity between treatment groups and in determining which covariate(s) should be included in the model for adjustment. Distribution of baseline attributes simply presented in tabular form is however, rather preferred. Regarding covariate selection, such approach that makes use of information on the degree of correlation between the covariate(s) and the outcome variable is more in tandem with statistical principle(s) than that based on tests of significance. Stratification and minimisation are not alternatives to covariate adjusted analysis; in fact they establish the need for one. Intention-to-treat is the preferred approach for the evaluation of primary outcome measures and researchers have responsibility to report whether or not the procedure was followed. A major use of results from subgroup analysis is to generate hypothesis for future clinical trials. Since RCTs are gold standard in the comparison of medical interventions, researchers cannot afford the practices of distorted allocation or statistical procedures in this all important experimental design method.

1. Introduction

Randomised controlled trials (RCTs), also referred to as randomised clinical trials, have been described as the gold standard and widely accepted as the best trial design for comparing two or more medical therapies or health care interventions[1,2]. This claim is however correct only if the trial is appropriately designed, conducted and reported[3]. RCTs offer solutions to some of the issues that have been raised against observational studies. For example, as earlier observed[4], treatment differences identified from observational designs, rather than from experimental clinical

*Corresponding author: Dr Bolaji Emmanuel Egbewale, Department of Community Medicine, College of Health Sciences, Ladoke Akintola University of Technology, Ogbomoso, Nigeria.
 Tel: +2348132797209
 E-mail: begbewale@gmail.com

trials are subject to methodological weaknesses, including selection bias, confounding and cohort effects-variations in characteristics of an area of study over time among individuals who are defined by some shared temporal experience, such as year of birth. Previous studies have reckoned, however, that a well conducted observational study can be more valuable than RCTs with distorted randomisation, as statistical considerations and overall interpretations usually take bias in non-experimental studies into account[5,6].

Furthermore, distorted view in RCTs is not limited to allocation of participants; it is also a thing of concern in certain statistical procedures involved in the delivery of a meaningful RCT. Since the authenticity of RCTs depends largely on the correctness of the design, conduct and reporting, it is important to make clear concepts and current trends of procedures involved, especially those that have considerable statistical notions. For example, the way and manner of comparing treatment groups for their similarities in prognostic factors at baseline differ among clinical researchers. While some adopt a practice of using tests of significance by using *P*-values to measure baseline comparability between groups or justify their choice of covariates to adjust in the mainstream analysis, others renounce such practice and regard it as unnecessary. The aim of this paper is to clarify issues and present clear information regarding certain statistical concepts and procedures that are necessary for the delivery of meaningful RCTs. Previous studies have focused on selected aspects of the design of RCTs[7,8]. The simplest and perhaps the most popular type of clinical trial is the two-arm parallel design, in which the study participants on recruitment to the trial are randomised to either of the two treatment groups[1,9-12].

2. Method

This is a narrative synthesis of concepts and acceptable practices of selected statistical issues in RCTs. Searches for literatures were conducted electronically and manually where necessary. In total, 44 literatures including articles and books on the subjects were accessed from various sources: library text materials, PubMed bibliographic database and literature suggestions from friends and colleagues. Five statistical issues, namely, baseline comparability, selection of covariate for adjustment, covariate adjustment, intention-to-treat analysis (ITT) and subgroup analysis were considered in this paper.

3. Baseline comparability

This is a single concept that has generated much controversy among trialists, statisticians, and clinical investigators who have

the responsibility of measuring and interpreting treatment effects. Various authors have remarked that randomisation guarantees unbiased allocation of treatments to study participants and does not ensure for a particular trial that the patients or study participants in each treatment group will have similar characteristics[9,13]. This then suggests that randomisation, at best, secures unbiased treatment allocation, but not necessary balance. The view was shared by other researchers who also noted that the procedure provides foundation for statistical tests in practice[11]. Since in practice following randomisation, some important covariates may not be balanced between treatment groups especially when the sample size is small; it is therefore a usual practice in clinical trial experiments to present baseline information on prognostic factors[14-16]. The first part of the result section is often devoted to the tabular presentation of the baseline characteristics of the study participants. This practice allows for quick judgement of the success or otherwise of the randomisation procedure, and as a result, provides basic information on which confidence on subsequent treatment comparison hinges.

