Blinding and Randomization

Anton Bespalov, Karsten Wicke, and Vincent Castagné

Contents

Abstract

Most, if not all, guidelines, recommendations, and other texts on Good Research Practice emphasize the importance of blinding and randomization. There is, however, very limited specific guidance on when and how to apply blinding and randomization. This chapter aims to disambiguate these two terms by discussing what they mean, why they are applied, and how to conduct the acts of randomization and blinding. We discuss the use of blinding and randomization as the means against existing and potential risks of bias rather than a mandatory practice that is to be followed under all circumstances and at any cost. We argue

A. Bespalov (✉)
Partnership for Assessment and Accreditation of Scientific Practice, Heidelberg, Germany

Pavlov Medical University, St. Petersburg, Russia
e-mail: anton.bespalov@paasp.net

K. Wicke
AbbVie, Ludwigshafen, Germany

V. Castagné
Porsolt, Le Genest-Saint-Isle, France

A. Bespalov et al. (eds.), *Good Research Practice in Non-Clinical Pharmacology and Biomedicine*, Handbook of Experimental Pharmacology 257,
https://doi.org/10.1007/164_2019_279

that, in general, experiments should be blinded and randomized if (a) this is a confirmatory research that has a major impact on decision-making and that cannot be readily repeated (for ethical or resource-related reasons) and/or (b) no other measures can be applied to protect against existing and potential risks of bias.

Keywords
Good Research Practice · Research rigor · Risks of bias

> ‘When I use a word,’ Humpty Dumpty said in rather a scornful tone, ‘it means just what I choose it to mean – neither more nor less.’
>
> Lewis Carroll (1871)
>
> *Through the Looking-Glass, and What Alice Found There*

1 Randomization and Blinding: Need for Disambiguation

In various fields of science, outcome of the experiments can be intentionally or unintentionally distorted if potential sources of bias are not properly controlled. There is a number of recognized risks of bias such as selection bias, performance bias, detection bias, attrition bias, etc. (Hooijmans et al. 2014). Some sources of bias can be efficiently controlled through research rigor measures such as randomization and blinding.

Existing guidelines and recommendations assign a significant value to adequate control over various factors that can bias the outcome of scientific experiments (chapter “Guidelines and Initiatives for Good Research Practice”). Among internal validity criteria, randomization and blinding are two commonly recognized bias-reducing instruments that need to be considered when planning a study and are to be reported when the study results are disclosed in a scientific publication.

For example, editorial policy of the Nature journals requires authors in the life sciences field to submit a checklist along with the manuscripts to be reviewed. This checklist has a list of items including questions on randomization and blinding. More specifically, for randomization, the checklist is asking for the following information: “If a method of randomization was used to determine how samples/animals were allocated to experimental groups and processed, describe it.” Recent analysis by the NPQIP Collaborative group indicated that only 11.2% of analyzed publications disclosed which method of randomization was used to determine how samples or animals were allocated to experimental groups (Macleod, The NPQIP Collaborative Group 2017). Meanwhile, the proportion of studies mentioning randomization was much higher – 64.2%. Do these numbers suggest that authors strongly motivated to have their work published in a highly prestigious scientific journal ignore the instructions? It is more likely that, for many scientists (authors, editors, reviewers), a statement such as “subjects were randomly assigned to one of the N treatment conditions” is considered to be sufficient to describe the randomization procedure.

Table 1 Example of an allocation schedule that is a pseudo-randomization

Group A	Group B	Group C	Group D
Mouse 1	Mouse 2	Mouse 3	Mouse 4
Mouse 5	Mouse 6	Mouse 7	Mouse 8
Mouse 9	Mouse 10	Mouse 11	Mouse 12
Mouse 13	Mouse 14	Mouse 15	Mouse 16

For the field of life sciences, and drug discovery in particular, the discussion of sources of bias, their impact, and protective measures, to a large extent, follows the examples from the clinical research (chapter "Learning from Principles of Evidence-Based Medicine to Optimize Nonclinical Research Practices"). However, clinical research is typically conducted by research teams that are larger than those involved in basic and applied preclinical work. In the clinical research teams, there are professionals (including statisticians) trained to design the experiments and apply bias-reducing measures such as randomization and blinding. In contrast, preclinical experiments are often designed, conducted, analyzed, and reported by scientists lacking training or access to information and specialized resources necessary for proper administration of bias-reducing measures.

As a result, researchers may design and apply procedures that reflect *their* understanding of what randomization and blinding are. These may or may not be the correct procedures. For example, driven by a good intention to randomize 4 different treatment conditions (A, B, C, and D) applied a group of 16 mice, a scientist may design the experiment in the following way (Table 1).

The above example is a fairly common practice to conduct "randomization" in a simple and convenient way. Another example of common practice is, upon animals' arrival, to pick them haphazardly up from the supplier's transport box and place into two (or more) cages which then constitute the control and experimental group(s). However, both methods of assigning subjects to experimental treatment conditions violate the randomness principle (see below) and, therefore, should not be reported as randomization.

Similarly, the use of blinding in experimental work typically cannot be described solely by stating that "experimenters were blinded to the treatment conditions." For both randomization and blinding, it is essential to provide details on what exactly was applied and how.

