



# Out of Control? Managing Baseline Variability in Experimental Studies with Control Groups

Paul Moser

## Contents

1	What Are Control Groups?	102
2	Basic Considerations for Control Groups	103
2.1	Attribution of Animals to Control Groups	103
2.2	What Group Size for Control Groups?	103
2.3	Controls and Blinding	105
3	Primary Controls	106
3.1	Choosing Appropriate Control Treatments: Not All Negative Controls Are Equal ...	106
3.2	Vehicle Controls	107
3.3	Sham Controls	108
3.4	Non-neutral Control Groups	109
3.5	Controls for Mutant, Transgenic and Knockout Animals	109
4	Positive Controls	110
5	Secondary Controls	112
5.1	Can Baseline Values Be Used as Control?	112
5.2	Historical Control Values	113
6	When Are Control Groups Not Necessary?	114
7	Conclusion	114
	References	114

## Abstract

Control groups are expected to show what happens in the absence of the intervention of interest (negative control) or the effect of an intervention expected to have an effect (positive control). Although they usually give results we can anticipate, they are an essential component of all experiments, both in vitro and in vivo, and fulfil a number of important roles in any experimental design. Perhaps most importantly they help you understand the influence of variables that you cannot fully eliminate from your experiment and thus include them in

---

P. Moser (✉)  
Cerbascience, Toulouse, France  
e-mail: [pmoser@cerbascience.com](mailto:pmoser@cerbascience.com)

your analysis of treatment effects. Because of this it is essential that they are treated as any other experimental group in terms of subjects, randomisation, blinding, etc. It also means that in almost all cases, contemporaneous control groups are required. Historical and baseline control groups serve a slightly different role and cannot fully replace control groups run as an integral part of the experiment. When used correctly, a good control group not only validates your experiment; it provides the basis for evaluating the effect of your treatments.

---

**Keywords**

Baseline values · Blinding · Historical controls · Negative control groups · Positive control groups · Sham controls · Vehicle

---

## 1 What Are Control Groups?

As Donald Rumsfeld famously said about weapons of mass destruction, there are known knowns, known unknowns and unknown unknowns. This is also true of experiments. Through good experimental design, we try to eliminate as much as possible the influence of the first two, the things we know about. The purpose of the control group is to understand the influence of the third. The term sounds comforting, as if we have managed to somehow rein in the experiment and submitted the study to our will. If anything, the opposite is true: the control group is a tacit acknowledgement, not only of all the things that we can't control but of those things that we are not even aware of, the unknown unknowns.

The choice of appropriate control groups is intimately tied to the aims of your study. To unambiguously demonstrate that your experimental treatment has (or has not) had an effect in your test system, there needs to be a value against which you can compare it. A good control group allows you to do this – a bad control group means you cannot make valid comparisons to evaluate the activity of your test condition and, even worse, means you may end up drawing invalid conclusions.

Several types of control groups have been described in the literature, including positive, negative, sham, vehicle and comparative (Johnson and Besselsen 2002). These can broadly be classed into negative and positive controls. Note that in many studies these terms are used very loosely and, as Kramer and Font (2017) rightly point out, a description of what the control group is being used for is better than a label such as positive or negative which might be misleading. What are generally referred to as negative controls include vehicle and sham groups and are expected to show what happens in the absence of the intervention of interest. These controls are necessary because all studies are open to unexpected effects.

In contrast, positive controls are expected to have an effect. They are used to show that the study can detect an effect in the desired direction and thus that the experimental protocol is sensitive to interventions expected to have an effect. They might also be used to show the magnitude of effect that is possible with an active substance. Positive controls are in a comparable position to your test treatment: they need a good negative control to be of any use. Also, just like test treatments, they need to be subject to the same randomisation and blinding procedures and must be included in the experimental design and analysis.

There should also be a distinction between what we could call primary and secondary controls. Both have a value but in different areas. A primary control is what we typically think of as a control, i.e. a group that undergoes all experimental procedures except for the variable being investigated and which are specific to the experimental question being studied. In contrast, a secondary control, such as historical control values, could be used to check the conformity of the experiment or, in the case of baseline data, could be used to verify the homogeneity of the treatment groups at the start of the experiment. In almost all cases, the presence of a primary control group is essential to be able draw valid conclusions from the study.