On the other hand, tests of significance that utilize *P*-value to determine the statistical significance of the observed baseline difference in patients' characteristics are also being adopted. However, this practice has suffered wide criticism, and has been regarded as unnecessary. The consensus regarding baseline comparison of patients' characteristics appears to be that researchers should present the distributions of such baseline information of treatment groups in a table, thus, allowing readers to see the extent of similarities of the groups[17,18]. Furthermore, this practice allows physicians to infer the results to particular patients[19]. Many authors disapprove of the use of hypothesis tests as means of comparing baseline characteristics across groups[20-23]. They contest the practice whereby tests of significance used to assess comparability in respect of the magnitude of baseline imbalance. Their argument is that there is no need for such tests, as a proper randomisation procedure ensures that groups' difference is entirely due to chance, and all such tests seek to establish is that the observed difference could or could not have been due to chance. They also argue that researchers who use hypothesis tests to compare baseline characteristics report fewer significant results than expected by chance, thus suspecting a foul play in reporting. The procedure of hypothesis tests on baseline characteristics has been described as not only clearly absurd but also as unnecessary and might also be harmful[16,20,22].

4. Selection of covariates for adjustment

Irrespective of the method adopted at the design stage to bring about balance at baseline, the view shared by most clinical

investigators (especially statisticians) is to further account for baseline imbalance by applying relevant statistical method. However, a major bone of contention is which covariate(s) to include in the statistical adjustment model. There are basically three different views on this issue of covariate selection for statistical adjustment. Perhaps the most popular one of these is the use of baseline tests of significance to determine which covariate to include in the model for adjustment. In this case, study groups are compared in a wide range of baseline variables; those with statistically significant difference between groups are automatically accounted for in the analysis, and those that are not significant are ignored[24]. It was observed by previous researchers that about 50% of clinical trial experiments published in four leading medical journals adopted this method[25]. However, this idea has suffered major set-back over the years, and its use has been greatly discouraged by methodologists[26].

The basic argument against the afore-mentioned approach is that, since study participants are randomly allocated to treatment groups in the first instance, then, any observed difference must have been due to chance. It then appears absurd to again test whether the observed difference is purely by chance or not, which is what the test of significance does. In addition, ignoring baseline covariates that have prognostic influence but not significantly different between groups is also an argument against the correctness of the use of hypothesis testing approach for covariate selection[20-23,25,27,28]. In fact, as variously submitted, a significant imbalance will not matter if a factor does not predict outcome; whereas, a non-significant imbalance can benefit from covariate adjustment[25,28].

The second view reflects the importance attached to the prognostic strength of covariates. However, there are two variants of this idea; the first bases the covariate prognostic importance on the level of correlation between the particular covariate and the outcome variable. The usual practice here is that, if there is a weak correlation, say, $r < 0.3$, adjusting for the imbalance in such a covariate is not necessary even with a significant baseline difference in the covariate between the treatment groups. This appears to support the idea that non-significance does not matter if the covariate-outcome correlation is strong. Just like significance testing, examining strength of relationship between the baseline and the outcome variables is a data-driven procedure. Analysts should examine the correlation between the covariates and the outcome of interest before deciding on selection of such covariates for adjustment. A classic example is a trial of primary biliary cirrhosis that had a non-significant baseline imbalance in a strong prognostic variable, serum bilirubin unadjusted, and adjusted analyses yielded $P = 0.1$ and $P = 0.02$, respectively, for the treatment differences in survival[29]. This example touches on the importance of recognising the prognostic strength of baseline variables rather than the statistical significance of imbalances.