The purpose of this chapter is to disambiguate these two terms by discussing what they mean, why they are applied, and how to conduct the acts of randomization and blinding. We discuss the use of blinding and randomization as the means against existing and potential risks of bias rather than a mandatory practice that is to be followed under all circumstances and at any cost.

2 Randomization

Randomization can serve several purposes that need to be recognized individually as one or more of them may become critical when considering study designs and conditions exempt from the randomization recommendation.

First, randomization permits the use of probability theory to express the likelihood of chance as a source for the difference between outcomes. In other words, randomization enables the application of statistical tests that are common in biology and pharmacology research. For example, the central limit theorem states that the sampling distribution of the mean of any independent, random variable will be normal or close to normal, if the sample size is large enough. The central limit theorem assumes that the data are sampled randomly and that the sample values are independent of each other (i.e., occurrence of one event has no influence on the next event). Usually, if we know that subjects or items were selected randomly, we can assume that the independence assumption is met. If the study results are to be subjected to conventional statistical analyses dependent on such assumptions, adequate randomization method becomes a must.

Second, randomization helps to prevent a potential impact of the selection bias due to differing baseline or confounding characteristics of the subjects. In other words, randomization is expected to transform any systematic effects of an uncontrolled factor into a random, experimental noise. A random sample is one selected without bias: therefore, the characteristics of the sample should not differ in any systematic or consistent way from the population from which the sample was drawn. But random sampling does not guarantee that a particular sample will be exactly representative of a population. Some random samples will be more representative of the population than others. Random sampling does ensure, however, that, with a sufficiently large number of subjects, the sample becomes more representative of the population.

There are characteristics of the subjects that can be readily assessed and controlled (e.g., by using stratified randomization, see below). But there are certainly characteristics that are not known and for which randomization is the only way to control their potentially confounding influence. It should be noted, however, that the impact of randomization can be limited when the sample size is low.[1] This needs to be kept in mind given that most nonclinical studies are conducted using small sample sizes. Thus, when designing nonclinical studies, one should invest extra efforts into analysis of possible confounding factors or characteristics in order to judge whether or not experimental and control groups are similar before the start of the experiment.

Third, randomization interacts with other means to reduce risks of bias. Most importantly, randomization is used together with blinding to conceal the allocation sequence. Without an adequate randomization procedure, efforts to introduce and maintain blinding may not always be fully successful.

2.1 Varieties of Randomization

There are several randomization methods that can be applied to study designs of differing complexities. The tools used to apply these methods range from random

[1] https://stats.stackexchange.com/questions/74350/is-randomization-reliable-with-small-samples.

number tables to specialized software. Irrespective of the tools used, reporting on the randomization schedule applied should also answer the following two questions:

- Is the randomization schedule based on an algorithm or a principle that can be written down and, based on the description, be reapplied by anyone at a later time point resulting in the same group composition? If yes, we are most likely dealing with a "pseudo-randomization" (e.g., see below comments about the so-called Latin square design).
- Does the randomization schedule exclude any subjects and groups that belong to the experiment? If yes, one should be aware of the risks associated with excluding some groups or subjects such as a positive control group (see chapter "Out of Control? Managing Baseline Variability in Experimental Studies with Control Groups").

An answer "yes" to either of the above questions does not automatically mean that something incorrect or inappropriate is being done. In fact, a scientist may take a decision well justified by their experience with and need of particular experimental situation. However, in any case, the answer "yes" to either or both of the questions above mandates the complete and transparent description of the study design with the subject allocation schedule.

2.1.1 Simple Randomization

One of the common randomization strategies used for between-subject study designs is called simple (or unrestricted) randomization. Simple random sampling is defined as the process of selecting subjects from a population such that just the following two criteria are satisfied:

- The probability of assignment to any of the experimental groups is equal for each subject.
- The assignment of one subject to a group does not affect the assignment of any other subject to that same group.

With simple randomization, a single sequence of random values is used to guide assignment of subjects to groups. Simple randomization is easy to perform and can be done by anyone without a need to involve professional statistical help. However, simple randomization can be problematic for studies with small sample sizes. In the example below, 16 subjects had to be allocated to 4 treatment conditions. Using Microsoft Excel's function RANDBETWEEN (0.5;4.5), there were 16 random integer numbers from 1 to 4 generated. Obviously, this method has resulted in an unequal number of subjects among groups (e.g., there is only one subject assigned to group 2). This problem may occur irrespective of whether one uses machine-generated random numbers or simply tosses a coin.

Subject ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Group ID	4	1	1	3	3	1	4	4	3	4	3	3	4	2	3	1

An alternative approach would be to generate a list of all treatments to be administered (top row in the table below) and generate a list of random numbers (as many as the total number of subjects in a study) using a Microsoft Excel's function RAND() that returns random real numbers greater than or equal to 0 and less than 1 (this function requires no argument):

Treatment	1	1	1	1	2	2	2	2	3	3	3	3	4	4	4	4
Random number	0.76	0.59	0.51	0.90	0.64	0.10	0.50	0.48	0.22	0.37	0.05	0.09	0.73	0.83	0.50	0.43

The next step would be to sort the treatment row based on the values in the random number row (in an ascending or descending manner) and add a Subject ID row:

Subject ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Treatment	3	3	2	3	3	4	2	2	4	1	1	2	4	1	4	1
Random number	0.05	0.09	0.10	0.22	0.37	0.43	0.48	0.50	0.50	0.51	0.59	0.64	0.73	0.76	0.83	0.90

There is an equal number of subjects (four) assigned to each of the four treatment conditions, and the assignment is random. This method can also be used when group sizes are not equal (e.g., when a study is conducted with different numbers of genetically modified animals and animals of wild type).

