The Cost of Imbalance in Clinical Trials

Sylvain Chassang*
Princeton University and NBER

Rong Feng
New York University

September 7, 2020

Abstract

Clinical trials following the "gold standard" of random assignment frequently use independent lotteries to allocate patients to treatment and control arms. Unfortunately, independent assignment can generate treatment and control arms that are unbalanced (i.e. treatment and control populations with significantly different demographics). This is regrettable since other assignment methods such as matched pair designs ensure balance across arms while maintaining randomization and permitting inference.

This paper seeks to measure the cost of imbalance with respect to gender in a sample of roughly 2000 clinical studies. We document significant imbalance: 25% of experiments have at least 26% more men in one treatment arm than in the other. In addition, clinical trials with greater imbalance have more dispersed treatment effects, indicating that imbalance reduces the informativeness of experiments. A simple structural model suggests that for a typical experiment, using a balanced random design could deliver informativeness gains equivalent to increasing the sample size by 18%.

KEYWORDS: clinical trials, balance, gender, informativeness.

*Corresponding author; E-mail: chassang@princeton.edu.

1 Introduction

It is well known that randomized controlled trials in which treatment is assigned independently across patients can result in treatment and control groups whose observable characteristics are significantly different (Treasure and MacRae (1998), Bruhn and McKenzie (2009), Morgan and Rubin (2015), Banerjee et al. (2020)). Figure 1 illustrates the issue. Consider an experiment where we shuffle eight patients, four men and four women, into two equal-sized treatment arms. It is entirely possible that three women are assigned to arm 1 and three men are assigned to arm 2. As a result arm 1 is 75% female and arm 2 is 75% male.

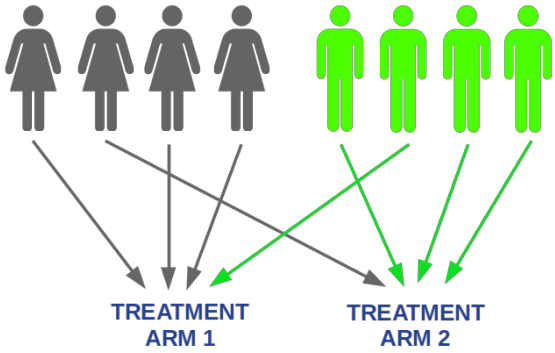

Figure 1: Random treatment assignment leading to equal-sized but unbalanced treatment arms.

Imbalance with respect to characteristics such as gender is problematic because it makes experiments less informative. When men and women have different medical outcomes on average, differences in mean outcomes between the two arms may be driven by imbalance with respect to gender, rather than by treatment efficacy. This is particularly problematic since medical databases such as clinicaltrials.gov report unconditional average treatment effects but not treatment effects conditional on gender.

Importantly, there exist simple and well understood random assignment procedures that ensure balance with respect to target characteristics (Bugni et al., 2018). Figure 2 illustrates

a matched-pair design in which patients of each gender are matched in pairs (say by order of arrival), and one member of each pair is randomly assigned to each treatment arm. This guarantees that gender proportions are equal across both arms while maintaining appropriate randomization.

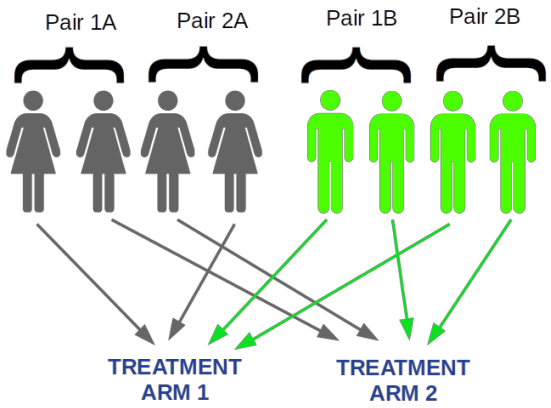

Figure 2: A matched-pair design; one member of each pair is randomly assigned to each treatment arm.

This paper has two objectives: (i) to assess the prevalence of balance issues in medical trials; (ii) to evaluate the informational benefits from adopting balanced experiment designs.

2 Data and Descriptive Statistics

We study the issue of balance with respect to gender in medical trials registered with the clinicaltrials.gov database. We focus on gender for three reasons: (i) it is systematically collected; (ii) it is likely to affect medical outcomes; (iii) experiment designs ensuring balance over binary characteristics are readily available, and well understood (Bugni et al., 2018). Naturally, similar concerns apply to other demographics, such as age, or ethnicity.