Although this review discusses primarily *in vivo* experiments, almost all the points discussed apply equally to *in vitro* experiments. In fact, *in vitro* studies are in the envious position where additional control groups can be added for little additional cost and usually without invoking the ethical questions that are important to consider for *in vivo* studies. In all other ways, *in vitro* studies require the same attention to experimental design as *in vivo* studies as far as blinding, randomisation and statistical analysis are concerned (Festing 2001). This should also apply to any negative and positive controls included in the study.

In this brief review, I will only discuss the use of prospective control groups, which are most appropriate for non-clinical studies. Retrospective controls, which are often used in epidemiological clinical studies, to study such things as drug use during pregnancy, substance abuse etc., and where a randomised clinical trial would be unethical (e.g. Andrade 2017; Szekér et al. 2017), raise other issues which are not particularly relevant to non-clinical studies and which will not be discussed here.

---

## 2 Basic Considerations for Control Groups

### 2.1 Attribution of Animals to Control Groups

Correct randomisation to avoid bias is a basic but essential part of any experimental design (see chapter “Blinding and Randomization”) that applies equally to control groups. Control groups should be treated like any other experimental group within an experiment. Subjects for the control groups must come from the same population as the other groups so that the experiment is carried out on a homogenous population. This means that normally they should not be historical values nor baseline values. As we will see below, there are specific circumstances where this is either not possible or where some flexibility is permitted.

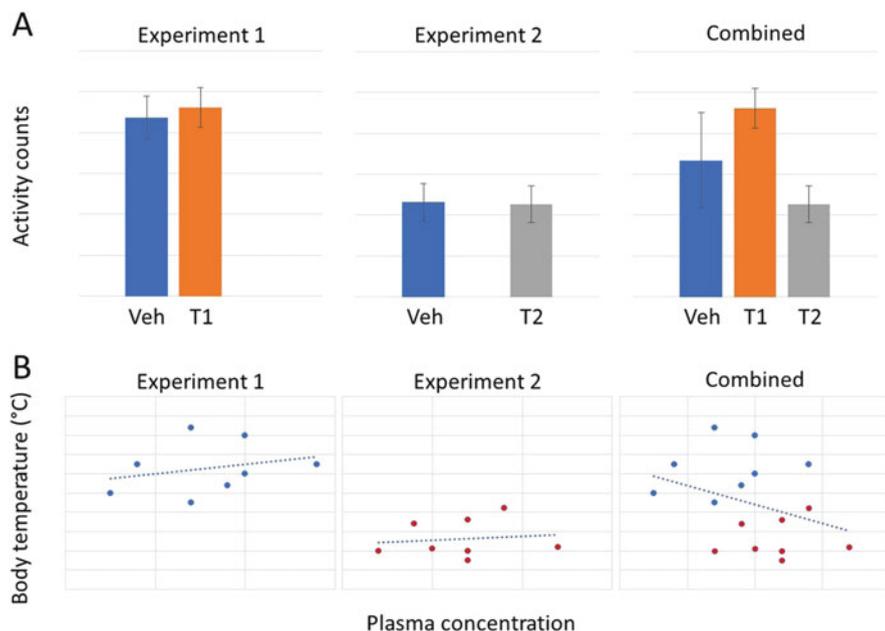
### 2.2 What Group Size for Control Groups?

As discussed elsewhere (see chapter “Building Robustness into Translational Research”), it is fundamentally important that your experiment be adequately powered. But does changing the relative sizes of individual groups in an experiment affect our ability to detect an effect? The majority of nonclinical studies use similar group sizes for all treatment groups, including controls, and there are articles on

experimental design that present only this option (e.g. Haimez 2002; Aban and George 2015; Singh et al. 2016). Bate and Karp (2014) have looked more closely at the question of relative group sizes, and they show that the traditional, balanced approach is indeed the best experimental design when all pairwise comparisons are planned. However, when the planned comparisons are of several treatment groups with a single control group, there is a small gain in sensitivity by having relatively more animals in the control group. For example, in an experiment with 30 subjects divided into 4 treatment groups and 1 control group, the power to detect a given effect size is between 2 and 5% greater if there are 10 subjects in the control group and 5 in each treatment group as compared to 5 equal groups of 6. By contrast, if the treatment groups are increased in size relative to the control group (for the same total number of subjects per experiment), there is a marked loss of power which can be 25% lower in some worst-case scenarios. It is therefore clear that the control group should never have fewer subjects than the treatment groups.