However, such randomization schedule may still be problematic for some types of experiments. For example, if the subjects are tested one by one over the course of 1 day, the first few subjects could be tested in the morning hours while the last subjects – in the afternoon. In the example above, none of the first eight subjects is assigned to group 1, while the second half does not include any subject from group 3. To avoid such problems, block randomization may be applied.

2.1.2 Block Randomization

Blocking is used to supplement randomization in situations such as the one described above – when one or more external factors change or may change during the period when the experiment is run. Blocks are balanced with predetermined group assignments, which keeps the numbers of subjects in each group similar at all times. All blocks of one experiment have equal size, and each block represents all independent variables that are being studied in the experiment.

The first step in block randomization is to define the block size. The minimum block size is the number obtained by multiplying numbers of levels of all independent variables. For example, an experiment may compare the effects of a vehicle and three doses of a drug in male and female rats. The minimum block size in such case would be eight rats per block (i.e., 4 drug dose levels × 2 sexes). All subjects can be divided into N blocks of size X∗Y, where X is a number of groups or treatment conditions (i.e., 8 for the example given) and Y – number of subjects per treatment condition per block. In other words, there may be one or more subjects per treatment condition per block so that the actual block size is multiple of a minimum block size (i.e., 8, 16, 24, and so for the example given above).

The second step is, after block size has been determined, to identify all possible combinations of assignment within the block. For instance, if the study is evaluating effects of a drug (group A) or its vehicle (group B), the minimum block size is equal to 2. Thus, there are just two possible treatment allocations within a block: (1) AB and (2) BA. If the block size is equal to 4, there is a greater number of possible treatment allocations: (1) AABB, (2) BBAA, (3) ABAB, (4) BABA, (5) ABBA, and (6) BAAB.

The third step is to randomize these blocks with varying treatment allocations:

Block number	4	3	1	6	5	2
Random number	0.015	0.379	0.392	0.444	0.720	0.901

And, finally, the randomized blocks can be used to determine the subjects' assignment to the groups. In the example above, there are 6 blocks with 4 treatment conditions in each block, but this does not mean that the experiment must include 24 subjects. This random sequence of blocks can be applied to experiments with a total number of subjects smaller or greater than 24. Further, the total number of subjects does not have to be a multiple of 4 (block size) as in the example below with a total of 15 subjects:

Block number	4				3				1				6			
Random number	0.015				0.379				0.392				0.444			
Subject ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	–
Treatment	B	A	B	A	A	B	A	B	A	A	B	B	B	A	A	–

It is generally recommended to blind the block size to avoid any potential selection bias. Given the low sample sizes typical for preclinical research, this recommendation becomes a mandatory requirement at least for confirmatory experiments (see chapter "Resolving the Tension Between Exploration and Confirmation in Preclinical Biomedical Research").

2.1.3 Stratified Randomization

Simple and block randomization are well suited when the main objective is to balance the subjects' assignment to the treatment groups defined by the independent variables whose impact is to be studied in an experiment. With sample sizes that are large enough, simple and block randomization may also balance the treatment groups in terms of the unknown characteristics of the subjects. However, in many experiments, there are baseline characteristics of the subjects that do get measured and that may have an impact on the dependent (measured) variables (e.g., subjects' body weight). Potential impact of such characteristics may be addressed by specifying inclusion/exclusion criteria, by including them as covariates into a statistical analysis, and (or) may be minimized by applying stratified randomization schedules.

It is always up to a researcher to decide where there are such potentially impactful covariates that need to be controlled and what is the best way of dealing with them. In case of doubt, the rule of thumb is to avoid any risk, apply stratified randomization, and declare an intention to conduct a statistical analysis that will isolate a potential contribution of the covariate(s).

It is important to acknowledge that, in many cases, information about such covariates may not be available when a study is conceived and designed. Thus, a decision to take covariates into account often affects the timing of getting the randomization conducted. One common example of such a covariate is body weight. A study is planned, and sample size is estimated before the animals are ordered or bred, but the body weights will not be known until the animals are ready. Another example is the size of the tumors that are inoculated and grow at different rates for a pre-specified period of time before the subjects start to receive experimental treatments.

For most situations in preclinical research, an efficient way to conduct stratified randomization is to run simple (or block) randomization several times (e.g., 100 times) and, for each iteration, calculate means for the covariate per each group (e.g., body weights for groups A and B in the example in previous section). The randomization schedule that yields the lowest between-group difference for the covariate would then be chosen for the experiment. Running a large number of iterations does not mean saving excessively large volumes of data. In fact, several tools used to support randomization allow to save the seed for the random number generator and re-create the randomization schedule later using this seed value.

