‘Is it ‘random’ or ‘haphazard’? Demonstrating effects of non-random allocation by simulation

Penny S Reynolds[1],
[1]Department of Anesthesiology, Statistics in Anesthesiology Research (STAR)-Core, College of Medicine, University of Florida, 1600 SW Archer Rd., Gainesville FL 32610

Abstract
Randomization is frequently misunderstood or neglected by preclinical investigators. I used a typical data set for swine models of preclinical research to show how improper randomization of treatment allocation adversely affects hypothesis tests and the underlying null distributions of the test statistics. Simulations were used to examine effects of true randomization (completely randomized design, restricted randomization, randomized complete blocks) vs pseudo-randomization (alternation, false “blocking”) on error estimates and F-distributions in the presence of systematic trend. True randomization and blocking protected against systematic trend, but pseudo-randomization resulted in reference distribution collapse. Thus, no meaningful inferential test can be based on non-random ‘designs’. Both investigators and analysts must be made aware that hypothesis tests based on non-randomized data will be both biased and invalid.

Key Words: F-values, *p*-values, preclinical research, random allocation

1. Introduction

Randomization is the assignment of treatments or test interventions to subjects or experimental units such that every possible assignment of treatments has the same probability of occurrence. Randomization is essential for minimizing the effects of undetected bias[1-3] and is the basis for exact tests of significance and obtaining unbiased estimators of intervals and treatment effects; randomization is therefore the “cornerstone” of null hypothesis significance testing (NHST)[4,5]. To ensure equal probabilities of assignment, randomization methods require generation of a reproducible randomization schedule (the best methods are computer-based procedures with seed numbers); allocation bias is further minimised by blinding of investigators to sequence allocation. ‘Randomization’ thus has both technical and practical meaning.[1,6] Unfortunately, the concept of randomization appears to be widely misunderstood by non-statistician investigators, although the topic has been extensively addressed for decades [1,4,5,7-9]. Both clinical and preclinical investigators often fail in practice to discriminate a true random sequence from those that are ‘quasi-random’, ‘alternating’, ‘unplanned’, or ‘haphazard’ [10-12]. A more insidious problem is the uncritical analysis of data by statistical consultants and data analysts. Statisticians analysing data “sight unseen” may be unaware of methods by which data were sampled and collected, and thus fail to account for lack of randomization in the original design.

The objective of this study was to illustrate the consequences of non-random systematic allocation for NHST and inference in the presence of systematic trend, with specific application to preclinical animal-based research.

2. Simulations

2.1 Simulation rationale

Body weight is a major determinant of nearly all aspects of animal physiology and functional

morphology;[13] furthermore, many interventions and translational dose conversions are weight-based.[14] However, weight data with pronounced trend and heterogeneity are common in many animal studies. If animal growth rates are rapid, animals processed at different times may vary considerably in size even if baseline weights at colony entry were similar. For example, Yorkshire swine may gain 4-5 kg/week during the rapid growth phase; 10 week old animals weighing 30 kg may weigh ≥100 kg 8-10 weeks later. Therefore, to reduce selection bias and confounding, animal weights should be uniformly distributed across treatment groups through the use of specific design tools, such as matching, randomization, and blocking.[7]

Unfortunately, systematic trend may be a major confounding factor in many animal studies. For example, a recent survey of swine models of military-relevant therapeutics [11] indicated that more than half of surveyed studies showed a difference of 13-75 kg between the smallest and largest animals, but almost none reported appropriate mitigation measures. In addition, many studies claiming to be 'randomized' described allocation strategies that actually consisted of either alternation of treatments to sequential subjects, or alternatively, 'lumping' the same treatments into 'blocks', with sequential allocation of experimental interventions first, followed by sequential allocation of controls.[11] A number of systematic reviews indicate that inadequate or no randomization in preclinical studies is common. [8,15,16]

2.2. Simulation procedures

2.2.1. Data

Weight data typical of swine studies were generated from published growth charts for finisher pigs aged between 7 and 22 weeks and weighing from 18 to 110 kg (*http://www.hendersons.co.uk/pigequip/Pig_growth_rate.html*). The original 13 observations for body weights (W) were expanded to obtain a test population N = 100 by interpolating values from 18 to 100 kg (*proc expand*, SAS 9.4, SAS Inc., Cary NC), for a linear weight gradient of approximately 0.8 kg/day (r = 0.99; Figure 1).