The third known principle that guide covariate selection for the purpose of adjustment of baseline imbalance appears to be a variant of the second with selection being on the basis of covariates that have been found a priori to be prognostic in relation to the outcome variable. This includes evidence of suitable covariate-outcome correlation ($r \geqslant 0.3$) from previous research or pilot studies[16,27,28]. The decision on which covariate is selected for adjustment is taken before the trial starts and usually specified in the protocol. This agrees with the recommendation of the International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for human use guideline[11]. The idea of using covariates identified a priori would also cover statistically adjusting for stratification or minimisation factors.

5. Covariate adjustment

In practice, simple randomisation may not ensure balance in some important covariates. If any unbalanced covariates are strongly correlated with the study outcomes, their presence may make it difficult to interpret the results of statistical tests for the treatment effect[11]. Thus, it is important that such imbalances are corrected or adjusted. Other studies have recorded a beneficial effect of covariate adjustment over the unadjusted even for covariate with moderate correlation with the outcome variable[13,30]. The procedures for controlling the covariate imbalance can either be at the design stage or during statistical analysis; adjustment at the design stage includes the use of such techniques as minimisation and stratification. The procedure for adjustment during statistical analysis accounts for covariate imbalance at the analysis stage by using relevant statistical method appropriate for the purpose.

In the case of a single post treatment assessment of a continuous outcome variable, methods for adjustment at the statistical analysis stage are: change score analysis that determines group effect based on the difference between the baseline and the post treatment score (basic adjustment) and analysis of covariance, which is a model-based adjustment that includes the baseline of the outcome variable in the model. Statistical adjustment can also be performed by the use of logistic regression or by pooling the stratified analyses, using, for example, a Mantel Haenszel test. In many clinical trials, both design methods that reduce covariates imbalance and statistical adjustment during analysis are used simultaneously. Previous researchers have observed that for a given set of covariates, even though the stratification or minimisation methods will make the treatment groups comparable in these variables[15], they do not completely remove the effect of imbalance. As a result, the stratification or minimisation factors need to be incorporated in the model for adjustment. Statistical adjustment can have a profound influence on effect estimates and tests of significance. For example, covariate-

adjusted estimates are not only more precise than the unadjusted, but making the odds ratio or hazard ratio for logistic regression analyses and hazards models become further away from the null. Statistical adjustment for strong predictors of outcome achieves more valid treatment effect estimates and significance tests[25].

In addition, with respect to chance imbalance between treatment groups in a baseline covariate (especially when the baseline covariates is strongly correlated with the outcome), an adjusted estimate of the treatment effect accounts for this observed imbalance while an unadjusted analyses does not[25,31]. A further benefit of the covariate adjusted analysis can be the creation of a predictive model which combines the influences of treatment and prognostic covariates in estimating the expected outcome for individual patients[31]. Moreover, the validity of an unadjusted analysis relies on the assumption that there are no important imbalances involving measured and unmeasured baseline covariates across treatment groups. When imbalances occur on measured predictors of outcome variables, adjusted analyses should be performed[32]. It should be added that even if the groups have similar characteristics, it might still be desirable to adjust for another variable if we know in advance that the variable is strongly related to prognosis. Although, a primary reason for adjustment for imbalance in one or more covariates is the removal of chance bias, adjusting for a prognostic variable may also lead to greater power of the trial[9,28,33].

6. ITT analysis

ITT analysis is the strategy for the analysis of RCTs that compares patients in terms of the groups to which they were originally randomly assigned[34]. This implies that patients are always analysed in the group to which they were initially randomised even if they drop out of the study[34-36]. This principle is fundamental to the experimental nature of RCTs as it ensures that the ideal structure for comparison created by random assignment of participants into treatment group is not distorted. There is wide agreement that the most appropriate analysis set for the primary effectiveness analyses of any confirmatory (phase III) clinical trial is the ITT; this form of analysis assesses the overall clinical effectiveness most relevant to the real-life use of the therapy[12].