It can sometimes be tempting to add a second experiment to increase the  $n$  values or to expand the dose-range of the treatment being tested. The practice of combining experiments is relatively widespread but not always apparent from the data presentation. Combining data sets from two studies having identical treatment groups is particularly hard to spot and unless it is a part of the original experimental design should be considered a form of p-hacking (Simmons et al. 2011; Head et al. 2015), a practice where data is manipulated until it reaches significance. Somewhat easier to spot, but more dangerous, is the combining of experiments involving different experimental groups. This issue is highlighted by Lew (2008) who shows that incorrect conclusions can be drawn if the individual experiments are not analysed separately. An example of this is shown in Fig. 1a. Another issue with combining data from different experiments is the possibility of misinterpretation due to Simpson's paradox (Ameringer et al. 2009). In this case the combination of studies leads to a different conclusion to that drawn from the individual component studies analysed separately (Fig. 1b). Simpson's paradox is caused by the unequal distribution of a confounding variable between the different experiments and is often a consequence of unequal group sizes. There are few, if any, published examples from non-clinical work, but it has been highlighted as an issue in clinical trials, as in the example for the antidiabetic drug rosiglitazone described by Rucker and Schumacher (2008) where they describe a meta-analysis of several studies which reached the opposite conclusion to each individual study. Another example concerning the effectiveness of two treatments for kidney stones is presented by Julious and Mullee (1994).

Ultimately, the biggest issue with combining experiments and so inflating the  $n$  value for the control group is the same as for using historical controls (see below): it removes one of the fundamental reasons for including a control group, namely, the control of unknown variables affecting a particular study. If the number of animals in the control group is different from the treated group (as in the example described by Lew 2008) or if the df of the ANOVA do not correspond to equal groups sizes, this should be suspected. Unfortunately, I can use an example from my work where two experiments were combined to broaden the dose range, as seen with the second graph in Figure 2 of Moser and Sanger (1999). The graph shows five doses of



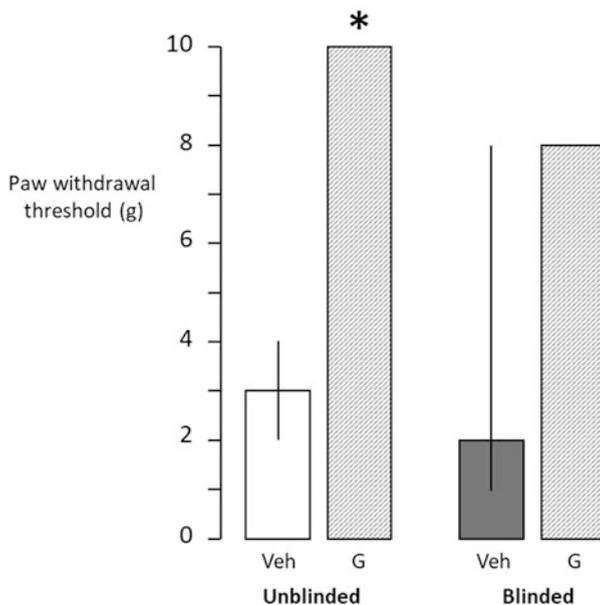
**Fig. 1** Misinterpretation of experimental outcome that results from combining data from two separate experiments where the control values differ. Panel **a** shows two experiments (mean  $\pm$  SD,  $n = 6$ ) where the two treatment doses do not differ from vehicle ( $t$ -test), whereas the combined data set (now with  $n = 12$  for the vehicle group) results in a significant biphasic effect (both doses  $p < 0.05$  vs vehicle, Dunnett's test). Panel **b** shows two experiments correlating changes in body temperature with plasma levels of drug. In each experiment the drug shows a weak tendency to increase body temperature. However, when they are combined, the drug appears to have the opposite effect, to reduce body temperature. This is an example of Simpson's paradox. Although both sets of figures were constructed from simulated data, they highlight the importance of analysing separate experiments separately

pindolol tested in the forced swim test, but combines two experiments evaluating four doses each. In our defence, we clearly mentioned this, both in the methods and in the figure legend, and we compared the two experiments statistically for homogeneity. In retrospect, however, it would have been better to present the two experiments in separate graphs (or at least as two separate data sets on the same graph) with separate statistical analyses. Transparency in such cases is always the better option when there is nothing to hide.