Source. We searched the clinicaltrials.gov database for all non-gender-specific interventional trials meeting the following joint conditions: (i) trials consisting of exactly two treat-

ment arms; (ii) trials reporting gender counts for both arms;[1] (iii) trials with exactly one primary outcome, reporting means and standard deviations, (iv) trials with least 20 patients. At the time of writing this paper, the total number of such studies is 2042.

Summary statistics. Roughly 42% of the studies in our universe were held in North America, while 35% were held Europe. Sample sizes range from 21 to more than 15 000. Such large sample sizes are rare: 58% of experiments have less than 100 patients, and 75% have less than 200 patients. These relatively small sample sizes explain why independent random assignment can generate unbalanced treatment arms. Both private for-profit and not-for-profit sponsors are well represented: 54% of studies have private non-profit lead sponsors, such as Mass. General Hospital, the Mayo Clinic, or Duke University; 40% of the studies have lead sponsors from industry, such as GlaxoSmithKline, Novartis, or Novo Nordisk. The remainder of studies in our sample have public sponsors, such as the NIH.

Documenting imbalance. Our main imbalance measure, Absolute Imbalance, is the magnitude of the difference between the share of men across the two treatment arms:

$$\text{Absolute Imbalance} = \left| \frac{\text{Number of Men in Arm 1}}{\text{Number of Patients in Arm 1}} - \frac{\text{Number of Men in Arm 2}}{\text{Number of Patients in Arm 2}} \right|.$$

We also report Relative Imbalance, which corresponds to the relative increase in the share of men from one treatment arm to the other:

$$\text{Relative Imbalance} = \frac{\text{Max Share of Men across Arms}}{\text{Min Share of Men across Arms}} - 1.$$

We find that medical trials frequently suffer from significant imbalance with respect to gender. Absolute Imbalance, i.e. the difference in the share of men across treatment arms, is greater than 5.1 percentage points for 50% of experiments, and greater than 10.5 percentage

[1] We use patient counts at the onset of the trial, before any potential attrition.

points for 25% of experiments.

Numbers for Relative imbalance are more expressive: 50% of experiments have at least 12% more men in one arm than the other; 25% of experiments have at least 26% more men in one arm than in the other. In fact, as Figure 3 shows, there is a long tail of experiments with large degrees of imbalance. For experiments with sample size less than 100 (which represents 58% of experiments in our sample), imbalance is even more prevalent: 25% have an Absolute Imbalance greater than 13.4 percentage points, and a Relative Imbalance greater than 34%.

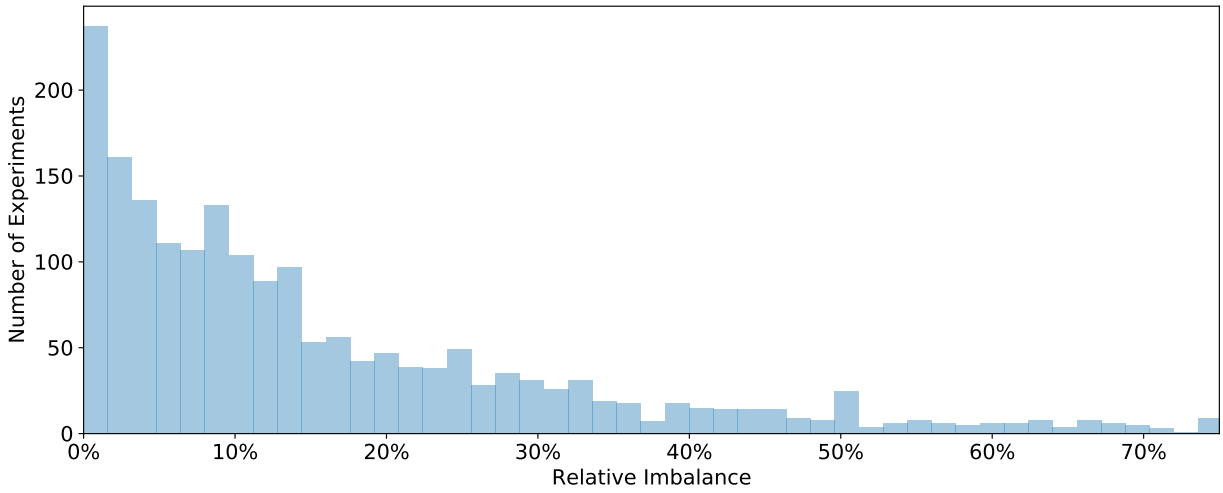

Figure 3: Distribution of Relative Imbalance.