2.2.2. Simulation conditions

This was designed as a 'uniformity trial' simulation for a hypothetical three-group trial. A uniformity trial is essentially a 'trial without treatments', and can be used to check for uncontrolled variation, heterogeneity assumptions, and performance of statistical inference methods and associated tests of significance[17,18]. Because true treatment effects are zero, any differences will be the results of variation in experimental units.

Samples of n = 36 were repeatedly drawn from the test population using simple random sampling without replacement (SAS *proc surveyselect*), and linear trend in W was maintained by sorting on ascending values of W.

Three treatments (A, B, C) were assigned to subjects in each replicate using three randomized, and two non-randomized, allocation scenarios. Randomized allocation scenarios were: (a) completely randomized (CR), with varying sample size imbalance; (b) 'restricted' randomized (RR), with randomization constrained to produce equal sample sizes per treatment arm (n_t = 12); (c) randomized complete block (RCB). Random treatment assignments were generated for the CRD scenario using the SAS macro *RandBetween (https://blogs.sas.com/content/iml/2015/10/05/random-integers-sas.html*), and for constrained randomization scenarios using the SAS *ranuni* call routine with fixed starting seeds. The two non-random assignments were (d) 'alternating' (ABC ABC); and (e) false 'blocking', where the same treatments were assigned consecutively (AAA, BBB, CCC) resulting in 'blocks' of the same treatment.

The statistical models used to assess treatment effects under the null hypothesis H_0: $\mu_1 = \mu_2 = \mu_3$ were based on one-way ANOVA $y_i = \mu + \tau_i + e_{ij}$, where $y_i = W_i$, τ_i are 'treatments' i = 1, 2, 3, treatment assignment 1:1:1; and $e_{ij} \sim N(0, \sigma_e^2)$ denotes the experimental unit error. I examined the effects of blocking (RCB) with the model was $y_{ij} = \mu + \beta_j + \tau_j + e_{ij}$, where τ_{ij} are treatments i = 1, 2, 3, and $e_{ij} \sim N(0, \sigma_e^2)$. Block size was either $n_b = 3$ (12 blocks) or 6 (6 blocks), with randomization within block; it was assumed blocking was performed on a variable unrelated to weight. Block effects β_j were modelled as random with $\beta_j \sim N(0, \sigma^2_b)$.

I performed hypothesis tests for treatment effect on each replicate with SAS *proc mixed* to obtain $F_{1-\alpha, v1, v2}$ (where $\alpha = 0.05$, $\nu_1 = 2$, and $\nu_2 = 33$ for fixed treatment effects [$\nu_2 = 28$ for RCBD $n_b = 6$], P-values, residual mean square error MSE, and treatment differences.. Empirical distributions were plotted for each allocation scenario; under the null hypothesis, F-values should approach the theoretical distribution for $F_{2,33}$ (or $F_{2,28}$ for RCBD), and p-values should be uniformly distributed on the interval [0, 1][19]. Estimates were summarised by the 5th, 50th, and 95th percentiles on the simulations.

2.2.3. Number of simulations

I performed 4000 simulations for each scenario. The number of replicates was based on assessment of persistent stability and bias of the cumulative type I error rate for a three-arm CR trial and $F_{2, 33}$. Stability was visually assessed by plotting the cumulative error rate vs. simulation replicates and determining the number of simulations for stable convergence at $\alpha = 0.05$ with bias range 0.045-0.055. (Figure 2).

3. Results

3.1. F distributions and p-values

Simulated distributions for randomized allocation strategies (CRD, RR, RCBD) approximated the expected F-distributions, and p-values were uniformly distributed (Figure 3, Table 1). In contrast, systematic non-randomized allocations (alternating, false 'block') resulted in highly anomalous F-distributions, with over-representation of either extremely high or low p-values. F-distributions tended to 'collapse' towards extreme values unrelated to the expected theoretical distribution, but related to the direction of bias (Figure 4; Table 1).

3.2. Treatment differences and precision

Summary statistics for residuals and group differences are shown in Table 2. For randomized allocation, blocking resulted in increased precision for estimates of effect sizes. In contrast, grouping of similar weights in sequential treatment clusters by false block allocation resulted in biased and greatly exaggerated treatment differences that in practice would be flagged as highly statistically significant. Problems with treatment alternation in the presence of systematic trend are more subtle. Variation was somewhat increased (slightly increasing risk of obscuring true treatment differences if they existed), but overall summary statistics appeared to resemble those obtained for randomized allocation designs CR and RR. However, examination of distributional data indicates that the major underlying problem with alternation is invalidation of the null density distribution.