### 2.3 Controls and Blinding

Blinding is an essential part of any experiment (see chapter "Blinding and Randomization") and must equally apply to all control groups. If you know a particular animal has received a control substance, then you will be at risk of biasing that animal's data. This was demonstrated in a study at Pfizer by Machin et al. (2009)

**Fig. 2** Paw withdrawal threshold measured using von Frey filaments in the tibial nerve transection model of neuropathic pain. Vehicle pretreatment is compared with gabapentin (100 mg/kg PO) under unblinded and blinded conditions (\*:  $p < 0.001$  Mann-Whitney  $U$ -test;  $n = 6$ ). The data shown are median values; the bar indicates the range. Graph drawn from data presented in Machin et al. (2009)



who showed that carrying out an experiment blinded greatly increased the variability of the control group in a study of neuropathic pain where tactile analgesia was evaluated using von Frey filaments (Fig. 2). Although the effect of the positive control, gabapentin, appears similar in the two conditions, bear in mind that this is largely because of a ceiling effect and does not mean that the positive control group suffers less from such bias. Indeed, in their abstract Machin et al. (2009) state that typical values in non-lesioned animals are around 8 g, so it is possible that carrying out the study unblinded has exaggerated the effect of the positive control. Holman et al. (2015) have studied the effect of blinding on effect size and shown that over 83 studies, the effect of blinding is to reduce, in some cases quite dramatically, Hedges's  $g$  (i.e. the difference between positive and negative control means divided by the pooled standard deviation). Their review underlines the importance of applying blinding to all experimental groups in a study.

### 3 Primary Controls

#### 3.1 Choosing Appropriate Control Treatments: Not All Negative Controls Are Equal

By definition, we expect a positive control to produce an effect in our test, but a negative control to be essentially indistinguishable from vehicle. Indeed, vehicle is often the negative control most people use. I would argue that a vehicle group is not strictly a negative control group but more an evaluation of the baseline against

which to judge the effects of other control groups and treatments. It is there to evaluate the effects of things you cannot fully control or interventions you cannot avoid such as animal handling for substance administration, housing conditions, etc. One reason this neutral control cannot strictly speaking be a negative control is that other experimental variables may interact with those unavoidable interventions.

There are some documented examples of this. For example, it has been shown that elements of the surgical procedure necessary for implanting intracerebroventricular cannulae in rats were sufficient to change the behavioural response resulting from an interaction between calcium antagonists and physostigmine (Bourson and Moser 1989). These experiments showed that the potentiation of physostigmine-induced yawning by nifedipine was abolished by sham-lesioning procedures in rats, whereas the nifedipine potentiation of apomorphine-induced yawning was unaffected. The study also demonstrated that the presurgical drug treatment (i.e. desmethylinipramine and pentobarbital) or 7 days isolation was alone sufficient to reduce the yawning response to physostigmine and abolish its potentiation by nifedipine.

### 3.2 Vehicle Controls

These appear to be the simplest of control groups – simple administration of the vehicle used to formulate your test substance under the same conditions (pretreatment time, volume, concentration, etc.). This apparent simplicity can make it easy to overlook many of the issues surrounding the choice of vehicle. Much of the data we have for effects of vehicles has come from toxicology studies where control animals might receive the vehicle for up to 2 years in the case of carcinogenicity studies. Under these circumstances the tolerability and long-term toxicity of the vehicle are the main concerns, and there are several publications indicating the maximum tolerated doses of a wide range of potential vehicles (ten Tije et al. 2003; Gad et al. 2006). However, while these concentrations may be tolerated, that does not mean they are without behavioural effects. Castro et al. (1995) examined the effects of several commonly used vehicles on locomotion in mice and found marked effects of Tween, DMSO and ethanol-containing vehicles at levels well below those indicated by Gad et al. (2006) as being well-tolerated. Matheus et al. (1997) looked at the effects of Tween, propylene glycol and DMSO on elevated plus-maze behaviour in rats following their injection into the dorsal periaqueductal grey. Interestingly, whereas Castro et al. (1995) found DMSO to reduce locomotion, Matheus et al. (1997) found it to increase arm entries. DMSO, at concentrations above 15%, has also been found to modify sleep architecture in rats (Cavas et al. 2005).