Imbalance appears to be relatively lower for industry-run trials than non-profit-run trials: 25% of industry-run trials have Relative Imbalance greater than 23%; 25% of non-profit-run trials have Relative Imbalance greater that 31%. In addition, imbalance appears to be lower in Phase III than in Phase II trials: 25% of Phase III trials have a Relative Imbalance greater than 21%; 25% of Phase II trials have a Relative Imbalance greater than 29%.

Our data also lets us evaluate the likely share of experiments already using a balanced assignment protocol similar to the matched-pair design illustrated by Figure 2. Balanced designs guarantee that Absolute Imbalance will be essentially 0. Such low levels of Absolute

Imbalance are very unlikely if a balanced assignment procedure is not used. In our data, 13.3% of experiments have Absolute Imbalance under 1 percentage point. This provides a ballpark estimate of the number of experiments generated using an intentionally balanced design.

Effect size and imbalance. The significant degree of imbalance we document is only problematic if gender affects medical outcomes. We first provide anecdotal evidence that this is the case, before turning to a structural model specifying a relationship between imbalance, sample size, and effect dispersion. We define Effect Size as

$$\text{Effect Size} = \left| \frac{\text{Mean Outcome in Arm 1} - \text{Mean Outcome in Arm 2}}{\text{Standard Error of Outcomes}} \right|.$$

In words, the Effect Size is the absolute value of the estimated treatment effect, re-expressed in units of standard errors of the distribution of outcomes. We note that Effect Size is a convex function of estimated treatment effects. As a result, more noisily estimated treatment effects increases the expected Effect Size.[2]

Figures 4 and 5 show that greater Absolute Imbalance is associated with a greater effect size. Because Figure 4 is somewhat obscured by the density of points with low imbalance and low effect size, Figure 5 provides an easier-to-read summary statistic: the share of experiments whose Effect Size is greater than one standard-deviation.

A linear regression of Effect Size on Absolute Imbalance yields a significant positive coefficient equal to 1.1 (t-stat: 4.45, p-value<0.001, CI: [0.61, 1.57]).[3]

[2]We note that the findings illustrated by Figures 4 and 5 could be spuriously generated by omitted variable bias: smaller sample size is associated with both greater Absolute Imbalance and greater expected Effect Size. The structural model of treatment effects described in Section 3 addresses this issue by appropriately controlling for sample size.

[3]We remove outliers and focus on the sample of data such that: Absolute Imbalance is between 2 percentage points and 25 percentage points, Effect Size is below 3SD.

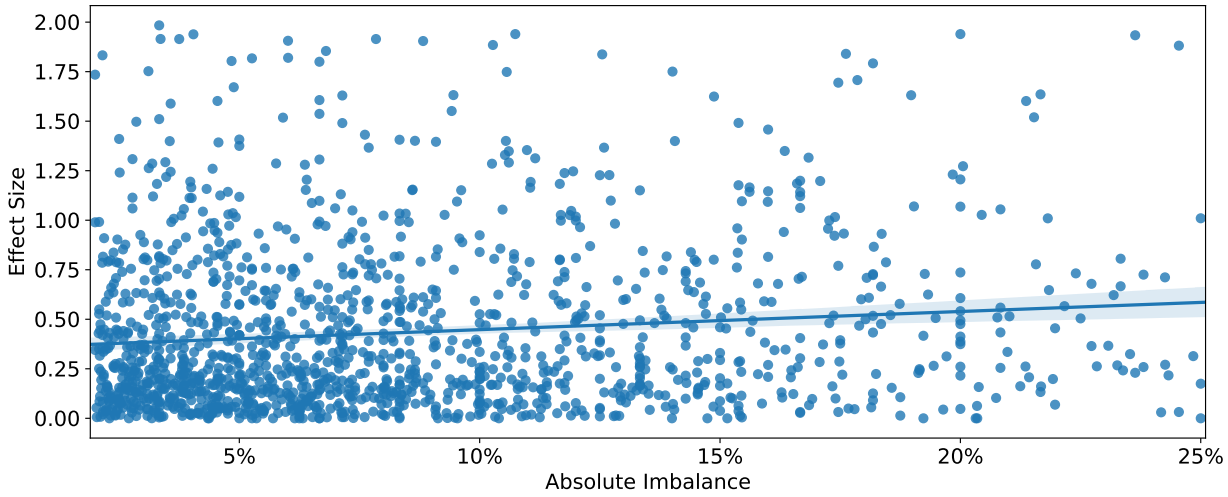

Figure 4: Greater Absolute Imbalance is associated with greater Effect Size.