Although stratified randomization is a relatively simple technique that can be of great help, there are some limitations that need to be acknowledged. First, stratified randomization can be extended to two or more stratifying variables. However, given the typically small sample sizes of preclinical studies, it may become complicated to implement if many covariates must be controlled. Second, stratified randomization works only when all subjects have been identified before group assignment. While this is often not a problem in preclinical research, there may be situations when a large study sample is divided into smaller batches that are taken sequentially into the study. In such cases, more sophisticated procedures such as the covariate adaptive randomization may need to be applied similar to what is done in clinical research (Kalish and Begg 1985). With this method, subjects are assigned to treatment groups by taking into account the specific covariates and assignments of subjects that have already been allocated to treatment groups. We intentionally do not provide any further examples or guidance on such advanced randomization methods as they should preferably be developed and applied in consultation with or by biostatisticians.

Table 2 A Latin square design as a common example of a pseudo-randomization

Subject	Consecutive tests (or study periods)			
	1	2	3	4
#1	A	B	C	D
#2	B	C	D	A
#3	C	D	A	B
#4	D	A	B	C

2.1.4 The Case of Within-Subject Study Designs

The above discussion on the randomization schedules referred to study designs known as between-subject. A different approach would be required if a study is designed as within-subject. In such study designs also known as the crossover, subjects may be given sequences of treatments with the intent of studying the differences between the effects produced by individual treatments. One should keep in mind that such sequence of testing always bears the danger that the first test might affect the following ones. If there are reasons to expect such interference, within-subjects designs should be avoided.

In the simplest case of a crossover design, there are only two treatments and only two possible sequences to administer these treatments (e.g., A-B and B-A). In nonclinical research and, particularly, in pharmacological studies, there is a strong trend to include at least three doses of a test drug and its vehicle. A Latin square design is commonly used to allocate subjects to treatment conditions. Latin square is a very simple technique, but it is often applied in a way that does not result in a proper randomization (Table 2).

In this example, each subject receives each of the four treatments over four consecutive study periods, and, for any given study period, each treatment is equally represented. If there are more than four subjects participating in a study, then the above schedule is copied as many times as need to cover all study subjects.

Despite its apparent convenience (such schedules can be generated without any tools), resulting allocation schedules are predictable and, what is even worse, are not balanced with respect to first-order carry-over effects (e.g., except for the first test period, D comes always after C). Therefore, such Latin square designs are not an example of properly conducted randomization.

One solution would be to create a complete set of orthogonal Latin Squares. For example, when the number of treatments equals three, there are six (i.e., 3!) possible sequences – ABC, ACB, BAC, BCA, CAB, and CBA. If the sample size is a multiple of six, then all six sequences would be applied. As the preclinical studies typically involve small sample sizes, this approach becomes problematic for larger numbers of treatments such as 4, where there are already 24 (i.e., 4!) possible sequences.

The Williams design is a special case of a Latin square where every treatment follows every other treatment the same number of times (Table 3).

The Williams design maintains all the advantages of the Latin square but is balanced (see Jones and Kenward 2003 for a detailed discussion on the Williams squares including the generation algorithms). There are six Williams squares

Table 3 An example of a Williams design

	Consecutive tests (or study periods)			
Subject	1	2	3	4
#1	A	B	C	D
#2	B	D	A	C
#3	C	A	D	B
#4	D	C	B	A

possible in case of four treatments. Thus, if there are more than four subjects, more than one Williams square would be applied (e.g., two squares for eight subjects).

Constructing the Williams squares is not a randomization yet. In studies based on within-subject designs, subjects are not randomized to treatment in the same sense as they are in the between-subject design. For a within-subject design, the treatment sequences are randomized. In other words, after the Williams squares are constructed and selected, individual sequences are randomly assigned to the subjects.

2.2 Tools to Conduct Randomization

The most common and basic method of simple randomization is flipping a coin. For example, with two treatment groups (control versus treatment), the side of the coin (i.e., heads, control; tails, treatment) determines the assignment of each subject. Other similar methods include using a shuffled deck of cards (e.g., even, control; odd, treatment), throwing a dice (e.g., below and equal to 3, control; over 3, treatment), or writing numbers of pieces of paper, folding them, mixing, and then drawing one by one. A random number table found in a statistics book, online random number generators (random.org or randomizer.org), or computer-generated random numbers (e.g., using Microsoft Excel) can also be used for simple randomization of subjects. As explained above, simple randomization may result in an unbalanced design, and, therefore, one should pay attention to the number of subjects assigned to each treatment group. But more advanced randomization techniques may require dedicated tools and, whenever possible, should be supported by professional biostatisticians.

Randomization tools are typically included in study design software, and, for in vivo research, the most noteworthy example is the NC3Rs' Experimental Design Assistant (www.eda.nc3rs.org.uk). This freely available online resource allows to generate and share a spreadsheet with the randomized allocation report after the study has been designed (i.e., variables defined, sample size estimated, etc.). Similar functionality may be provided by Electronic Laboratory Notebooks that integrate study design support (see chapter "Electronic Lab Notebooks and Experimental Design Assistants").

Randomization is certainly supported by many data analysis software packages commonly used in research. In some cases, there is even a free tool that allows to conduct certain types of randomization online (e.g., QuickCalcs at www.graphpad.com/quickcalcs/randMenu/).

