Journal of Clinical Epidemiology 84 (2017) 105–113

Journal of Clinical Epidemiology

Simple randomization did not protect against bias in smaller trials

Tri-Long Nguyen[a,b,c,d], Gary S. Collins[e], André Lamy[d], Philip J. Devereaux[b,d,f], Jean-Pierre Daurès[a], Paul Landais[a,g], Yannick Le Manach[b,c,d,*]

[a]*Laboratory of Biostatistics, Epidemiology, Clinical Research and Health Economics, UPRES EA2415, University of Montpellier, Montpellier, France*
[b]*Department of Clinical Epidemiology and Biostatistics, Michael DeGroote School of Medicine, Faculty of Health Sciences, McMaster University, Hamilton, Canada*
[c]*Department of Anesthesia, Michael DeGroote School of Medicine, Faculty of Health Sciences, McMaster University, Hamilton, Canada*
[d]*Population Health Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, Perioperative Medicine and Surgical Research Unit, Perioperative Research Group, Population Health Research Institute, McMaster University, Hamilton, Canada*
[e]*Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Botnar Research Centre, University of Oxford, Windmill Road, Oxford, United Kingdom*
[f]*Department of Medicine, Michael DeGroote School of Medicine, Faculty of Health Sciences, McMaster University, Hamilton, Canada*
[g]*Department of Biostatistics, Clinical Research and Medical Informatics, Nîmes University Hospital, Nîmes, France*

Accepted 21 February 2017; Published online 28 February 2017

Abstract

Objectives: 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 reanalyzed two published multicenter, large, and simple trials: the International Stroke Trial (IST) and the Coronary Artery Bypass Grafting (CABG) Off- or On-Pump Revascularization Study (CORONARY). 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 CABG on a composite primary outcome of death, nonfatal stroke, nonfatal myocardial infarction, or new renal failure requiring dialysis at 30 days (CORONARY). 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. Later, 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: Clinical trial; Randomization; Sample size; Covariate balance; Bias; Causal inference

1. 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 that 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, first described by Jakob Bernoulli as his golden theorem [4].

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

* Corresponding author. Tel./fax: +1-905-527-7327.
E-mail address: yannick.lemanach@phri.ca (Y. Le Manach).

http://dx.doi.org/10.1016/j.jclinepi.2017.02.010
0895-4356/

What is new?

Key findings
In small trials, simple randomization may not correctly remove covariate imbalance.

What this adds to what was known?
In small trials, random covariate imbalance leads to accidental bias.

What is the implication and what should change now?
Including at least 1,000 participants should ensure covariate balance.

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 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 reanalyzed two multicenter, large, and simple trials assessing the effect of aspirin and heparin after stroke [11] and comparing off-pump vs. 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.

2. 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*).

2.1. Law of large numbers in RCTs

To illustrate how the properties of the law of large numbers may ensure the covariate balance and impact the results of RCTs, we reanalyzed two published studies. The IST, including 19,435 adult patients suspected of acute ischemic 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 vs. 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 (SMDs), which are not affected by the sample size variation. For a variable (or a category of a discrete variable) W, the absolute SMD across the treatment groups is defined by:

$$\mathrm{SMD} = \frac{\left|\mathrm{Mean}(W)_{treatment} - \mathrm{mean}(W)_{control}\right|}{\sqrt{\frac{\mathrm{Variance}(W)_{treatment} + \mathrm{variance}(W)_{control}}{2}}}.$$

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 and during sleep), conscious level (unconscious, drowsy, and alert), cardiac rhythm (sinus rhythm and atrial fibrillation), systolic blood pressure, stroke syndrome (total anterior, partial anterior, posterior circulation, and lacunar), leg weakness (present,