Accepted Manuscript

Simple Randomization Did not Protect Against Bias in Smaller Trials

Tri-Long Nguyen, Pharm.D. M.Sc., Gary S. Collins, Ph.D., André Lamy, M.D., P.J. Devereaux, M.D. Ph.D., Jean-Pierre Daurès, M.D. Ph.D., Paul Landais, M.D. Ph.D., Yannick Le Manach, M.D. Ph.D.

PII: S0895-4356(16)30344-4

DOI: [10.1016/j.jclinepi.2017.02.010](https://doi.org/10.1016/j.jclinepi.2017.02.010)

Reference: JCE 9332

To appear in: *Journal of Clinical Epidemiology*

Received Date: 12 August 2016

Revised Date: 26 January 2017

Accepted Date: 21 February 2017

Please cite this article as: Nguyen T-L, Collins GS, Lamy A, Devereaux PJ, Daurès J-P, Landais P, Le Manach Y, Simple Randomization Did not Protect Against Bias in Smaller Trials, *Journal of Clinical Epidemiology* (2017), doi: 10.1016/j.jclinepi.2017.02.010.

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Simple randomization did not protect against bias in smaller trials

Tri-Long Nguyen Pharm.D. M.Sc.,¹ Gary S. Collins Ph.D.,² André Lamy M.D.,³ P.J. Devereaux M.D. Ph.D.,⁴ Jean-Pierre Daurès M.D. Ph.D.,⁵ Paul Landais M.D. Ph.D.,⁶ Yannick Le Manach M.D. Ph.D.⁷

¹ Pharm.D., M.Sc., Ph.D. candidate, Laboratory of Biostatistics, Epidemiology, Clinical Research and Health Economics, EA2415, Montpellier University, Montpellier, France; Departments of Anesthesia & Clinical Epidemiology and Biostatistics, Michael DeGroote School of Medicine, Faculty of Health Sciences, McMaster University and the Perioperative Research Group, Population Health Research Institute, Hamilton, Canada.

² Ph.D., Professor, Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Botnar Research Centre, University of Oxford, Windmill Road, Oxford, United Kingdom.

³ M.D., Professor, Population Health Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, Perioperative Medicine and Surgical Research Unit, McMaster University, Hamilton, Canada.

⁴ M.D., Ph.D., Professor, Population Health Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, Perioperative Medicine and Surgical Research Unit & Departments of Clinical Epidemiology and Biostatistics and Medicine, Michael DeGroote School of Medicine, Faculty of Health Sciences, McMaster University, Hamilton, Canada.

⁵ M.D., Ph.D. Professor, Montpellier University, Faculty of Medicine, EA2415, Biostatistics, Epidemiology, Clinical Research and Health Economics, France.

⁶ M.D., Ph.D. Professor and Chairman, Montpellier University, Faculty of Medicine, EA2415, Department of Biostatistics, Clinical Research and Medical Informatics, Nîmes University Hospital, France.

⁷ M.D., Ph.D., Assistant Professor, Departments of Anesthesia & Clinical Epidemiology and Biostatistics, Michael DeGroote School of Medicine, Faculty of Health Sciences, McMaster University and the Perioperative Research Group, Population Health Research Institute, Hamilton, Canada.

Corresponding author:

Yannick LE MANACH, M.D., Ph.D.;

Departments of Anesthesia & Clinical Epidemiology and Biostatistics; Michael DeGroote School of Medicine; Faculty of Health Sciences, McMaster University. Hamilton, Ontario, Canada.

Population Health Research Institute; David Braley Cardiac, Vascular and Stroke Research Institute; Perioperative Medicine and Surgical Research Unit; Hamilton, Ontario, Canada.

Email: yannick.lemanach@phri.ca

Phone: (905) 527-7327

Author contributions:

- **Study concept and design:** Nguyen, Collins, Le Manach.
- **Analysis and interpretation:** Nguyen, Collins, Le Manach.
- **Drafting of the manuscript:** Nguyen, Collins, Le Manach.
- **Critical revision of the manuscript for important intellectual content:** Nguyen, Collins, Lamy, Devereaux, Daurès, Landais, Le Manach.

Word count:

- **Abstract:** 245
- **Text:** 3,597

ABSTRACT

Objective: By removing systematic differences across treatment groups, simple randomization is assumed to protect against bias. However, random differences may remain if the sample size is insufficiently large. We sought to determine the minimal sample size required to eliminate random differences, thereby allowing an unbiased estimation of the treatment effect.