It is a recommendation of the Consolidated Standard of Reporting Trials that authors should indicate whether analyses were performed on an ITT basis[37]. The only safe way to deal with all forms of protocol violation is to apply ITT; included here are patients who actually receive a treatment other than the one allocated, and patients who do not take their treatment (known as non-compliers). However, whether ITT principle also applies if it is discovered after the trial has begun that a patient was not eligible for the trial is opened to debate.

A different analysis strategy commonly used (as a secondary evaluation) is to exclude patients who have not adhered to the allocated management strategy for whatever reason. This form of analysis is called per-protocol analysis, efficacy analysis, explanatory analysis, or analysis by treatment administered. It only describes the outcomes of the participants who adhered to the research protocol. Previous study observe that per-protocol analysis becomes a problem especially when the reasons for non-adherence to the protocol are related to prognosis[38]. Empirical evidence suggests that participants who adhere tend to do better than those who do not adhere, even after adjustment for all known prognostic factors and irrespective of assignment to active treatment or placebo[39]. Thus, excluding non-compliers participants from the analysis leaves those who are destined to have a better outcome and destroys the unbiased comparison afforded by randomization. However, a relationship between a higher methodological quality of the trials and the reporting of the ITT has also been established[40].

7. Subgroup analysis

One of the reasons for collecting substantial baseline data from patients in a RCT is that subgroup analyses (treatment outcome comparisons for patients subdivided by baseline strata) may be carried out[25]. This is to assess whether treatment difference in outcome or lack of it depends on certain characteristics of patients. The results from such group-specific assessment can be used to generate hypothesis for future study[25,33]. Subgroup analyses are important if there are potentially large differences between stratified groups in the risk of a poor outcome with or without treatment, if there are practical questions about when to treat, or if there are doubts about benefit in specific groups such as elderly people which are leading to potentially inappropriate over- or under-treatment[41]. Since patients recruited into a clinical trial are not a homogeneous sample, their response to treatment and the differing impact on them of different treatments may well vary in ways that affect the choice of which treatment is best for which patients. It was argued that if in truth, there are specific subgroups of patients for which a new treatment is more or less effective or harmful than that is indicated by the overall comparison with standard treatment in the trial as a whole, there is a scientific and ethical obligation to try and identify such subgroups[31].

However, most trials only have sufficient statistical power to detect the overall main effect difference in response between treatment groups, so that if subgroup effects do exist, they may well go undetected because the trial was not large enough[8,31]. Smaller sample sizes within subgroups lead to greater standard errors and reduced power relative to the overall clinical trial, resulting in an increased risk of a false-negative result; whereas, the multiplicity of hypotheses tests that results from examining multiple subgroups

will lead to an increased risk of a false positive result: inflation of type I error[9]. The author further reckoned that to look for effects in subgroups is never a good way to rescue a study in which the primary ITT analysis fails to show an overall effect. The suggested approach to a sub-group analysis is to compare the difference between the treatments for the sub-groups of interest. The interaction can be examined within an appropriate multiple regression model, whether the outcome is continuous, binary or survival time.

The unadjusted strategy yields an average treatment effect without any consideration for heterogeneity in prognosis among patients. Although both covariate adjustment and subgroup analyses consider heterogeneity and attempt to provide more individualized estimates of treatment effect, they are substantially different[42]. The difference is that, while covariate adjustment obtains a single more individualised treatment effect estimate, which is assumed to be applicable to all patients[31,43], subgroup analyses provide multiple treatment effect estimates, assuming that treatment effects differ between particular groups of patients[44]. For the reason aforementioned, though subgroup analyses are sometimes performed, they rarely have enough power to detect differential treatment effect. It has however been variously observed that tests of interaction are underused and subgroup analyses are commonly over-interpreted[9,43]. Researchers should therefore be wary of this.

