Global Journal of Health Science; Vol. 14, No. 7; 2022
ISSN 1916-9736 E-ISSN 1916-9744
Published by Canadian Center of Science and Education

Cumulative Equivalence: Controlling for Inter-Individual Differences at Baseline Characteristic Testing of RCTs

John Damiao[1]

[1] College of Health Professions, Pace University, Pleasantville, NY, USA

Correspondence: John Damiao, College of Health Professions, Pace University, Pleasantville, NY, USA. Tel: 914-773-3199. E-mail: jdamiao@pace.edu

Received: March 14, 2022 Accepted: June 6, 2022 Online Published: June 15, 2022

doi:10.5539/gjhs.v14n7p32 URL: https://doi.org/10.5539/gjhs.v14n7p32

Abstract

Randomized control trials (RCTs) are regarded as the gold standard for intervention research. The randomization process is intended to establish comparability between groups, so that the study outcomes can be attributable to the intervention, rather than group differences. The purpose of this paper is to emphasize the inherent risks of conducting multiple tests in the establishment of equivalency at baseline while omitting the cumulative effect of small group differences in RCTs. Randomization does not thoroughly prevent differences in group averages at the specific characteristic level. Any baseline differences that benefit the intervention group when accumulated over multiple categories of demographic characteristics described herein as cumulative inequivalence can significantly impact the internal validity of RCTs. This paper describes a procedure for assessing for cumulative inequivalence, as well as procedures such as re-randomization prior to intervention to establish comparability and thus promote cumulative equivalence of RCTs.

Keywords: Randomized control trials, cumulative inequivalence, baseline testing, comparability

1. Introduction

Randomized control trials (RCTs) are commonly regarded as the gold standard for health intervention research studies (Victoria, et al. 2004). The process for random assignment of study participants is used to assure both groups are equal, thus reducing study outcomes attributable only to the intervention (Cartwright, 2007; Lim & In, 2019; Saint-Mont, 2015; Sidani, 2015). Randomization can take many forms, as described by Lim & In (2019). *Simple randomization* describes a process in which a roll of a die, coin flip or random-number generator are used to assign participants to groups. *Block randomization* is the process by which participants are assigned in blocks such as to assure that the variability within blocks is less than the variability between blocks, such as gender. *Unbalanced randomization* is applied for ethical or cost reasons, or when high attrition rates can be expected for a certain study. Similar to block randomization, *stratified randomization* can be applied when participants present with covariates that could affect the study outcome or treatment. Lastly, *adaptive randomization* is a method of guiding the allocation process to minimize an imbalance between groups.

Commonly, researchers demonstrate this homogeneity via demographic calculations consisting of group means or frequency comparisons, which are typically emphasized in a table format. However, this procedure typically only occurs at the individual variable level and are likely to result in researchers making erroneous assumptions of equivalency. Incorporating the commonly applied alpha (α) = .05 at the post-intervention phase may help to protect against type I errors; but this procedure is ineffective at protecting against group differences at the pre-intervention phase. In other words, α = .05 at the post-intervention phase makes sense as a threshold for reaching statistical significance, as it suggests that the probability of making a type I error is only 5%. We want to see a large enough group difference at the post-intervention phase that we can confidently agree that the group difference is large enough to be attributed to the intervention. However, accepting this large of a group difference at the pre-intervention as a measure of group equivalence is obviously illogical, as it is too low of a threshold for comparability. Thus, incorporating an α = .10 to increase the confidence of equivalency appears much more reasonable. Furthermore, failure to account for a cumulative group difference over multiple variables of interest may impact baseline equivalency and the integrity of the RCT outcomes. In other words, while randomization was originally thought to relieve the experimenter from the anxiety of determining how data might be impacted by group differences; randomization alone may be responsible for providing researchers a false sense of security

(Saint-Mont, 2015).

The methodological rigor of the RCT is largely dependent on the strength and validity of the establishment of both the control and intervention groups as equivalent. The purpose of this paper is to emphasize the inherent risks of conducting multiple tests in the establishment of such equivalency while omitting the cumulative effect of small group differences. A method for assessment and correction for cumulative inequivalence to further strengthen the internal validity of RCTs is described.