4. Conclusions

The null distribution of the test statistic is the appropriate reference distribution for NHST only if treatment allocation was actually randomized. These simulations clearly illustrate that inferential tests based on non-randomized treatment allocation will be invalid because the

underlying distributions from non-randomized allocations no longer model the appropriate density function for $F_{1-\alpha, \upsilon 1, \upsilon 2}$. As a result, inferential statistics cannot provide valid probability statements about treatment effects.

The consequences of systematic non-randomized treatment allocation for statistical inference were identified as early as the 1930s, and intermittently ever since (at least in the agricultural literature; see, for example [9,20,21]). However, it is still not uncommon to see published assertions in the biomedical literature that randomization provides 'no statistical advantage' (other than improved allocation concealment) over systematic methods such as alternation[22]. In contrast, this study emphasizes the real problem associated with non-randomization is invalidation of test statistics. The presence of systematic trend results in highly directional bias, and contributes to the extreme F-values and compressed distributions noted here. On the other hand, randomization and blocking of experimental units are highly efficient methods of accounting for spatial trend; blocking also increase estimate precision and reduces experimental error[21].

Preclinical studies that are poorly conducted and reported produce biased results, usually in the direction of exaggerated treatment effects[15,16,23,24]. Much of this bias would be avoidable if experiments were properly planned and designed prior to data collection[25]; the same avoidable biases may also contribute to the poor translation potential of much animal research[8]. The most serious problem would appear to be the proliferation of studies intended to be experimental tests of specific hypotheses, but analyzed as if treatment allocation was randomized properly when it was not. A 2009 survey showed that very few (12%) preclinical animal-based research reported random treatment allocation[15]. More recently (2018), over 60% of studies in a survey of swine preclinical research reported random treatment allocation. However, these claims could be directly assessed by calculating the p-value for treatment differences for studies reporting baseline summary statistics. Because random allocation of subjects to groups should result in an expected treatment difference of zero, the expected p-value distribution should be uniform (Figure 3). Instead, the surveyed studies were characterized both by over-representation of small p-values and inadequate reporting of methodology, making it impossible to determine what was actually done.[11] Non-uniform p-value distributions for baseline data have been used to detect non-randomization and, in some cases, provide supporting evidence for research fraud and misconduct.[26,27] It is not necessarily implied that the surveyed swine studies were fraudulent. It is likely these results reflect persistent systematic error of the type modelled here, and lack of investigator knowledge about the intent and practical implementation of randomization, coupled with 'boilerplate' statistical methods writing. Nevertheless, inaccurate reporting casts doubt on the validity of results.

Clearly, a high priority for statistical educators should be the instruction of investigators on randomization of experimental units as a critical design component of experimental studies. In addition, applied biostatisticians should be alerted to the necessity of enquiring into the provenance of the data handed over to them for analysis, especially with respect to strategies of data sampling and collection. Observational studies, which are non-randomized by definition, have numerous alternatives to conventional statistical methods of analysis and probabilistic interpretations[28]. Non-randomized experimental studies cannot be analyzed the same way.

Acknowledgements

I thank C. Garvan and T. Vasilopoulos (STAR-Core, Department of Anesthesiology, University of Florida) for helpful discussions during this project.