Study Design and Setting: We re-analyzed two published multicenter, large and simple trials: the International Stroke Trial (IST) and the CORONARY study. We reiterated 1,000 times the analysis originally reported by the investigators in random samples of varying size. We measured the covariates balance across the treatment arms. We estimated the effect of Aspirin and Heparin on death or dependency at 30 days after stroke (IST), and the effect of off-pump coronary-artery bypass grafting on a composite primary outcome of death, nonfatal stroke, nonfatal myocardial infarction, or new renal failure requiring dialysis at 30 days (CORONARY study). In addition, we conducted a series of Monte Carlo simulations of randomized trials to supplement these analyses.

Results: Randomization removes random differences between treatment groups when including at least 1,000 participants, thereby resulting in minimal bias in effects estimation. Below, substantial bias is observed. In a short review, we show such an enrollment is achieved in 41.5% of phase 3 trials published in the highest-impact medical journals.

Conclusions: Conclusions drawn from completely randomized trials enrolling a few participants may not be reliable. In these circumstances, alternatives such as minimization or blocking should be considered for allocating the treatment.

KEYWORDS: Randomization; sample size; bias.

INTRODUCTION

The Randomized Controlled Trial (RCT) is considered as the gold standard to provide evidence for clinical decision-making [1]. In such an experimental design, the treatment allocation is completely at random with respect to the patient's baseline characteristics. By doing so, the treatment groups are assumed to be balanced on measured and unmeasured confounders, allowing an unbiased estimation of the treatment effect [2]. However, although systematic differences between groups are removed, random differences can remain. One could speculate that simple randomization removes the random differences across the treatment groups only in sufficiently large trials. Rubin wrote: "complete randomization may not be good enough", except "in very large experiments" [3]. Such as flipping a coin, it appears intuitive that simple randomization may not guarantee the balance in small trials, contrary to large trials. This refers to the Law of Large Numbers, firstly described by Jakob Bernoulli as his "Golden Theorem" [4]. According to this Law, the sample mean of a random variable converges to the population mean as the sample size approaches infinity. In RCTs, both treated and control patients are sampled from the same population. If the sample size is sufficiently large, the distribution of the baseline characteristics within each treatment group converges to the population distribution. In these circumstances, the treatment and the control groups are similar on average. This theorem remains often ignored in medical field. Because the probability of removing the random differences across the treatment groups increases with larger sample size, enrolling enough participants is of greatest importance to provide reliable evidence of the treatment effect. As previously demonstrated, this risk of bias induced by covariate imbalance (sometimes called "accidental bias") converges asymptotically to null [5]. Furthermore, large-scale trials reduce variation in estimated effects and allow researchers to draw conclusions with a greater confidence [6]. Although this need of large RCTs has been

supported for years [7, 8], the actual sample sizes of contemporary RCTs remain small [9, 10].

In this study, we investigated the minimal sample size required to remove the random differences between treatment groups in RCTs. We aimed to demonstrate that discarding the Law of Large Numbers leads to substantial bias in the treatment effect measurement. To this end, we re-analyzed two multicenter, large and simple trials assessing the effect of Aspirin and Heparin after stroke [11], and comparing off-pump *versus* on-pump Coronary Artery Bypass Grafting (CABG) [12]. In addition, we conducted a short review to illustrate our findings to the actual enrollment of participants in RCTs published in high-impact journals.

METHODS

We evaluated how the Law of Large Numbers guarantees the covariate balance and thus the unbiased estimation of the treatment effect in RCTs, by using two multicenter, large and simple trials: the International Stroke Trial (IST) [11] and the CABG Off or On Pump Revascularization Study (CORONARY) [12]. In addition, we conducted a series of Monte Carlo simulations, through which the “true treatment effect” was known. Then, we explored the actual enrollment in RCTs published in the five leading general medical journals (*Annals of Internal Medicine*, the *British Medicine Journal*, the *Journal of the American Medical Association*, *The Lancet*, and the *New England Journal of Medicine*).

Law of Large Numbers in RCTs

To illustrate how the properties of the Law of Large Number may ensure the covariate balance and impact the results of RCTs, we re-analyzed two published studies. The IST, including 19,435 adult patients suspected of acute ischaemic stroke from 36 countries, evaluated the effect of Aspirin and Heparin on death or dependency at 6 months, by a two-by-two factorial design [11]. The database of this trial has been made available by the investigators [13]. The CORONARY study, including 4,752 senior patients undergoing a CABG from 19 countries, compared the procedure off-pump *versus* on-pump regarding a composite primary outcome of death, nonfatal stroke, nonfatal myocardial infarction, or new renal failure requiring dialysis at 30 days [12]. For each RCT, we generated random subsamples with replacement from the original data set, by varying the sample size (n from 50 to 5,000 subjects, stepwise 50), which replicated RCTs with different enrollments. For each size, 1,000 RCTs were simulated, in which we measured the covariate balance across the treatment groups and estimated the treatment effect. The balance reflects the average