Someone interested to have a nearly unlimited freedom in designing and executing different types of randomization will benefit from the resources generated by the R community (see https://paasp.net/resource-center/r-scripts/). Besides being free and supported by a large community of experts, R allows to save the scripts used to obtain randomization schedules (along with the seed numbers) that makes the overall process not only reproducible and verifiable but also maximally transparent.

2.3 Randomization: Exceptions and Special Cases

Randomization is not and should never be seen as a goal per se. The goal is to minimize the risks of bias that may affect the design, conduct, and analysis of a study and to enable application of other research methods (e.g., certain statistical tests). Randomization is merely a tool to achieve this goal.

If not dictated by the needs of data analysis or the intention to implement blinding, in some cases, pseudo-randomizations such as the schedules described in Tables 1 and 2 may be sufficient. For example, animals delivered by a qualified animal supplier come from large batches where the breeding schemes themselves help to minimize the risk of systematic differences in baseline characteristics. This is in contrast to clinical research where human populations are generally much more heterogeneous than populations of animals typically used in research.

Randomization becomes mandatory in case animals are not received from major suppliers, are bred in-house, are not standard animals (i.e., transgenic), or when they are exposed to an intervention before the initiation of a treatment. Examples of intervention may be surgery, administration of a reagent substance inducing long-term effects, grafts, or infections. In these cases, animals should certainly be randomized after the intervention.

When planning a study, one should also consider the risk of between-subject cross-contamination that may affect the study outcome if animals receiving different treatment(s) are housed within the same cage. In such cases, the most optimal approach is to reduce the number of subjects per cage to a minimum that is acceptable from the animal care and use perspective and adjust the randomization schedule accordingly (i.e., so that all animals in the cage receive the same treatment).

There are situations when randomization becomes impractical or generates other significant risks that outweigh its benefits. In such cases, it is essential to recognize the reasons why randomization is applied (e.g., ability to apply certain statistical tests, prevention of selection bias, and support of blinding). For example, for an in vitro study with multi-well plates, randomization is usually technically possible, but one would need to recognize the risk of errors introduced during manual pipetting into a 96- or 384-well plate. With proper controls and machine-read experimental readout, the risk of bias in such case may not be seen as strong enough to accept the risk of a human error.

Another common example is provided by studies where incremental drug doses or concentrations are applied during the course of a single experiment involving just one subject. During cardiovascular safety studies, animals receive first an infusion of a vehicle (e.g., over a period of 30 min), followed by the two or three concentrations of the test drug, and the hemodynamics is being assessed along with the blood samples taken. As the goal of such studies is to establish concentration-effect relationships, one has no choice but to accept the lack of randomization. The only alternatives would be to give up on the within-subject design or conduct the study over many days to allow enough time to wash the drug out between the test days. Needless to say, neither of these options is perfect for a study where the baseline characteristics are a critical factor in keeping the sample size low. In this example, the desire to conduct a properly randomized study comes into a conflict with ethical considerations.

A similar design is often used in electrophysiological experiments (in vitro or ex vivo) where a test system needs to be equilibrated and baselined for extended periods of time (sometimes hours) to allow subsequent application of test drugs (at ascending concentrations). Because a washout cannot be easily controlled, such studies also do not follow randomized schedules of testing various drug doses.

The low-throughput studies such as in electrophysiology typically go over many days, and every day there is a small number of subjects or data points added. While one may accept the studies being not randomized in some cases, it is important to stress that there should be other measures in place that control potential sources of bias. It is a common but usually unacceptable practice to analyze the results each time a new data point has been added in order to decide whether a magic *P* value sank below 0.05 and the experiment can stop. For example, in one recent publication, it was stated: "For optogenetic activation experiments, cell-type-specific ablation experiments, and in vivo recordings (optrode recordings and calcium imaging), we continuously increased the number of animals until statistical significance was reached to support our conclusions." Such an approach should be avoided by clear experimental planning and definition of study endpoints.

The above examples are provided only to illustrate that there may be special cases when randomization may not be done. This is usually not an easy decision to make and even more difficult to defend later. Therefore, one should always be advised to seek a professional advice (i.e., interaction with the biostatisticians or colleagues specializing in the risk assessment and study design issues). Needless to say, this advice should be obtained before the studies are conducted.

In the ideal case, once the randomization was applied to allocate subjects to treatment conditions, the randomization should be maintained through the study conduct and analysis to control against potential performance and outcome detection bias, respectively. In other words, it would not be appropriate first to assign the subjects, for example, to groups A and B and then do all experimental manipulations first with the group A and then with the group B.

3 Blinding

In clinical research, blinding and randomization are recognized as the most important design techniques for avoiding bias (ICH Harmonised Tripartite Guideline 1998; see also chapter "Learning from Principles of Evidence-Based Medicine to Optimize

Nonclinical Research Practices"). In the preclinical domain, there is a number of instruments assessing risks of bias, and the criteria most often included are randomization and blinding (83% and 77% of a total number of 30 instruments analyzed, Krauth et al. 2013).

While randomization and blinding are often discussed together and serve highly overlapping objectives, attitude towards these two research rigor measures is strikingly different. The reason for a higher acceptance of randomization compared to blinding is obvious – randomization can be implemented essentially at no cost, while blinding requires at least some investment of resources and may therefore have a negative impact on the research unit's apparent capacity (measured by the number of completed studies, irrespective of quality).