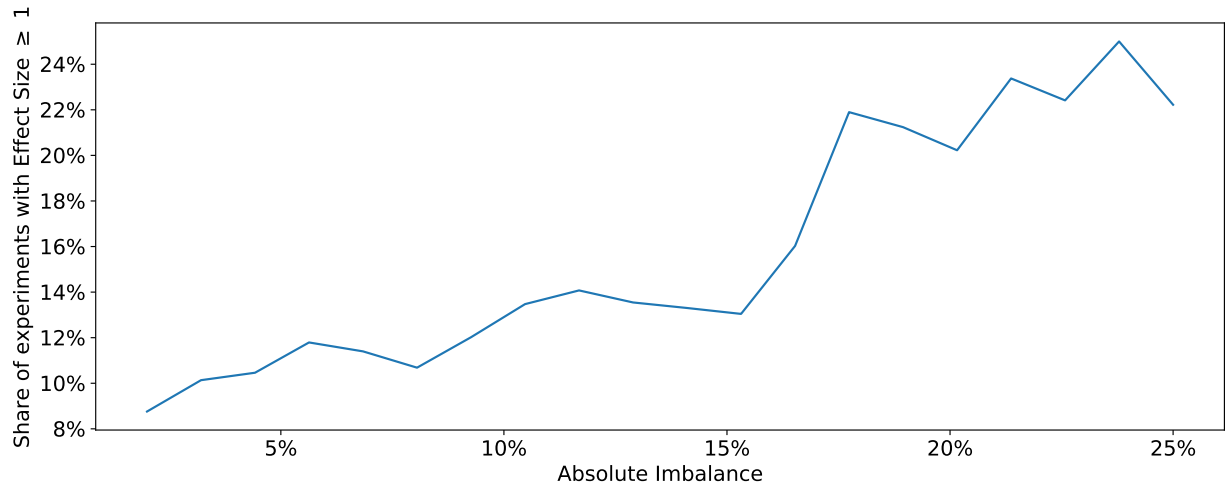

Figure 5: The share of experiments with Effect Size ≥ 1 SD increases with Absolute Imbalance.

3 Measuring the Cost of Imbalance

In this section, we derive a structural relationship between the dispersion of treatment effects, imbalance and sample size. This structural model allows us to express informational gains from using balanced designs in terms of increased sample size. We then estimate the model in our data and evaluate possible informational gains from using intentionally balanced designs.

3.1 A structural model of treatment effects

Treatment effects for a single experiment. Consider a given experiment consisting of a treatment and control arm. For any patient whose identity is denoted by i, let $Y_i \in \mathbb{R}$ denote the patient's outcome. Let $Gender_i \in \{0, 1\}$ denote their gender (with 0 corresponding to female, and 1 to male). Finally, let $Treatment_i \in \{0, 1\}$ denote treatment status, with 0 corresponding to the control group, and 1 to the treatment group.

We assume that for a given experiment, outcomes are related to treatment status and gender by a linear Gaussian model:

$$Y_i = scale \times (\alpha + \beta \times Treatment_i + \gamma \times Gender_i + \varepsilon_i) \tag{1}$$

with *scale* a positive scaling parameter (capturing among other things variation in units), α, β and γ constant parameters specific to the treatments and conditions being studied, and $\varepsilon_i \sim \mathcal{N}(0, \sigma)$ a normally distributed idiosyncratic error term.

Let N_0 and N_1 denote the number of patients respectively assigned to the control and treatment groups. Let $\overline{Y}_0$ and $\overline{Y}_1$ denote the average outcome for patients in the control and treatment groups. Finally, let $\overline{G}_0$ and $\overline{G}_1$ denote the respective share of men in the control and treatment groups. After averaging and taking differences, equation (1) implies that

$$\begin{aligned}\overline{Y}_1 - \overline{Y}_0 &= \frac{1}{N_1} \sum_{i \in \text{ Treated}} Y_i - \frac{1}{N_0} \sum_{j \in \text{Control}} Y_j \\ &= scale \times (\beta + \gamma \times (\overline{G}_1 - \overline{G}_0) + \Delta\varepsilon)\end{aligned}$$

with $\Delta\varepsilon$ a Gaussian error term with distribution $\mathcal{N}\left(0, \sqrt{\frac{\sigma^2}{N_1} + \frac{\sigma^2}{N_0}}\right)$

