Francis I. Proctor Foundation Guide to Randomization

Contributors: Ben Arnold

Compiled: 2021-02-26

Contents

Welcome!

This is a guide for best (essential!) practices in trial randomization and masking. At the Proctor Foundation, we lead many randomized, controlled trials to study intervention effects. Many of our trials are masked (aka "blinded"), whereby treatment allocation is concealed from participants, investigators, and/or outcome assessors. This short guide is a compendium our team's best practices around these activities.

IMPORTANT: The randomization and masking steps are among the most important activities to ensure a trial's validity. Jeopardizing one or both can undermine a trial. Team members involved in generating an allocation sequence and masking a trial should work directly with at least one of our faculty biostatisticians. At present, those faculty members include Ben Arnold and Travis Porco.

Chapter 1

Randomization

Contributors: Ben Arnold

1.1 Introduction

Random allocation of treatment to units (individuals or clusters) is perhaps the single strongest design tool we have in epidemiology and clinical research to estimate the causal effect of a treatment on outcomes. Randomization ensures that individuals who receive treatment are, on average, exchangeable with those who do not (Altman and Bland, 1999). Without randomization, individuals who seek or receive treatment are almost inevitably different from those who do not, often in immeasurable ways. This leads to confounding of the treatment-outcome relationship.

Importantly (Altman and Bland, 1999):

> "The term random does not mean the same as haphazard but has a precise technical meaning. By random allocation we mean that each patient has a known chance, usually an equal chance, of being given each treatment, but the treatment to be given cannot be predicted."

Below, we provide guidance on how to generate random sequences.

Many trials are masked, where the treatment group is kept secret from participants and/or investigators. The next chapter focuses on best practices for masking.

1.2 Randomization Overview

In practice, generating an allocation sequence involves the following steps:

Table 1.1: Randomization Steps

Steps	
1	Finalize the study design and randomization plan, including specifics about allocation ratio, any blocking/stratification, and masking.
2	Create a randomization subdirectory within the trial's project directory to save the randomization files. If randomization is masked, you will need to save the randomization files in a separate, tightly controlled directory (sync'd to the cloud for secure backup).
3	Write a script to generate a random sequence. If the trial is masked, use temporary letters.
4	Assess randomization diagnostics to ensure that the randomization sequence behaves as expected.
5	Share the randomization sequence and diagnostics with the PI and trial's biostatistician. Have at least 2 people review the randomization script and diagnostics to check for any errors.
6	If the trial is masked, work with the trial's unmasked biostatistician to assign the final letters to each treatment group using the agreed upon, private mapping between letters and treatment group.
7	Set a new seed, and generate the final sequence. Store the sequence in a `.csv` file in the randomization directory and lock the file to ensure it cannot be over-written.

1.3 A few comments on computing

Like all of our data science workflows, generating a random sequence needs to be transparent and reproducible. See Proctor's handbook on data science. The principles we describe there are germane for randomization sequences as well!

https://proctor-ucsf.github.io/dcc-handbook/intro.html

All of the examples in this guide use R software. There are surely many other effective ways to generate sequences in other software, but R includes many convenient functions for pseudo-random number generation.

Our advice for generating the sequence is to not rely on any packages beyond base `R`. The R language and packages evolve rapidly. Using base R ensures that functions will behave consistently over time. For example, many tidyverse packages such as the `dplyr` package are incredible for data manipulation. `dplyr` includes many convenient pseudo-random sampling routines, but as of this writing its syntax rapidly evolves – some commonly used functions seem to

be replaced or deprecated every few months. This makes the code more fragile. As an exception, we do use `ggplot2` for graphics in randomization diagnostics examples below.

1.4 Creating a Randomization Directory

The files used to generate a randomization sequence should live in one directory. They should be clearly labeled. They should include lots of comments and documentation to orient a new reader to their contents. The directory should include, at minimum:

Table 1.2: Checklist for the Randomization Directory

	Checklist
___ 1.	A metadata README file that describes all files in the directory.
___ 2.	The script that generates the randomization sequence.
___ 3.	Randomization diagnostics (typically an output file from the script).
___ 4.	The randomization sequence(s) generated, stored as `.csv` files.
___ 5.	The key that maps group labels to masked codes (masked trials only)

For unmasked randomization, we recommend creating a subdirectory within a trial's parent directory called `Randomization`. In the hypothetical `MyTrial` study below, create a new subdirectory nested within:

```
~/Box Sync/MyTrial/Randomization
```

If the randomization sequence is masked, then we need to keep the randomization files separate from the main trial directory. Otherwise team members who should not know the mapping between treatment labels and masking codes could discover the link. The sequence could also be discoverable, and if a team member is involved in treating patients that could bias the allocation.

In masked trials, we recommend creating a parallel, shadow directory for the trial with restricted access permissions. We typically use the suffix `-unmasked-materials` to identify the restricted access directories. They should live on the same encrypted server as the trial's main directory, e.g.:

```
~/Box Sync/MyTrial-unmasked-materials/Randomization
```

Note that an `-unmasked-materials` directory can contain other, sensitive, restricted access materials. Such as: interim analyses requested by the trial's Data and Safety Monitoring Committee.

Table 1.3: Key Points for Restricted Access Directories in Masked Trials

Key Points	
1	The trial's biostatistician should control access to a trial's directory of unmasked materials
2	At least 3 team members should have access to unmasked materials at all times

1.5 Unrestricted randomization

1.6 Blocked randomization

1.7 Stratified randomization

1.8 Randomization diagnostics

References

Bibliography

Altman, D. G. and Bland, M. J. (1999). Treatment allocation in controlled trials: why randomise? *BMJ*, 318(7192):1209–1209.