Since the costs and resources are not an acceptable argument in discussions on ethical conduct of research, we often engage a defense mechanism, called rationalization, that helps to justify and explain why blinding should not be applied and do so in a seemingly rational or logical manner to avoid the true explanation. Arguments against the use of blinding can be divided into two groups.

One group comprises a range of factors that are essentially psychological barriers that can be effectively addressed. For example, one may believe that his/her research area or a specific research method has an innate immunity against any risk of bias. Or, alternatively, one may believe that his/her scientific excellence and the ability to supervise the activities in the lab make blinding unnecessary. There is a great example that can be used to illustrate that there is no place for beliefs and one should rather rely on empirical evidence. For decades, compared to male musicians, females have been underrepresented in major symphonic orchestras despite having equal access to high-quality education. The situation started to change in the mid-1970s when blind auditions were introduced and the proportion of female orchestrants went up (Goldin and Rouse 2000). In preclinical research, there are also examples of the impact of blinding (or a lack thereof). More specifically, there were studies that reveal substantially higher effect sizes reported in the experiments that were not randomized or blinded (Macleod et al. 2008).

Another potential barrier is related to the "trust" within the lab. Bench scientists need to be explained what the purpose of blinding is and, in the ideal case, be actively involved in development and implementation of blinding and other research rigor measures. With the proper explanation and engagement, blinding will not be seen as an unfriendly act whereby a PI or a lab head communicates a lack of trust.

The second group of arguments against the use of blinding is actually composed of legitimate questions that need to be addressed when designing an experiment. As mentioned above in the section on randomization, a decision to apply blinding should be justified by the needs of a specific experiment and correctly balanced against the existing and potential risks.

3.1 Fit-for-Purpose Blinding

It requires no explanation that, in preclinical research, there are no double-blinded studies in a sense of how it is meant in the clinic. However, similar to clinical research, blinding in preclinical experiments serves to protect against two potential sources of bias: bias related to blinding of personnel involved in study conduct including application of treatments (performance bias) and bias related to blinding of personnel involved in the outcome assessment (detection bias).

Analysis of the risks of bias in a particular research environment or for a specific experiment allows to decide which type of blinding should be applied and whether blinding is an appropriate measure against the risks.

There are three types or levels of blinding, and each one of them has its use: assumed blinding, partial blinding, and full blinding. With each type of blinding, experimenters allocate subjects to groups, replace the group names with blind codes, save the coding information in a secure place, and do not access this information until a certain pre-defined time point (e.g., until the data are collected or the study is completed and analyzed).

3.1.1 Assumed Blinding

In the assumed blinding, experimenters have access to the group or treatment codes at all times, but they do not know the correspondence between group and treatment before the end of the study. With the partial or full blinding, experimenters do not have access to the coding information until a certain pre-defined time point.

Main advantage of the assumed blinding is that an experiment can be conducted by one person who plans, performs, and analyzes the study. The risk of bias may be relatively low if the experiments are routine – e.g., lead optimization research in drug discovery or fee-for-service studies conducted using well-established standardized methods.

Efficiency of assumed blinding is enhanced if there is a sufficient time gap between application of a treatment and the outcome recording/assessment. It is also usually helpful if the access to the blinding codes is intentionally made more difficult (e.g., blinding codes are kept in the study design assistant or in a file on an office computer that is not too close to the lab where the outcomes will be recorded).

If introduced properly, assumed blinding can guard against certain unwanted practices such as remeasurement, removal, and reclassification of individual observations or data points (three evil Rs according to Shun-Shin and Francis 2013). In preclinical studies with small sample sizes, such practices have particularly deleterious consequences. In some cases, remeasurement even of a single subject may skew the results in a direction suggested by the knowledge of group allocation. One should emphasize that blinding is not necessarily an instrument against the remeasurement (it is often needed or unavoidable) but rather helps to avoid risks associated with it.

3.1.2 Partial Blinding

There are various situations where blinding (with no access to the blinding codes) is implemented not for the entire experiment but only for a certain part of it, e.g.:

- No blinding during the application of experimental treatment (e.g., injection of a test drug) but proper blinding during the data collection and analysis
- No blinding during the conduct of an experiment but proper blinding during analysis

For example, in behavioral pharmacology, there are experiments where subjects' behavior is video recorded after a test drug is applied. In such cases, blinding is applied to analysis of the video recordings but not the drug application phase. Needless to say, blinded analysis has typically to be performed by someone who was not involved in the drug application phase.

A decision to apply partial blinding is based on (a) the confidence that the risks of bias are properly controlled during the unblinded parts of the experiment and/or (b) rationale assessment of the risks associated with maintaining blinding throughout the experiment. As an illustration of such decision-making process, one may imagine a study where the experiment is conducted in a small lab (two or three people) by adequately trained personnel that is not under pressure to deliver results of a certain pattern, data collection is automatic, and data integrity is maintained at every step. Supported by various risk reduction measures, such an experiment may deliver robust and reliable data even if not fully blinded.

Importantly, while partial blinding can adequately limit the risk of some forms of bias, it may be less effective against the performance bias.