2. The Problem of Cumulative Inequivalence

Between-group baseline comparisons are intended to hold constant the variables that may influence the study outcomes (Sidani, 2015. p57), however, Cartwright (2007) argues caution in assuming this process results in equivalency between groups relative to the construct of interest. Similarly, Victoria and colleagues (2004) regard RCTs as the gold standard for intervention research, but also caution the potential for selection bias, and potential for confounding factors that may go unrecognized.

The RCT as a methodological design is recognized for its tight control of internal validity based on the assumption that randomly assigned participants result in equal groups. Randomization, however, is unable to control for inter-individual differences, as baseline measures are typically analyzed and reported in terms of group averages (Shadish & Ragesdale, 1996; Heinsman & Shadish, 1996). Furthermore, randomization does not thoroughly prevent differences in group averages at the specific characteristic level (Lim & In, 2019). This would be permissible if both groups have a fairly even share of benefits among several categories, but if one group experiences a significant benefit in most categories, these advantages can compound to create essentially unequal groups.

For example, a small decrease in the age, body-mass index, and cholesterol parameters of the intervention group may not be statistically significant at the individual characteristic level; but when compounded over all categories, can provide this group with a significant advantage over the control. Thus, there is an inherent risk for cumulative inequivalence at the random assignment phase, well before the study even begins.

2.1 Assessing Cumulative Inequivalence

When researchers conduct multiple tests, they increase the risk of inflating the type I error. It is common practice to apply a statistical correction to create a more conservative threshold for determining statistical significance; however, similar statistical corrections that are expected of outcome analysis are not typically applied to the pre intervention equivalency comparison phase. Hence, the same concerns applied to the inflation of type I errors should be applied to the assumptions of successful equivalence at randomization, albeit through a different statistical procedure.

When analysis requires multiple tests to be run, researchers are met with a statistical dilemma. Multiple calculations of a statistical procedures (i.e., multiple t-test) comes at the theoretical cost of inflating the chance of a type I error. It is common practice for researchers to apply a Bonferroni correction in these instances to create a more conservative threshold for determining statistical significance. This is accomplished by dividing the commonly used $\alpha = .05$ by the number of tests being conducted. In other words, if a researcher conducts three t-tests, each t-test should result in $p \leq .017$ ($.05 \div 3$).

Theoretically, a Bonferroni correction applied to the characteristic analysis phase in the same way would be an incorrect procedure as it would further inflate the erroneous assumption of equivalence, as the goal is to stay above the $p = .05$ level. For example, it is naïve to assume similarity between two groups when the mean age results in a $p = .08$. Traditional hypothesis testing tells us that there is an 8% chance of differences occurring just by chance. However, this also means that if we were to assume these groups were inherently different, we would *only* be wrong 8% of the time. It should be apparent to anyone conducting RCTs that this probability for assuming equivalence between groups should be unacceptable. The inverse correction consisting of multiplying the α by the number of tests as is done with the Bonferroni method appears feasible for comparisons when few categories are involved (<10). However, for comparisons with many categories, this procedure would make demonstrating such equivalence a procedural challenge and thus discouraging a thorough evaluation of pre-intervention equivalence.

The present author proposes a frequency assessment using a chi square (goodness of fit) analysis in which the intervention and control groups must be relatively equal in terms of attributes that would benefit either group. Due to the typically small number of categories analyzed in baseline testing, $\alpha = .10$ is a more suitable threshold for detecting inequivalence.

For example, a baseline demographic characteristic analysis may suggest equivalency of groups in terms of age,

income, presence of disease, caregiver support, etc.; but they are not likely to be identical either. A chi square should be conducted to assure that neither group has a statistical benefit among all categories combined. See Table 1 for a comparison of equivalence and inequivalence.