References

1. Altman DG, Bland JM. Treatment allocation in controlled trials: why randomise? *BMJ*. 1999;318:1209.
2. Jadad AR, Enkin MW. Bias in randomized controlled trials. In: *Randomized Controlled Trials: Questions, answers, and musings.* Second ed.: BMJ Books, Blackwell Publishing; 2008:160.
3. Schulz KF, Grimes DA. Generation of allocation sequences in randomised trials: chance, not choice. *The Lancet.* 2002;359:515-519.
4. Cox DR. Randomization in the design of experiments. *International Statistical Review.* 2009;77(3):415-429.
5. Kunz R, Vist G, Oxman AD. Randomisation to protect against selection bias in healthcare trials. *Cochrane Database of Systematic Reviews.* 2007;2007(2):Art. No.: MR000012.
6. Altman DG, Bland JM. How to randomise *BMJ.* 1999;319(7211):703-704.
7. O'Connor AM, Sargeant JM. Critical appraisal of studies using laboratory animal models. *ILAR J.* 2014;55(3):405-417.
8. Hirst JA, Howick J, Aronson JK, et al. The need for randomization in animal trials: an overview of systematic reviews. *PLoS One.* 2014;9(6):e98856.
9. Greenberg BG. Why randomize? *Biometrics.* 1951;7(4):309-322.
10. Hannon BA, Oakes JM, Allison DB. Alternating assignment was incorrectly labeled as randomization. *J Alzheimers Disease* 2019;71(1):1.
11. Reynolds PS, Garvan CW. Gap analysis of swine-based hemostasis research: "Houses of brick or mansions of straw?" *Mil Med.* 2019;in press.
12. Hall TW, Herron TL, Pierce BJ, Witt TJ. The effectiveness of increasing sample size to mitigate the influence of population characteristics in haphazard sampling. *Auditing: A Journal of Practice & Theory.* 2001;20(1):169-185.
13. Schmidt-Nielsen K. *Scaling: Why is animal size so important?* Cambridge: Cambridge University Press; 1984.
14. Nair AB, Jacob S. A simple practice guide for dose conversion between animals and human. *J Basic Clin Pharm.* 2016;16(7):27-31.
15. Kilkenny C, Parsons N, Kadyszewski E, et al. Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PLoS One.* 2009;4(11):e0007824.
16. Bebarta V, Luyten D, Heard K. Emergency medicine animal research: does use of randomization and blinding affect the results? *Acad Emerg Med* 2003;10(6):684-687.
17. Cochran WG. A catalogue of uniformity trial data *Supplement to the Journal of the Royal Statistical Society.* 1937;4(2):233-253.
18. Richter C, Kroschewski B. Geostatistical models in agricultural field experiments: investigations based on uniformity trials. *Agronomy J.* 2012;104(1):91-105.
19. Murdoch DJ, Tsai Y-L, Adcock J. P-values are random variables *The American Statistician.* 2012;62(3):242-245.
20. Piepho HP, Möhring J, Williams ER. Why randomize agricultural experiments? *J Agro Crop Sci.* 2013;199:374-383.
21. Edmondson RN. Past developments and future opportunities in the design and analysis of crop experiments. *Journal of Agricultural Science.* 2005;143:27-33.
22. Chalmers I, Clarke M. J Guy Scadding and the move from alternation to randomisation. *Journal of the Royal Society of Medicine.* 2016;109(7):282-283.
23. Kimmelman J, Henderson VC. Assessing risk/benefit for trials using preclinical evidence: a proposal. *J Med Ethics* 2016;42(1):50-53.

24. Macleod MR, Lawson McLean A, Kyriakopoulou A, et al. Risk of bias in reports of *in vivo* research: a focus for improvement. *PLOS Biology* 2015;13(11):e1002301.
25. Festing MFW, Altman DG. Guidelines for the design and statistical analysis of experiments using laboratory animals. *ILAR Journal.* 2002;432(4):244-258.
26. Carlisle JB. Data fabrication and other reasons for non-random sampling in 5087 randomised, controlled trials in anaesthetic and general medical journals. *Anaesthesia.* 2017;72:944-952.
27. Carlisle JB, Dexter F, Pandit JJ, Shafer SL, Yentis SM. Calculating the probability of random sampling for continuous variables in submitted or published randomised controlled trials *Anaesthesia.* 2015;70:844-858.
28. Greenland S. Randomization, statistics, and causal inference. *Epidemiology.* 1990;1(6):421-429.