difference between the two treatment groups of a sample, which is assumed to be minimal when the covariates converge to their population mean. As the balance is a sample property rather than a population property, we did not assess covariates balance with significance tests [14-18]. We used absolute standardized mean differences, which are not affected by the sample size variation. For a variable (or a category of a discrete variable) W , the absolute standardized mean difference across the treatment groups is defined by:

$$SMD = \frac{|mean(W)_{treatment} - mean(W)_{control}|}{\sqrt{\frac{variance(W)_{treatment} + variance(W)_{control}}{2}}}$$

We used this metric for assessing the baseline characteristics measured in the original articles. In the IST, we reported the standardized differences in delay of administration, age, sex, onset (awake; during sleep), conscious level (unconscious; drowsy; alert), cardiac rhythm (sinus rhythm; atrial fibrillation), systolic blood pressure, stroke syndrome (total anterior; partial anterior; posterior circulation; lacunar), leg weakness (present; absent; not assessable), pre-randomization CT scan; appearance of CT scan (infarct visible; no infarct visible); pre-randomization antithrombotic therapy (Aspirin within previous 3 days; Heparin within previous 24 hours) [11]. In the CORONARY study, we reported the standardized differences in age, sex, body mass index, previous myocardial infarction, previous percutaneous coronary intervention, previous stroke, peripheral arterial disease, diabetes, renal failure requiring dialysis, congestive heart failure, hypertension, left ventricular ejection fraction (grade 1: $\geq 50\%$; grade 2: 35 to 49%; grade 3: 20 to 34%, grade 4: $<20\%$), EuroSCORE grade (0 to 2; 3 to 5; >5), urgent surgery, diseased vessels (left main; triple; double; single) [12].

Like in the original articles, we calculated the treatment effect as an absolute mean difference in the IST, and as a hazard ratio in the CORONARY study. For each sample size n , we

compared the estimated effects to the reference values. We calculated the mean squared error, which measures the bias and variance of the estimated effects:

$$MSE_n = \text{mean}[(\text{estimated effect}_n - \text{reference value})^2]$$

Though we recognize this metric did not allow us to distinguish accidental bias (related to the covariates imbalance) from random errors (related to the sample size), these two sources of error must be minimized to allow reliable conclusions from RCTs. Hazard ratios were log-transformed in this calculation. We considered the reference values as the treatment effects originally presented by the investigators. As reported by the IST collaborative group, the effects of Aspirin and Heparin were, in absolute risk difference scale, equal to -1.25% and -0.03%, respectively [11]. As reported by Lamy *et al.*, the effect of off-pump CABG was equal to a hazard ratio of 0.95 [12].

As these analyses assumed that the reference values coincided with the “true treatment effects”, we conducted an additional series of Monte Carlo simulations to complete our approach.

Monte Carlo simulations

We emulated fictive RCTs in which the “true treatment effect” was known. We generated 14 baseline covariates following random normal distributions. From these, we created a set of binary (W_1 to W_{10}), continuous (W_{11} to W_{13}) and categorical (W_{14}) variables that were all designed to approximate those described in perioperative populations (*e.g.* age, comorbidities and biological markers) [19, 20]. We fixed different correlation coefficients between these covariates. An additional variable (W_{15}) was defined as the combination of others to reproduce the cardiac risk index proposed by Lee *et al* [21]. To simulate simple randomization, the treatment was allocated according to a random binomial distribution

B(0.5), with $Z=1$ and $Z=0$ denoting the treated and the control subjects, respectively. The occurrence of the outcome Y (binary variable) followed the model:

$$\begin{aligned} \text{logit}[p(Y)] = & \alpha_0 + \alpha_1 W_1 + \alpha_2 W_2 + \alpha_3 W_3 + \alpha_4 W_4 + \alpha_5 W_5 + \alpha_7 W_7 + \alpha_9 W_9 + \alpha_{10} W_{10} + \alpha_{11} W_{11} \\ & + \alpha_{12} W_{12} + \alpha_{13} W_{13} + \alpha_{15} W_{15} + \gamma_T Z \end{aligned}$$