8. Conclusions

In trial settings, it is no longer popular to use test of significance to assess baseline comparability of characteristics of study participants between treatment groups. It is sufficient to present distributions of baseline characteristics of treatment groups in a table, by so doing, readers are allowed to see the extent of similarities in the groups. Similarly, the practice of using tests of significant or *P*-value to determine which covariate(s) qualify for statistical adjustment is prone to yielding misleading results and have also been discouraged. A more rather acceptable approach will make use of the information on the extent of correlation between the covariate and the outcome variable. Usually, a covariate may qualify for statistical adjustment if its correlation with the outcome variable is greater than or equal to 0.3.

Furthermore, it should be noted that no design methods completely remove the effect of imbalance, therefore, stratification or minimisation factors should always be considered for inclusion in the statistical model for adjustment. Whenever protocol violation is observed, the only safe statistical approach that will help in dealing with it is ITT. By so doing, the experimental nature of RCTs which allows unbiased comparison of groups' treatment effect is preserved. Researchers are required to report whether primary

analyses were performed on an ITT basis or not. Lastly, subgroup analysis is essential especially when there is a huge difference in the observed treatment effect within identified groups in the trial participants. However, since the trial is only powered to detect a main effect, even if subgroup effect exists, they may as well pass unnoticed. A major use of results from subgroup analysis is to generate hypothesis for future clinical trials. Since RCTs are gold standard in the comparison of medical interventions, researchers cannot afford distorted allocation or statistical procedures in this all important experimental design method.

Conflict of interest statement

I declare that I have no conflict of interest.

Acknowledgements

References

[1] McLeod RS, Wright JG, Solomon MJ, Hu X, Walters BC, Lossing AI. Randomised controlled trials in surgery: issues and problems. *Surgery* 1996; **119**(5): 483-6.

[2] Grimes DA, Schulz KF. An overview of clinical research: the lay of the land. *Lancet* 2002; **359**(9300): 57-61.

[3] Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *Ann Intern Med* 2010; **152**: 726-32.

[4] Kang M, Ragan BG, Park JH. Issues in outcomes research: an overview of randomization techniques for clinical trials. *J Athl Train* 2008; **43**(2): 215-21.

[5] Torgerson DJ, Roberts C. Understanding controlled trials. Randomisation methods: concealment. *BMJ* 1999; **319**(7206): 375-6.

[6] Cotton PB. Randomization is not the (only) answer: a plea for structured objective evaluation of endoscopic therapy. *Endoscopy* 2000; **32**(5): 402-5.

[7] Field D, Elbourne D. The randomised controlled trial. *Curr Paediatr* 2003; **13**: 53-7.

[8] Egbewale BE. Random allocation in controlled clinical trials: a review. *J Pharm Pharm Sci* 2014; **17**(2): 248-53.

[9] Altman DG. *Practical statistics for medical research*. London:

Chapman and Hall; 1991.

[10] Overall JE, Doyle SR. Implications of chance baseline differences in repeated measurement designs. *J Biopharm Stat* 1994; **4**(2): 199-216.

[11] Tu DS, Shalay K, Pater J. Adjustment of treatment effect for covariates in clinical trials: statistical and regulatory issues. *Ther Innov Regul Sci* 2000; **34**: 511-23.

[12] Cook TH, DeMets DL. Introduction to statistical methods for clinical trials; Madison: Chapman & Hall/CRC; 2008.

[13] Senn SJ. The use of baselines in clinical trials of bronchodilators. *Stat Med* 1989; **8**(11): 1339-50.

[14] Begg CB. Suspended judgment. Significance tests of covariate imbalance in clinical trials. *Control Clin Trials* 1990; **11**(4): 223-5.

[15] Raab GM, Day S, Sales J. How to select covariates to include in the analysis of a clinical trial. *Control Clin Trials* 2000; **21**(4): 330-42.

[16] Altman DG. Comparability of randomised groups. *Statistician* 1985; **34**: 125-36.

[17] Lachin JM. Statistical properties of randomisation in clinical trials. *Control Clin Trials* 1988; **9**(4): 289-311.