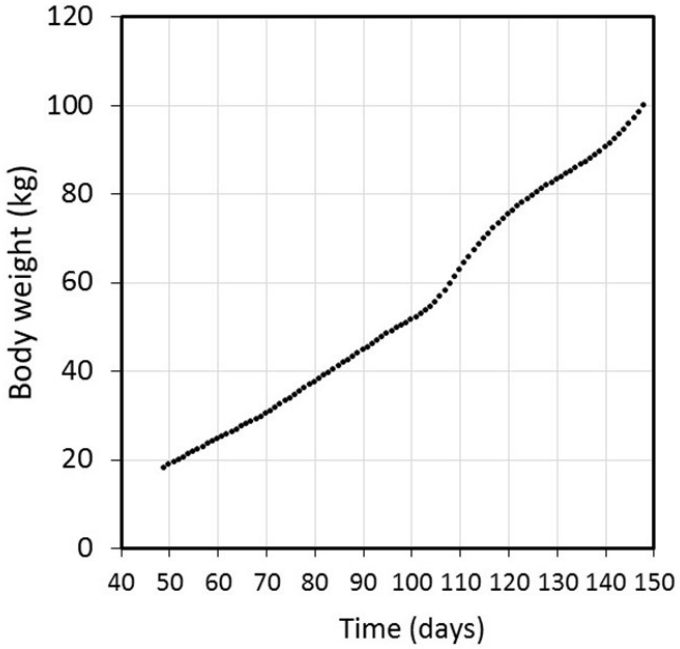

Figure 1. Generated test population N = 100 for body weight data.

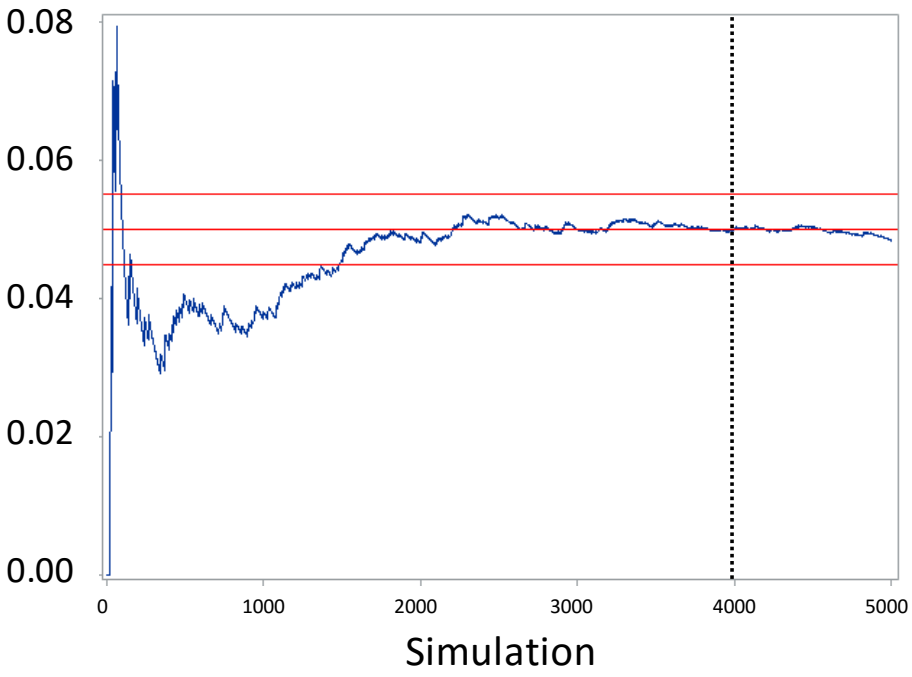

Figure 2: Cumulative type I error rate as a function of simulation replicate for completely randomized allocation designs (CRD) with total N = 36 and $\tau = 3$ treatment arms. Reference lines (red) are $\alpha = 0.05$; range 0.045-0.055.