Given parameters *scale*, β, and γ, the variance of individual outcomes Y_i in the treatment and control groups are equal to $scale^2 \times \sigma^2$:

$$Var(Y_i | i \in \text{Treatment}) = scale^2 \times \sigma^2 = Var(Y_i | i \in \text{Control})$$

Hence the weighted average of variances

$$\sigma_U^2 \equiv \frac{N_1 Var(Y_i|i \in \text{Treatment}) + N_0 Var(Y_i|i \in \text{Control})}{N_1 + N_0}$$

is also equal to $scale^2 \times \sigma^2$. Note that σ_U can be estimated using the sample standard deviations of outcomes in each group. We define the standardized estimated Treatment Effect as

$$\text{Treatment Effect} \equiv \frac{\overline{Y}_1 - \overline{Y}_0}{\sigma_U}.$$

The Effect Size reported in Figures 4 and 5 is the absolute value of the Treatment Effect.

Note that this estimated Treatment Effect will typically be different from the true efficacy of treatment b. We have that

$$\text{Treatment Effect} = b + c \times (\overline{G}_1 - \overline{G}_0) + \Delta e \tag{2}$$

with Δe a Gaussian error term with distribution $\mathcal{N}\left(0, \sqrt{\frac{1}{N_1} + \frac{1}{N_0}}\right)$ and b and c parameters specific to the experiment, respectively equal to β/σ_U and γ/σ_U.

Treatment effects across experiments. We now specify a data generating process capturing the distribution of treatment effects in our population of clinical trials. We assume that coefficients b and c are normally distributed across experiments, following distributions $b \sim \mathcal{N}(0, \sigma_b)$ and $c \sim \mathcal{N}(0, \sigma_c)$. Parameters σ_b, and σ_c characterize the overall population. Given equation (2), this implies that

$$\mathbb{E}\left[\left(\frac{\overline{Y}_1 - \overline{Y}_0}{\sigma_U}\right)^2 \middle| \overline{G}_0, \overline{G}_1, N_0, N_1\right] = \sigma_b^2 + \sigma_c^2 \times (\overline{G}_1 - \overline{G}_0)^2 + \frac{1}{N_1} + \frac{1}{N_0}.$$

Expressed differently,

$$\text{Treatment Effect}^2 = \sigma_b^2 + \sigma_c^2 \times \text{Absolute Imbalance}^2 + \frac{1}{N_1} + \frac{1}{N_0} + error. \tag{3}$$

Define the de-biased square effect size (DSES) as

$$\text{DSES} \equiv \text{Treatment Effect}^2 - \frac{1}{N_1} - \frac{1}{N_0}.$$

Equation (3) implies that we can recover population parameters σ_b^2 and σ_c^2 by regressing DSES on Absolute Imbalance^2:

$$\text{DSES} = \sigma_b^2 + \sigma_c^2 \times \text{Absolute Imbalance}^2 + error. \tag{4}$$

Cost of imbalance. We measure the cost of imbalance by expressing the information gains from setting imbalance to 0 in terms of an equivalent increase in sample size. This additional sample size captures the cost of using an unbalanced random assignment instead of a balanced assignment such as a matched-pair design.

Given a true efficacy parameter $b \sim \mathcal{N}(0, \sigma_b)$, and a given level of Absolute Imbalance, equation (2) implies that the Treatment Effect of an experiment with equally sized arms (i.e. with sample sizes $N_0 = N_1 = N/2$) provides a signal of true efficacy b with distribution $\mathcal{N}(b, \sigma_{TE})$ where

$$\sigma_{TE}^2 = \sigma_c^2 \times \text{Absolute Imbalance}^2 + \frac{4}{N}. \tag{5}$$

The variance of the signal provided by the experiment consists of both an idiosyncratic noise term $4/N$ that depends only on the sample size, as wells as a term corresponding to the confounding effect of imbalance with respect to gender.

For a given value of Absolute Imbalance, consider an experiment with sample size N in which imbalance would be entirely removed. The corresponding signal has the same variance as the unbalanced experiment with sample size N' satisfying

$$\begin{aligned} \frac{4}{N} &= \sigma_c^2 \times \text{Absolute Imbalance}^2 + \frac{4}{N'} \\ \iff N' &= \frac{4}{4 - N \times \sigma_c^2 \times \text{Absolute Imbalance}^2} \times N. \end{aligned} \tag{6}$$

This formula lets us measure the informational losses from using unbalanced designs.