The patient experienced the outcome ($Y=1$) if $p(Y)$ was greater than a random number generated from an uniform distribution $U(0,1)$. Introducing an error, this random process allowed us to consider a random outcome assignment rather than a fixed and deterministic one. In addition, we could compute Y_1 and Y_0 , denoting for each patient her/his potential outcome value that would be observed if, possibly contrary to the fact, she/he were treated and not treated, respectively. Having her/his own counterfactual outcome value, each patient was thus compared to herself/himself (*i.e.* “utopian experiment” with perfect balance). Given a trial, we could therefore assess the performance of the simple randomized experiment against the “utopian experiment” that exactly included the same set of patients. This allowed us to measure the loss of performance attributable to simple randomization. After calculating the average treatment effect as an absolute risk difference in both types of experiment (average difference in Y across the treatment groups and average difference between Y_1 and Y_0 , respectively), we reported the mean squared error for both estimators. Then, we computed the difference in relative bias (*i.e.* bias expressed as a percentage of the reference value) between the two estimators and calculated the average standard error of the randomization-based estimator which was compared with the empirical standard deviation of the estimated treatment effect across the simulations. This allowed us to separately evaluate the impact of the sample size on the performance of the randomization-based estimator and its variance estimator. The reference value (*i.e.* “true average treatment effect”) was equal to -2.55%. This marginal effect was obtained by fixing, in the outcome equation, a conditional odds ratio of 0.80 (log-odds ratio $\gamma_T = -0.23$) being consistent with real-life RCTs [22-24]. We varied the

sample size from 50 to 5,000 patients, stepwise 50 (1,000 generated samples per size). All parameters for data generation are described in *Supplementary Material*.

Number of trial participants in medical literature

To describe the actual enrollment in clinical trials, we extracted all of the studies indexed in PubMed as “clinical trials” and published on November 15th, 2015, in one of the top five highest-impact journals in Medicine (the *Annals of Internal Medicine*, the *British Medicine Journal*, the *Journal of the American Medical Association*, *The Lancet*, and the *New England Journal of Medicine*). Of these studies, we kept those reporting a link to ClinicalTrials.gov. After merging the two databases (PubMed and ClinicalTrials.gov) on the NCT registration number, we automatically extracted the number of enrolled participants in the interventional studies, in which the treatment was randomly allocated in a parallel or factorial design. Though some of these could use alternatives to simple randomization, such as blocked randomization or minimization on measured covariates, imbalance could remain in unmeasured covariates. We focused our analysis on the phase 3 trials, as they were especially designed to demonstrate the effectiveness of new treatments.

RESULTS

Law of Large Numbers in RCTs

We found a same “point of balance” in the IST and the CORONARY study (*Figure 1*). This threshold was found as the value of N , after which both the covariate balance and the mean squared error plateaued. In both trials, all of the baseline characteristics were correctly balanced across treatment groups only when enrolling 1,000 participants or more. Beyond this threshold, we reported minimal mean squared error in estimating the effects of Aspirin and Heparin (IST) and off-pump CABG (CORONARY study). Below, baseline characteristics were imbalanced, and substantial mean squared error in effects measurement were observed.

These results were confirmed in our series of Monte Carlo simulations, across which the average treatment and outcome prevalence were equal to 50.0% and 15.5%, respectively. The covariate balance was appropriately achieved in sample sizes greater than 1,000. In *Figure 2*, we showed that the additional mean squared error (*i.e.* loss of performance) attributable to simple randomization was minimized when enrolling at least 1,000 participants. This result appeared to be related to the ‘accidental’ bias caused by the covariate imbalance in simple randomization. On a relative scale, the additional bias attributable to simple randomization plateaued at $N = 1,000$ participants (*Figure 3*). In contrast, the estimated standard error was consistent with the empirical standard deviation regardless of the number of participants, thereby suggesting that the performance of the variance estimator was not affected by the sample size (*Figure 3*).

Number of trial participants in medical literature

At the date of November 15th, 2015, a total of 11,416 RCTs were published in the top five highest-impact medical journals. Of these, 2,405 reported a link to ClinicalTrials.gov, from

which 2,172 records matched with publications. Randomized allocation and parallel or factorial assignment were reported in 1,918 studies and 1,004 were phase 3 trials (*Figure 4*).

As illustrated in *Figure 5*, the number of published phase 3 RCTs substantially increased since the beginning of the century. However, the proportion of RCTs enrolling more than 1,000 participants appeared constant in time, with less than a half of phase 3 trials (417 RCTs, 41.5%). Unsurprisingly, this proportion varied according to the medical specialty (*Table 1*). Among the 215 phase 3 trials in cardiovascular diseases, 150 (69.8%) respected this threshold of 1,000 participants. This proportion was much lower in Oncology (58 trials among 187, 31.0%). No substantial difference was found across the age groups, but trials enrolling all of the age categories (child, adult and senior) were fewer to include 1,000 subjects (32 trials among 94, 34.0%). Phase 3 trials, which benefited from multiple funding sources were more likely to include 1,000 participants or more (120 among 252 trials, 47.6%).