Table 1. Example Comparison of Demographic Equivalence/Inequivalence at Baseline

Characteristic	1st randomization			2nd randomization		
	tx (n = 50)	control (n = 50)	x^2, p	tx (n = 50)	control (n = 50)	x^2, p
Age	**62.1 (9.8)**	67.9 (9.9)	≥ .05	68.2 (9.7)	**67.8 (10)**	≥ .05
Years since stroke M (SD)	**1.9 (2.5)**	2.9 (2.8)	≥ .05	**2.1 (2.4)**	2.8 (2.7)	≥ .05
Level of impairment M (SD)	**7.8 (4.5)**	8.2 (5.1)	≥ .05	**7.6 (4.6)**	8.0 (4.9)	≥ .05
Baseline LDL ≥ 190	37	**36**	≥ .05	39	**34**	≥ .05
Health condition Poor	**19**	21	≥ .05	**18**	22	≥ .05
Cardiovascular disease	36	**34**	≥ .05	37	**33**	≥ .05
Stress	**19**	22	≥ .05	**17**	24	≥ .05
Income ≥ 50 000	**17**	15	≥ .05	**18**	14	≥ .05
Tx Adherence	**44**	39	≥ .05	38	**42**	≥ .05
Caregiver support	**32**	33	≥ .05	**36**	30	≥ .05
Total Cumulative Categories with Positive attributes, %	**80**	20	.058*	60	40	.527

Note. * denotes statistical significance. x^2 (chi square) is calculated for each category, with no individual item suggesting significant differences between group when α = .05. However, the total cumulative comparison suggests inequivalence between tx (treatment) and control groups (p = .058) at α = 0.10 in the 1st randomization attempt, and no cumulative difference at the 2nd randomization (p = .527).

3. Correcting for Cumulative Inequivalence

In the unfortunate instance that cumulative inequivalence appears present, the researchers may correct this issue by randomly re-assigning participants to the control and intervention groups, followed by a reassessment of individual and cumulative inequivalence through the procedure described above. Similar to the adaptive randomization process (Lim & In, 2019), an active approach may be needed to enhance or ensure true randomization. However, here the key to correcting for cumulative inequivalence is simply to assess and re-assign participants (if needed) prior to application of the intervention. In other words, researchers must gather participant information, simulate random assignment, and assess equivalency prior to commencement of the study as per the following recommended procedures:

- Using content expertise and knowledge from existing literature, determine which baseline characteristics have the potential to bias study outcomes in terms of group difference.
- Gather baseline characteristic data from all participants prior to group assignment.
- Simulate random assignment of participants to groups.
- Compile baseline data per group as in Table 1.
- Calculate a between groups frequency assessment using a chi square (goodness of fit) analysis to assure a roughly equal distribution of 'beneficial' characteristics (The present author recommends an alpha of .10, rather than the typical .05).
- If baseline characteristics are not evenly distributed ($p < .10$), then proceed to simulate a random re-assignment of participants to groups until an even distribution of 'beneficial' characteristics among the groups has been achieved ($p > .10$).
- When a satisfactorily even distribution of 'beneficial' characteristics has been achieved in this simulation, participants may be formally assigned to these groups, and study procedures can commence.

4. Conclusion

While RCTs are the gold standard for intervention research, this randomization process does not guarantee comparability between groups at the assignment phase. An unequal distribution of 'beneficial' baseline characteristics could occur, creating a cumulative effect, thus biasing the effect of the study outcomes. Assessing and correcting for cumulative inequivalence requires additional statistical analysis at the pre-intervention level, which may be cumbersome for some research studies. However, this approach allows for a relatively simple and logical process for assuring an effective random assignment process, which ultimately increases the internal validity and rigor of RCTs.

Competing Interests Statement

The authors declare that there are no competing or potential conflicts of interest.

References

Cartwright, N. (2007). Are RCTs the gold standard? *BioSocieties, 2*(1), 11-20. https://doi.org/10.1017/S1745855207005029

Kaul, S., & Diamond, G. A. (2010). Trial and error: how to avoid commonly encountered limitations of published clinical trials. *Journal of the American College of Cardiology*, *55*(5), 415-427. https://doi.org/10.1016/j.jacc.2009.06.065

Lim, C. Y., & In, J. (2019). Randomization in clinical studies. *Korean journal of anesthesiology*, *72*(3), 221. https://doi.org/10.4097/kja.19049

Saint-Mont, U. (2015). Randomization does not help much, comparability does. *PloS one*, *10*(7), e0132102. https://doi.org/10.1371/journal.pone.0132102

Sidani, S. (2015). *Health Intervention Research* (P. 57). SAGE Publications LTD. https://doi.org/10.4135/9781473910140

Victora, C. G., Habicht, J. P., & Bryce, J. (2004). Evidence-based public health: moving beyond randomized trials. *American journal of public health*, *94*(3), 400-405. https://doi.org/10.2105/AJPH.94.3.400

Copyrights