3.1.3 Full Blinding

For important decision-enabling studies (including confirmatory research, see chapter "Resolving the Tension Between Exploration and Confirmation in Preclinical Biomedical Research"), it is usually preferable to implement full blinding rather than to explain why it was not done and argue that all the risks were properly controlled.

It is particularly advisable to follow full blinding in the experiments that are for some reasons difficult to repeat. For example, these could be studies running over significant periods of time (e.g., many months) or studies using unique resources or studies that may not be repeated for ethical reasons. In such cases, it is more rational to apply full blinding rather than leave a chance that the results will be questioned on the ground of lacking research rigor.

As implied by the name, full blinding requires complete allocation concealment from the beginning until the end of the experiment. This requirement may translate into substantial costs of resources. In the ideal scenario, each study should be supported by at least three independent people responsible for:

- (De)coding, randomization
- Conduct of the experiment such as handling of the subjects and application of test drugs (outcome recording and assessment)
- (Outcome recording and assessment), final analysis

The main reason for separating conduct of the experiment and the final analysis is to protect against potential unintended unblinding (see below). If there is no risk of unblinding or it is not possible to have three independent people to support the blinding of an experiment, one may consider a single person responsible for every step from the conduct of the experiment to the final analysis. In other words, the study would be supported by two independent people responsible for:

- (De)coding, randomization
- Conduct of the experiment such as handling of the subjects and application of test drugs, outcome recording and assessment, and final analysis

3.2 Implementation of Blinding

Successful blinding is related to adequate randomization. This does not mean that they should always be performed in this sequence: first randomization and then blinding. In fact, the order may be reversed. For example, one may work with an offspring of the female rats that received experimental and control treatments while pregnant. As the litter size may differ substantially between the dams, randomization may be conducted after the pups are born, and this does not require allocation concealment to be broken.

The blinding procedure has to be carefully thought through. There are several factors that are listed below and that can turn a well-minded intention into a waste of resources.

First, blinding should as far as possible cover the entire experimental setup – i.e., all groups and subjects. There is an unacceptable practice to exclude positive controls from blinding that is often not justified by anything other than an intention to introduce a detection bias in order to reduce the risk of running an invalid experiment (i.e., an experiment where a positive control failed).

In some cases, positive controls cannot be administered by the same route or using the same pretreatment time as other groups. Typically, such a situation would require a separate negative (vehicle) control treated in the same way as the positive control group. Thus, the study is only partially blinded as the experimenter is able to identify the groups needed to "validate" the study (negative control and positive control groups) but remains blind to the exact nature of the treatment received by each of these two groups. For a better control over the risk of unblinding, one may apply a "double-dummy" approach where all animals receive the same number of administrations via the same routes and pretreatment times.

Second, experiments may be unintentionally unblinded. For example, drugs may have specific, easy to observe physicochemical characteristics, or drug treatments may change the appearance of the subjects or produce obvious adverse effects. Perhaps, even more common is the unblinding due to the differences in the appearance of the drug solution or suspension dependent on the concentration. In such cases, there is not much that can be done but it is essential to take corresponding notes and acknowledge in the study report or publication. It is interesting to note that the unblinding is often cited as an argument against the use of blinding (Fitzpatrick et al. 2018); however, this argument reveals another problem – partial blinding schemes are often applied as a normative response without any proper risk of bias assessment.

Third, blinding codes should be kept in a secure place avoiding any risk that the codes are lost. For in vivo experiments, this is an ethical requirement as the study will be wasted if it cannot be unblinded at the end.

Fourth, blinding can significantly increase the risk of mistakes. A particular situation that one should be prepared to avoid is related to lack of accessibility of blinding codes in case of emergency. There are situations when a scientist conducting a study falls ill and the treatment schedules or outcome assessment protocols are not available or a drug treatment is causing disturbing adverse effects and attending veterinarians or caregivers call for a decision in the absence of a scientist responsible for a study. It usually helps to make the right decision if it is known that an adverse effect is observed in a treatment group where it can be expected. Such situations should be foreseen and appropriate guidance made available to anyone directly or indirectly involved in an experiment. A proper study design should define a backup person with access to the blinding codes and include clear definition of endpoints.

Several practical tips can help to reduce the risk of human-made mistakes. For example, the study conduct can be greatly facilitated if each treatment group is assigned its own color. Then, this color coding would be applied to vials with the test drugs, syringes used to apply the drug, and the subjects (e.g., apply solution from a green-labeled vial using a green-labeled syringe to an animal from a green-labeled cage or with a green mark on its tail). When following such practice, one should not forget to randomly assign color codes to treatment conditions. Otherwise, for example, yellow color is always used for vehicle control, green for the lowest dose, and so forth.

To sum up, it is not always lacking resources that make full blinding not possible to apply. Further, similar to what was described above for randomization, there are clear exception cases where application of blinding is made problematic by the very nature of the experiment itself.

4 Concluding Recommendations

Most, if not all, guidelines, recommendations, and other texts on Good Research Practice emphasize the importance of blinding and randomization (chapters "Guidelines and Initiatives for Good Research Practice", and "General Principles of Preclinical Study Design"). There is, however, very limited specific guidance on when and how to apply blinding and randomization. The present chapter aims to close this gap.