DISCUSSION

We showed that simple randomization protects against bias only in trials enrolling 1,000 patients or more. This threshold is necessary to insure the covariates balance across the treatment arms and subsequently, an unbiased estimation of the treatment effect. Enrolling fewer participants resulted in substantial bias in estimating the treatment effect. In addition, we found that not even a half of phase 3 RCTs published in the top five highest-impact journals actually achieved this number of inclusions.

In experimental design, complete randomization is often assumed to protect against bias, because it removes the systematic differences across the treatment groups on average [25]. As the treated population is similar to the control population, randomization may protect against selection bias [26]. As well, by balancing measured and unmeasured prognostic factors across the treatment groups, randomization may protect against confounding [2]. However, random differences can remain. These assumptions are therefore verified only under the Law of Large Numbers, that is to say, the treated and the control groups are similar on average only if the sample is large enough. Within very large samples, the average values of covariates converge to their population expectations, resulting in no difference across the treatment groups. Previous studies have shown the accidental bias due to these random differences asymptotically approaches the null [5]. Although a minimal sample size of 200 has been previously proposed to ensure an equal treatment allocation (i.e. proportions of treated and controls that approach 50%), no threshold was found to ensure adequate balance of the covariates [5]. We showed that this balance occurs in samples including 1,000 subjects or more. Below this value, the assumptions attributed to randomization might be no more verified, thereby leading to bias in estimated effects. In response to this issue, one could advocate adjusting for imbalanced covariates. However, such an approach is limited. First, it

assumes no unmeasured confounders [27]. Second, adjusting allows the estimation of a conditional effect, which is not necessarily equal to the marginal effect estimated in the RCT [27, 28]. Third, adjustment is a model-dependent approach, which is criticized for causal inference purpose. Indeed, as the estimate depends on the model's specification, the researcher might be tempted to report the result, which "fits its favorite hypothesis" [29, 30]. We recognize that the possibility of inclusion varies according to financial supports and medical specialties. Enrolling 1,000 participants may be difficult regarding certain conditions (e.g., rare diseases). In such situations, alternatives to simple randomization should be considered. For example, a greater power, efficiency and robustness of the fully blocked randomized experiment have been advocated [30-33]. In such a design, the imbalance is reduced to null, as the treated and the controls are blocked at the start on the covariates. Treatment allocation by minimization is also an alternative [34, 35]. As well, compared to parallel assignments, cross-over designs may reduce the variance of estimates [36].

In contrast with our findings, only a few RCTs actually enroll more than 1,000 participants. Califf *et al.* reported that 3.1% of studies registered on ClinicalTrials.gov between 2007 and 2010 included more than 1,000 subjects [9]. According to Hopewell *et al.* [10], the median sample sizes of RCTs indexed in PubMed database in 2000 and 2006, were 32 and 36 in each treatment group. In the top five highest-impact journals, we report a larger proportion of RCTs enrolling more than 1,000 participants in phase 3. However, more than a half should still be considered too small to provide fair evidences. This finding is quite alarming, as these studies were designed to provide evidences of effectiveness of new treatments. "Studies of lesser quality and smaller scope should not carry equal weight to large-scale comprehensive trials just because they are randomized" said Gregg Stone and Stuart Pocock [37]. This necessity of working on large samples has been supported for years [7, 8]. As previously written by Richard Peto, "useful trials must usually be capable of distinguishing between the

alternative possibilities of a small treatment effect and no treatment effect” [38], referring to large RCTs with sufficient statistical power to detect significant (small) effects. In addition, we argue that sample size calculation should not be based only on statistical power. As a simple example, if a treatment has a major effect on reducing a highly frequent outcome, a power-based sample size calculation might lead to including less than 1,000 participants. Admittedly, the result of such a RCT would be significant, but probably biased. We agree with Martin Bland that sample size calculation should not be dominated by the “tyranny of power” [39]. As long as “a p value is no substitute for a brain”, considering statistical power is necessary but not sufficient for sample size calculation. In addition to these issues, some authors have previously revealed important deficiencies in the quality of reporting of RCTs [40-42].