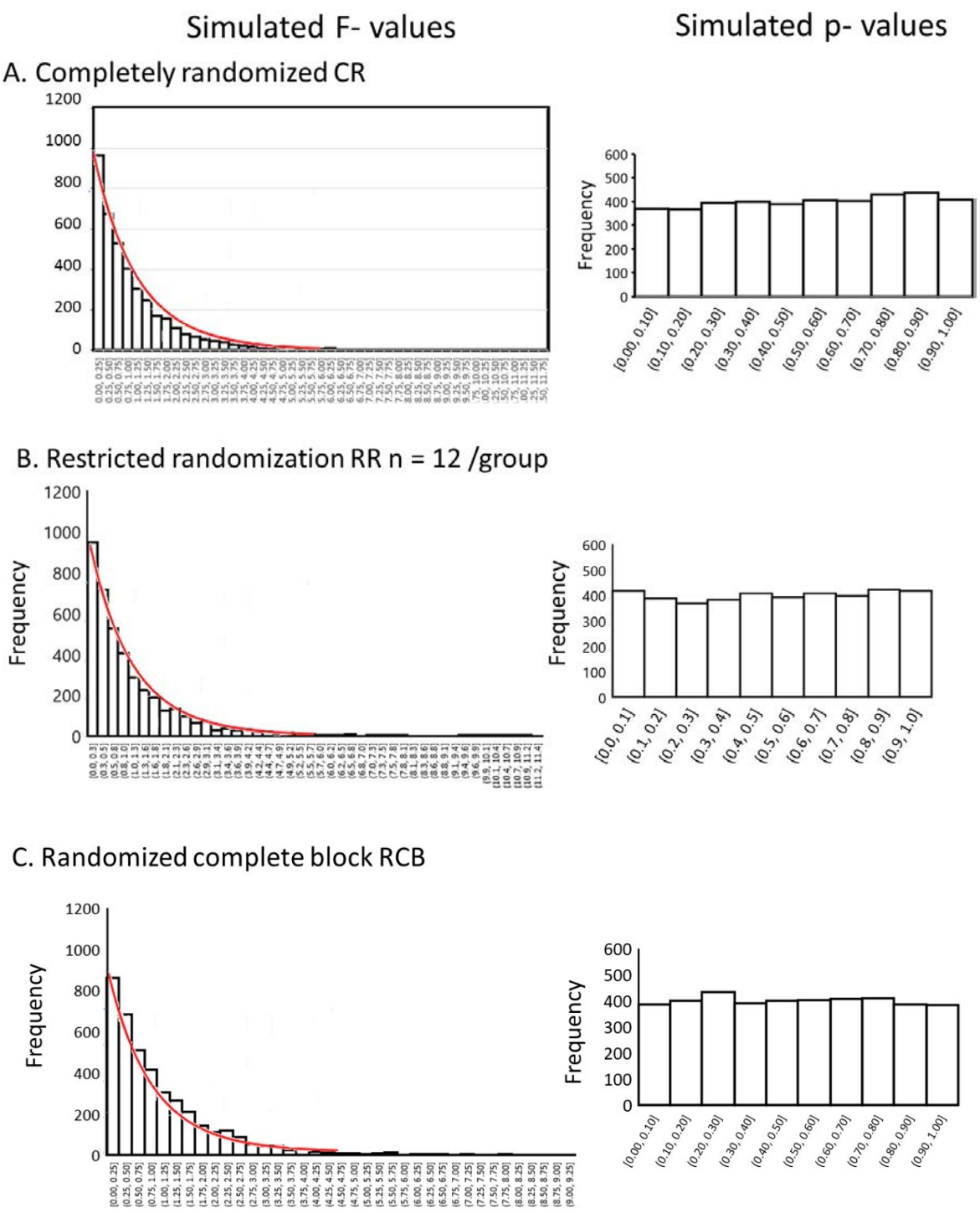

Figure 3. F-distributions and p-value distributions obtained from 4000 simulations of three randomized allocation schedules for three treatments and N = 36: (A) Completely randomized (CRD); (B) Restricted randomization (RRD); (C) Randomized complete block design with 6 blocks, random block effects (fixed effects not shown). The red line indicates the F density function for $F_{2,33}$ for CR and RR allocation