[18] Burgess DC, Gebski VJ, Keech AC. Baseline data in clinical trials. *Med J Aust* 2003; **179**(2): 105-7.

[19] Pocock S. Statistical aspect of clinical trial design. *Statistician* 1982; **31**: 1-18.

[20] Schulz KF, Chalmers I, Grimes DA, Altman DG. Assessing the quality of randomization from reports of controlled trials published in obstetrics and gynecology journals. *JAMA* 1994; **272**(2): 125-8.

[21] Schulz KF. Subverting randomisation in controlled trials. *JAMA* 1995; **274**(18): 1456-8.

[22] Senn SJ. *Statistical issues in drug development*. Chichester: John Wiley and sons; 1997.

[23] Matthews JNS. *An introduction to randomised controlled clinical trials*. London: Arnold; 2000.

[24] Meinert CL. *Clinical trials: design, conduct, and analysis*. New York: Oxford University Press; 1986.

[25] Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet* 2000; **355**(9209): 1064-9.

[26] Fayers PM, King MT. In reply to Berger "don't test for baseline imbalances unless they are known to be present?" *Qual Life Res* 2000; **18**(4): 401-402.

[27] Senn S. Testing for baseline balance in clinical trials. *Stat Med* 1994; **13**: 1715-26.

[28] Egbewale BE, Lewis M, Sim J. Bias, precision and statistical power of analysis of covariance in the analysis of randomized trials with baseline imbalance: a simulation study. *BMC Med Res Methodol* 2014; **14**: 49.

[29] Christensen E, Neuberger J, Crowe J, Altman DG, Popper H, Portmann B, et al. Beneficial effect of azathioprine and prediction of prognosis in primary biliary cirrhosis. Final results of an international trial. *Gastroenterology* 1985; **89**(5): 1084-91.

[30] Canner PL. Covariate adjustment of treatment effects in clinical trials. *Control Clin Trials* 1991; **12**(3): 359-66.

[31] Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Stat Med* 2002; **21**(19): 2917-30.

[32] The European Agency for the Evaluation of Medicinal Products. Points to consider on adjustment for baseline covariates. London: The European Agency for the Evaluation of Medicinal Products; 2003. [Online] Available from: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003639.pdf [Accessed on 12nd August, 2014]

[33] Wang D, Bakhai A. *Clinical trials: a practical guide to design, analysis and reporting*. London: Remedica Medical Education and Publishing; 2006.

[34] Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. *BMJ* 1999; **319**(7211): 670-4.

[35] Roland M, Torgerson DJ. What are pragmatic trials? *BMJ* 1988; **316**(7127): 285.

[36] Wright CC, Sim J. Intention-to-treat approach to data from randomized controlled trials: a sensitivity analysis. *J Clin Epidemiol* 2003; **56**(9): 833-42.

[37] Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *JAMA* 1996; **276**(8): 637-9.

[38] Montori VM, Guyatt GH. Intention-to-treat principle. *CMAJ* 2001; **165**(10): 1339-41.

[39] Horwitz RI, Viscoli CM, Berkman L, Donaldson RM, Horwitz SM, Murray CJ, et al. Treatment adherence and risk of death after a myocardial infarction. *Lancet* 1990; **336**(8714): 542-5.

[40] Ruiz-Canela M, Martínez-González MA, de Irala-Estévez J. Intention to treat analysis is related to methodological quality. *BMJ* 2000; **320**: 1007-8.

[41] Rothwell PM. Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet* 2005; **365**(9454): 176-86.

[42] Hernandez AV, Steyerberg EW, Habbema JD. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. *J Clin Epidemiol* 2004; **57**(5): 454-60.

[43] Robinson LD, Jewell NP. Some surprising results about covariates adjustment in logistic regression models. *Int Stat Rev* 1991; **58**: 227-40.

[44] Parker AB, Naylor CD. Subgroups, treatment effects, and baseline risks: some lessons from major cardiovascular trials. *Am Heart J* 2000; **139**(6): 952-61.