Parallel to the strengths of our study, some limitations have to be considered. First, our analysis was conducted only in RCTs with an equal allocation. We focused on this allocation type, because it is very popular in practice. Through two illustrative RCTs, we found consistent results regardless of the type of outcome (binary or censored), the scale of treatment effect (absolute risk difference or hazard ratio), the age group (adult or senior) and the type of intervention (medical or surgical). In spite of a large set of baseline covariates, we found a unique “point of balance” in both RCTs. Because this corresponds to the point minimizing both bias and variation of measurement effects, we assume that unmeasured covariates were also balanced. This “point of balance” depends on the covariates variance within the population and might be therefore greater in populations with a widespread distribution of patient’s characteristics (e.g., populations including all the categories of age: child, adult and senior). Second, we explored the enrollment only in RCTs registered on ClinicalTrials.gov. We used this large and publically accessible database to extract data of trials published in the top five highest-impact medical journals, which are generally

considered as the highest-quality studies. As discussed before, the prospective registration of such database may reduce publication bias [43, 44]. In our case, regarding the number of trial participants, only a few discrepancies are expected between the registry and publications [43, 45].

To conclude, simple randomization does not necessarily protect against bias without respect to the Law of Large Numbers. Insufficiently sized trials may exhibit substantial bias and wrong conclusions may be drawn.

ACKNOWLEDGEMENTS

We are thankful to Professor Peter A.G. Sandercock and the International Stroke Trial collaborative group for releasing the IST database in open-access. This trial was principally funded by the UK Medical Research Council, the UK Stroke Association, and the European Union BIOMED-1 program. Limited support for collaborators' meetings and travel was provided by Eli Lilly, Sterling Winthrop (now Bayer USA), Sanofi, and Bayer UK. Follow-up in Australia was supported by a grant from the National Heart Foundation and in Canada by a Nova Scotia Heart and Stroke Foundation grant. Czech Republic IST was supported by a grant from the IGA Ministry of Health. India IST was supported by the McMaster INCLEN program and the All India Institute of Medical Sciences. The IST in New Zealand was funded by the Julius Brendel Trust and the Lottery Grants Board. In Norway, the IST was supported by the Norwegian Council on Cardiovascular Disease and Nycomed (for insurance).

We are thankful to the CORONARY collaborative group. All funding was provided by the Canadian Institutes of Health Research. No manufacturers of off-pump or on-pump CABG supplies or devices had any role in the study. The data were gathered and analyzed by the Population Health Research Institute at McMaster University and Hamilton Health Sciences.

REFERENCES

- [1] Guyatt GH, Oxman AD, Kunz R, Vist GE, Falck-Ytter Y, Schunemann HJ, et al. What is "quality of evidence" and why is it important to clinicians? *Bmj*. 2008;336:995-8.
- [2] Altman DG, Bland JM. Statistics notes. Treatment allocation in controlled trials: why randomise? *Bmj*. 1999;318:1209.
- [3] Rubin DB. Comment: The Design and Analysis of Gold Standard Randomized Experiments. *Journal of the American Statistical Association*. 2008;103:1350-3.
- [4] Bernoulli J. *Ars Conjectandi: Usus & Applicationem Praecedentis Doctrinae in Civilibus, Moralibus & Oeconomicis*. 1793.
- [5] Lachin JM. Properties of simple randomization in clinical trials. *Controlled clinical trials*. 1988;9:312-26.
- [6] Pocock SJ, Hughes MD. Estimation issues in clinical trials and overviews. *Statistics in medicine*. 1990;9:657-71.
- [7] Yusuf S, Collins R, Peto R. Why do we need some large, simple randomized trials? *Statistics in medicine*. 1984;3:409-22.
- [8] Pocock SJ, Clayton TC, Stone GW. Design of Major Randomized Trials: Part 3 of a 4-Part Series on Statistics for Clinical Trials. *Journal of the American College of Cardiology*. 2015;66:2757-66.
- [9] Califf RM, Zarin DA, Kramer JM, Sherman RE, Aberle LH, Tasneem A. Characteristics of clinical trials registered in ClinicalTrials.gov, 2007-2010. *Jama*. 2012;307:1838-47.