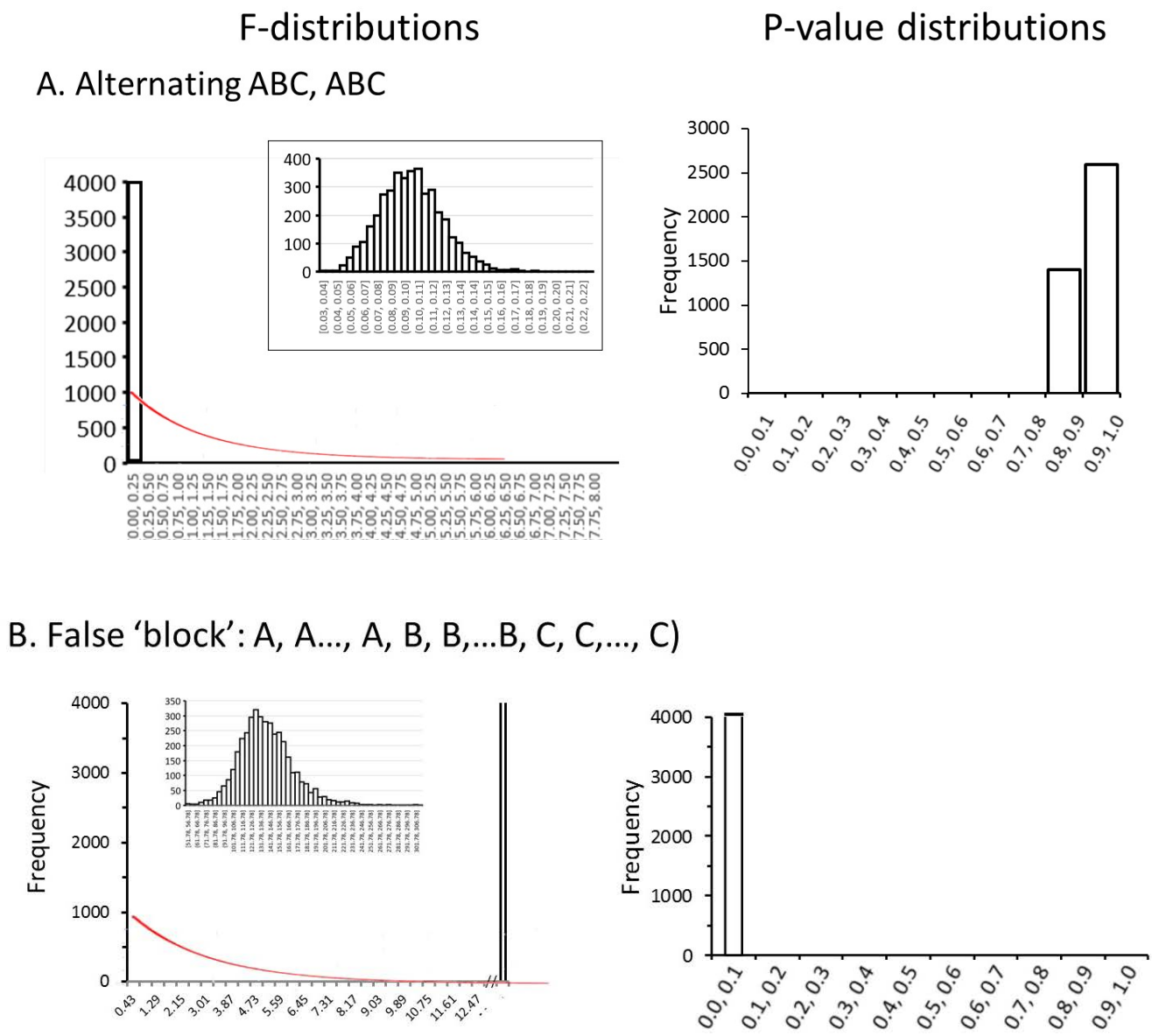

Figure 4. Effects of non-randomized allocation on F-distributions and p-values. Red lines indicate theoretical F-distribution for $F_{2,33}$. Insets show F-distributions over the restricted range of simulated values. Note F-distribution 'collapse' with over-representation of very large (A) or very small (B) p-values.

Table 1. F-value distributions (5^{th}, 50^{th}, and 95^{th} percentiles) based on 4000 simulations for randomized and non-randomized (systematic) allocation schemes

	Percentiles		
	5^{th}	50^{th}	95^{th}
1. Randomised allocation			
CRD	0.05	0.67	3.19
RRD	0.05	0.69	3.43
RCBD b = 12	0.05	0.72	3.43
b = 6	0.05	0.70	3.22
2. Non-randomised allocation			
Alternating	0.06	0.10	0.14
False block	97.29	137.64	192.91

Table 2. Summary estimates (5^{th}, $50^{th,}$ and 95^{th} percentiles) for MSE and group differences

			Group differences 50^{th} percentile			
Allocation method	MSE	5^{th}, 95^{th} percentiles	A-B	B-C	A-C	SE
1. Randomised						
CRD	604.67	475.75, 733.25	-0.24	-0.30	-0.46	10.25
RRD	604.28	472.14, 736.08	0.06	-0.18	-0.22	10.04
RCBD b = 12	7.42	4.51, 11.68	0.02	-0.40	-0.35	1.11
RCBD b = 6	22.09	15.54, 30.37	-0.03	0.05	0.02	1.92
2. Non-randomised						
Alternating	637.65	513.65, 760.32	-2.49	-1.92	-4.41	10.31
False block	69.24	50.42, 90.83	-23.63	-31.75	-55.83	3.38