Note that this assessment of the informativeness of experiments assumes that inference relies only on estimated Treatment Effects (as in Bugni et al., 2018). Indeed, it is frequently the only statistic of outcomes reported in clinical trial databases. Using either treatment effects conditional on gender, or controlling for gender when estimating treatment effects would help reduce the confounding impact of gender imbalance on inference.

3.2 Empirical findings

The impact of imbalance on treatment effect dispersion. We now estimate the relationship between Effect Size and Absolute Imbalance using various specifications. We focus on the subsample of data such that Absolute Imbalance is greater than 2 percentage points (exactly balanced experiments are likely to use intentionally balanced designs, suggesting that they may be different from other experiments), and less than 25 percentage points. In addition, we remove outliers by focusing on experiments with Effect Size less than 3 units of standard-deviation.

We first use ordinary least squares (OLS) to estimate the relationship between DSES and Absolute Imbalance^2 including the Inverse Sample Size as a control (specification 2), or not (specification 1). For greater robustness, we replicate specifications 1 and 2, but estimating the median of DSES. This specification is less sensitive to outliers.

Table 1 shows that under all specifications, Absolute Imbalance^2 has a positive and significant impact on DSES. The lowest estimate for coefficient σ_c^2 is equal to 1.24, corresponding to a value $\sigma_c \simeq 1.1$. This is the value of σ_c we retain to assess the value of balanced designs.

The cost of imbalance. The estimated parameter value $\sigma_c = 1.1$ and equation (6) let us estimate the potential gains from using balanced designs in a typical experiment from our sample of clinical trials. Set $N = 100$ (roughly corresponding to the median sample size of experiments). For experiments with sample size between 75 and 125 subjects, the mean

	OLS		Median	
	(1)	(2)	(3)	(4)
	DSES	DSES	DSES	DSES
Intercept	0.35	0.27	0.03	0.02
	(0.035)	(0.047)	(0.007)	(0.009)
	[<0.001]	[<0.001]	[<0.001]	[0.008]
Absolute Imbalance^2	6.91	5.03	1.66	1.24
	(2.22)	(2.36)	(0.44)	(0.48)
	[0.002]	[0.03]	[<0.001]	[0.008]
Inverse Sample Size	—	5.67	—	0.57
		(2.43)		(0.48)
		[0.019]		[0.237]
Number of Obs.	1466	1466	1466	1466
R^2	.007	.01	0.001	0.001

Table 1: The impact of Absolute Imbalance on Effect Size. Standard errors are given in parentheses, and p-values in brackets.

Absolute Imbalance is equal to 7.1 percentage points. Plugging-in these values in equation (6) yields $N' = 118$. In other terms, addressing balance issues (say using a matched-pair design) yields an increase in the informativeness of experiments equivalent to an 18% increase in sample size.

4 Conclusion

Imbalance with respect to gender is prevalent in medical trials and significantly reduces the informativeness of experiments. Better experiment designs would result in speedier trials, more reliable findings, and greater consistency across Phase II and Phase III trials.

The rarity of balanced random designs in practice is puzzling. One explanation may be lack of awareness from experimenters. Another explanation is that experimenters worry that regulators may view balanced designs as a deviation from the "gold standard." Indeed,

the European Medicines Agency (EMEA (2004)) cautions against the use of sophisticated methodologies attempting to achieve balance across many continuous characteristics. However, seeking balance on an important binary characteristic such as gender seems uncontroversial. In order to facilitate the adoption of sensible balanced experiment designs, regulators should clarify their position on this issue.

References

Banerjee, A. V., S. Chassang, S. Montero, and E. Snowberg (2020): "A Theory of Experimenters: Robustness, Randomization, and Balance," *American Economic Review*.

Bruhn, M. and D. McKenzie (2009): "In pursuit of balance: Randomization in practice in development field experiments," *American Economic Journal: Applied Economics*, 1, 200–232.

Bugni, F. A., I. A. Canay, and A. M. Shaikh (2018): "Inference under covariate-adaptive randomization," *Journal of the American Statistical Association*, 113, 1784–1796.

EMEA (2004): "Point to consider on adjustment for baseline covariates," *Statistics in Medicine*, 23, 701–709.

Morgan, K. L. and D. B. Rubin (2015): "Rerandomization to balance tiers of covariates," *Journal of the American Statistical Association*, 110, 1412–1421.

Treasure, T. and K. D. MacRae (1998): "Minimisation: the platinum standard for trials? Randomisation doesn't guarantee similarity of groups; minimisation does." *British Medical Journal*, 317, 362–363.