Generally speaking, experiments should be blinded and randomized if:

- This is a confirmatory research (see chapter "Resolving the Tension Between Exploration and Confirmation in Preclinical Biomedical Research") that has a

major impact on decision-making and that cannot be readily repeated (for ethical or resource-related reasons).
- No other measures can be applied to protect against existing and potential risks of bias.

There are various sources of bias that affect the outcome of experimental studies and these sources are unique and specific to each research unit. There is usually no one who knows these risks better than the scientists working in the research unit, and it is always up to the scientist to decide if, when, and how blinding and randomization should be implemented. However, there are several recommendations that can help to decide and act in the most effective way:

- Conduct a risk assessment for your research environment, and, if you do not know how to do that, ask for a professional support or advice.
- Involve your team in developing and implementing the blinding/randomization protocols, and seek the team members' feedback regarding the performance of these protocols (and revise them, as needed).
- Provide training not only on how to administer blinding and randomization but also to preempt any questions related to the rationale behind these measures (i.e., experiments are blinded not because of the suspected misconduct or lack of trust).
- Describe blinding and randomization procedures in dedicated protocols with as many details as possible (including emergency plans and accident reporting, as discussed above).
- Ensure maximal transparency when reporting blinding and randomization (e.g., in a publication). When deciding to apply blinding and randomization, be maximally clear about the details (Table 4). When deciding against, be open about the reasons for such decision. Transparency is also essential when conducting multi-laboratory collaborative projects or when a study is outsourced to another laboratory. To avoid any misunderstanding, collaborators should specify expectations and reach alignment on study design prior to the experiment and communicate all important details in study reports.

Blinding and randomization should always be a part of a more general effort to introduce and maintain research rigor. Just as the randomization increases the likelihood that blinding will not be omitted (van der Worp et al. 2010), other Good Research Practices such as proper documentation are also highly instrumental in making blinding and randomization effective.

To conclude, blinding and randomization may be associated with some effort and additional costs, but, under all circumstances, a decision to apply these research rigor techniques should not be based on general statements and arguments by those who do not want to leave their comfort zone. Instead, the decision should be based on the applicable risk assessment and careful review of potential implementation burden. In many cases, this leads to a relieving discovery that the devil is not so black as he is painted.

Table 4 Minimum reporting information for blinding and randomization procedures

Procedure	Technical report/laboratory notebook record	Scientific publication
Randomization	Type of randomization Block size (if applicable) Stratification variables (if applicable) Tools used for randomization Reference to the protocol followed Deviations from the protocol (if any)	Type of randomization Tools used for randomization Stratification variables (if applicable)
Blinding	Type of blinding Records of unblinding (if applicable) Reference to the protocol followed Deviations from the protocol (if any) Colleague(s) who provided blinding	Type of blinding Statement whether blinding integrity was maintained Statement whether blinding was provided by one of the co-authors

Acknowledgments The authors would like to thank Dr. Thomas Steckler (Janssen), Dr. Kim Wever (Radboud University), and Dr. Jan Vollert (Imperial College London) for reading the earlier version of the manuscript and providing comments and suggestions.

References

Carroll L (1871) Through the looking-glass, and what Alice found there. ICU Publishing

Fitzpatrick BG, Koustova E, Wang Y (2018) Getting personal with the “reproducibility crisis”: interviews in the animal research community. Lab Anim 47:175–177

Goldin C, Rouse C (2000) Orchestrating impartiality: the impact of “blind” auditions on female musicians. Am Econ Rev 90:715–741

Hooijmans CR, Rovers MM, de Vries RB, Leenaars M, Ritskes-Hoitinga M, Langendam MW (2014) SYRCLE’s risk of bias tool for animal studies. BMC Med Res Methodol 14:43

ICH Harmonised Tripartite Guideline (1998) Statistical principles for clinical trials (E9). CPMP/ICH/363/96, March 1998

Jones B, Kenward MG (2003) Design and analysis of cross-over designs, 2nd edn. Chapman and Hall, London

Kalish LA, Begg GB (1985) Treatment allocation methods in clinical trials a review. Stat Med 4:129–144

Krauth D, Woodruff TJ, Bero L (2013) Instruments for assessing risk of bias and other methodological criteria of published animal studies: a systematic review. Environ Health Perspect 121:985–992

Macleod MR, The NPQIP Collaborative Group (2017) Findings of a retrospective, controlled cohort study of the impact of a change in Nature journals’ editorial policy for life sciences research on the completeness of reporting study design and execution. bioRxiv:187245. https://doi.org/10.1101/187245

Macleod MR, van der Worp HB, Sena ES, Howells DW, Dirnagl U, Donnan GA (2008) Evidence for the efficacy of NXY-059 in experimental focal cerebral ischaemia is confounded by study quality. Stroke 39:2824–2829

Shun-Shin MJ, Francis DP (2013) Why even more clinical research studies may be false: effect of asymmetrical handling of clinically unexpected values. PLoS One 8(6):e65323

van der Worp HB, Howells DW, Sena ES, Porritt MJ, Rewell S, O'Collins V, Macleod MR (2010) Can animal models of disease reliably inform human studies? PLoS Med 7(3):e1000245