- [10] Hopewell S, Dutton S, Yu LM, Chan AW, Altman DG. The quality of reports of randomised trials in 2000 and 2006: comparative study of articles indexed in PubMed. *Bmj*. 2010;340:c723.
- [11] The International Stroke Trial (IST): a randomised trial of aspirin, subcutaneous heparin, both, or neither among 19435 patients with acute ischaemic stroke. International Stroke Trial Collaborative Group. *Lancet*. 1997;349:1569-81.
- [12] Lamy A, Devereaux PJ, Prabhakaran D, Taggart DP, Hu S, Paolasso E, et al. Off-pump or on-pump coronary-artery bypass grafting at 30 days. *The New England journal of medicine*. 2012;366:1489-97.
- [13] Sandercock PA, Niewada M, Czlonkowska A, International Stroke Trial Collaborative G. The International Stroke Trial database. *Trials*. 2011;12:101.
- [14] Senn S. Testing for baseline balance in clinical trials. *Statistics in medicine*. 1994;13:1715-26.
- [15] Senn S. Covariate imbalance and random allocation in clinical trials. *Statistics in medicine*. 1989;8:467-75.
- [16] Altman DG, Dore CJ. Randomisation and baseline comparisons in clinical trials. *Lancet*. 1990;335:149-53.
- [17] Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet*. 2000;355:1064-9.
- [18] Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in medicine*. 2002;21:2917-30.

- [19] Botto F, Alonso-Coello P, Chan MT, Villar JC, Xavier D, Srinathan S, et al. Myocardial injury after noncardiac surgery: a large, international, prospective cohort study establishing diagnostic criteria, characteristics, predictors, and 30-day outcomes. *Anesthesiology*. 2014;120:564-78.
- [20] Vascular Events In Noncardiac Surgery Patients Cohort Evaluation Study Investigators, Devereaux PJ, Chan MT, Alonso-Coello P, Walsh M, Berwanger O, et al. Association between postoperative troponin levels and 30-day mortality among patients undergoing noncardiac surgery. *Jama*. 2012;307:2295-304.
- [21] Lee TH, Marcantonio ER, Mangione CM, Thomas EJ, Polanczyk CA, Cook EF, et al. Derivation and prospective validation of a simple index for prediction of cardiac risk of major noncardiac surgery. *Circulation*. 1999;100:1043-9.
- [22] POISE Study Group, Devereaux PJ, Yang H, Yusuf S, Guyatt G, Leslie K, et al. Effects of extended-release metoprolol succinate in patients undergoing non-cardiac surgery (POISE trial): a randomised controlled trial. *Lancet*. 2008;371:1839-47.
- [23] Pearse RM, Harrison DA, MacDonald N, Gillies MA, Blunt M, Ackland G, et al. Effect of a perioperative, cardiac output-guided hemodynamic therapy algorithm on outcomes following major gastrointestinal surgery: a randomized clinical trial and systematic review. *Jama*. 2014;311:2181-90.
- [24] Schouten O, Boersma E, Hoeks SE, Benner R, van Urk H, van Sambeek MR, et al. Fluvastatin and perioperative events in patients undergoing vascular surgery. *The New England journal of medicine*. 2009;361:980-9.
- [25] Altman DG. Randomisation. *Bmj*. 1991;302:1481-2.

- [26] Odgaard-Jensen J, Vist GE, Timmer A, Kunz R, Akl EA, Schunemann H, et al. Randomisation to protect against selection bias in healthcare trials. The Cochrane database of systematic reviews. 2011:MR000012.
- [27] Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*. 1984;7:431–44.
- [28] Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. *Am J Epidemiol*. 1987;125:761-8.
- [29] Ho DE, Imai K, King G, Stuart EA. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis*. 2007;15:199–236.
- [30] King G, Nielsen R. Why Propensity Scores Should Not Be Used for Matching. 2016. Available at: <http://gking.harvard.edu/files/gking/files/psnot.pdf?m=1456683191> (accessed on November 30th, 2016)
- [31] Imai K, King G, Nall C. The Essential Role of Pair Matching in Cluster-Randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation. *Statistical Science*. 2009;24:29–53.
- [32] Imai K, King G, Stuart EA. Misunderstandings among Experimentalists and Observationalists about Causal Inference. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*. 2008;171:481-502.
- [33] Greevy R, Lu B, Silver JH, Rosenbaum P. Optimal multivariate matching before randomization. *Biostatistics*. 2004;5:263–75.
- [34] Pocock SJ, Simon R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics*. 1975;31:103-15.

- [35] Treasure T, MacRae KD. Minimisation: the platinum standard for trials?. Randomisation doesn't guarantee similarity of groups; minimisation does. *Bmj*. 1998;317:362-3.
- [36] Senn S. *Cross-Over Trials in Clinical Research*. Second ed 2002.
- [37] Stone GW, Pocock SJ. Randomized trials, statistics, and clinical inference. *Journal of the American College of Cardiology*. 2010;55:428-31.
- [38] Peto R. Clinical trial methodology. *Biomedicine* / [publiee pour l'AAICIG]. 1978;28 Spec No:24-36.
- [39] Bland JM. The tyranny of power: is there a better way to calculate sample size? *Bmj*. 2009;339:b3985.
- [40] Boutron I, Dutton S, Ravaud P, Altman DG. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *Jama*. 2010;303:2058-64.
- [41] Le Henanff A, Giraudeau B, Baron G, Ravaud P. Quality of reporting of noninferiority and equivalence randomized trials. *Jama*. 2006;295:1147-51.
- [42] Mathieu S, Boutron I, Moher D, Altman DG, Ravaud P. Comparison of registered and published primary outcomes in randomized controlled trials. *Jama*. 2009;302:977-84.
- [43] Hartung DM, Zarin DA, Guise JM, McDonagh M, Paynter R, Helfand M. Reporting discrepancies between the ClinicalTrials.gov results database and peer-reviewed publications. *Annals of internal medicine*. 2014;160:477-83.
- [44] Song F, Parekh S, Hooper L, Loke YK, Ryder J, Sutton AJ, et al. Dissemination and publication of research findings: an updated review of related biases. *Health technology assessment*. 2010;14:iii, ix-xi, 1-193.

- [45] Becker JE, Krumholz HM, Ben-Josef G, Ross JS. Reporting of results in ClinicalTrials.gov and high-impact journals. *Jama*. 2014;311:1063-5.

ACCEPTED MANUSCRIPT

FIGURE LEGENDS

Figure 1. Differences in baseline characteristics across the treatment groups (A, B, C) and error on the effects measurement (D, E, F). In panels (A, B, C), each colored curve refers to a baseline characteristic (see Methods section).

Figure 2. Monte Carlo simulations of randomized controlled trials. Differences in baseline characteristics across the treatment groups (A) and additional error on the effects measurement attributed to simple randomization (B). In panel (A), each colored curve refers to a baseline covariate.

Figure 3. Monte Carlo simulations of randomized controlled trials. Additional relative bias attributable to simple randomization in estimated treatment effects (A) and ratio of estimated standard error to empirical standard deviation of simple randomization-based estimated treatment effects (B).

Figure 4. Flow chart of randomized controlled trials recorded in ClinicalTrials.gov and published in the *Annals of Internal Medicine*, the *BMJ*, the *JAMA*, *The Lancet* or the *New England Journal of Medicine* at November 15th, 2015.

Figure 5. Phase 3 randomized controlled trials recorded in ClinicalTrials.gov and published in the *Annals of Internal Medicine*, the *BMJ*, the *JAMA*, *The Lancet* or the *New England Journal of Medicine* on November 15th, 2015.

Table 1. Randomized controlled trials recorded on ClinicalTrials.gov and published in the *Annals of Internal Medicine*, the *BMJ*, the *JAMA*, *The Lancet* or the *New England Journal of Medicine*, at the date of November 15th, 2015.

	N	≥ 1 000 participants n = 417 (41.5%)	< 1 000 participants n = 587 (58.5%)
<i>Discipline</i>			
Cardiovascular Disease	215	150 (69.8%)	65 (30.2%)
Oncology	187	58 (31.0%)	129 (69.0%)
Infectious Disease	173	79 (45.7%)	94 (54.3%)
Endocrinology	73	29 (39.7%)	44 (60.3%)
Rheumatology and Internal Medicine	67	3 (4.5%)	64 (95.5%)
Others	289	98 (33.9%)	191 (66.1%)
<i>Age group</i>			
Child only	77	35 (45.5%)	42 (54.5%)
Adult only	65	29 (44.6%)	36 (55.4%)
Senior only	7	3 (42.9%)	4 (57.1%)
Child and adult	38	16 (42.1%)	22 (57.9%)
Adult and senior	723	302 (41.8%)	421 (58.2%)
Child, adult and senior	94	32 (34.0%)	62 (66.0%)
<i>Funding source</i>			
Industry	520	216 (41.5%)	304 (58.5%)
NIH	41	15 (36.6%)	26 (63.4%)
U.S. Fed	11	4 (36.4%)	7 (63.6%)
Others	180	62 (34.4%)	118 (65.6%)
Multiple sources	252	120 (47.6%)	132 (52.4%)

Simple Randomization Should be Reserved for Sufficiently Large Trials. What Does “Large” Mean?

Highlights

- Simple randomization removes systematic differences across the treatment groups; random differences may remain in insufficiently large trials, however.
- In small samples, random covariate imbalance leads to ‘accidental bias’.
- When enrolling at least 1,000 participants, simple randomization may ensure the covariate balance and minimize the risk of measurement error.