Food reward is widely used as a motivating factor in many behavioural studies, particularly those studying operant behaviour. Modifying feeding conditions to modulate motivation has been shown to affect response rate in an operant discrimination task (Lotfizadeh et al. 2012), and it would therefore be expected that the use of high-calorie vehicles such as oils could have a similar effect. Although there do not appear to be any published examples of this, it is something I have observed

in rats trained on a delayed match to position task. The use of an oil vehicle almost completely abolished responding (Moser unpublished observation) although we did not establish if this was due to the high-calorie content of the vehicle or the effects of indigestion after a large volume of oil administered directly into the stomach.

In addition to intrinsic behavioural effects, many of these vehicles also interfere with the pharmacokinetics of the drugs being tested. ten Tije et al. (2003) have reviewed clinical and non-clinical effects of vehicles on the pharmacokinetics of co-administered chemotherapy agents, and Kim et al. (2007) have reported marked effects of vehicles on the ADME properties of the insecticide deltamethrin in rats. Some of the more striking examples include a 16-fold increase in danazol bioavailability in dogs when formulated in Tween 80 compared to a commercial formulation (Erllich et al. 1999) and up to a 40% decrease in the blood/plasma ratio for paclitaxel when formulated in cremephor EL (Loos et al. 2002).

### 3.3 Sham Controls

The term sham control is usually employed when it is considered that there is a part of the experimental protocol that we expect to have an impact on the outcome. Although it could be argued that the handling and injection procedure associated with administration of a drug vehicle could be expected to have some effect, the term sham control is usually applied when there is a surgical intervention of some sort. If, for example, the experiment involves a chemical lesion of the brain, the sham controls will undergo a similar surgical procedure without injection of the toxin, but including anesthesia, placement in a stereotaxic frame, incision of the scalp, trepanation, lowering of a dummy cannula into the brain (this is often omitted), suturing of the wound and appropriate postsurgical care. If the intervention is the surgical removal of tissue, then the sham control animals will be anesthetized and opened but without internal intervention and then sutured and given the same postsurgical care as the lesion group. Such groups are essential, as the anesthesia and postsurgical pain can be very stressful to laboratory animals (Hüske et al. 2016). In addition, anesthesia has also been shown to induce long-lasting effects on memory in rats (Culley et al. 2003). As part of a study to understand the processes involved in liver regeneration, Werner et al. (2014) reported on the effects of sham surgery procedures and anesthesia on the expression pattern of microRNAs in rat liver as compared to partial hepatectomy. They found 49 microRNAs modified by hepatectomy and 45 modified by sham laparotomy, with 10 microRNAs showing similar changes after both real and sham surgery. Anesthesia alone had much less effect, with only one microRNA changing in the same direction as surgery. The impact of sham surgery has also been highlighted by Cole et al. (2011) who compared the effects of standard sham procedures used in research on traumatic brain injury (craniotomy by drill or manual trepanation) with the effects of anesthesia alone. They found that the traditional sham control induced significant pro-inflammatory, morphological and behavioural changes and that these could confound interpretation in brain injury models.

### 3.4 Non-neutral Control Groups

Many experimental designs require more than one control group. There are many situations that require a neutral control group (sham, untreated etc.), a group undergoing an intervention (drug challenge, lesion etc.) and a positive control or comparator group (a treatment known to reverse the effects of the intervention – see below). Comparison of the neutral control with the intervention control shows that the intervention has had an effect – and the intervention group then becomes the point of comparison for treatments aimed at reversing its effects. This could be the effect of a lesion on behaviour (e.g. Hogg et al. 1998), a drug challenge such as amphetamine to increase locomotion (e.g. Moser et al. 1995) or an environmental change such as housing conditions, diet, light cycle, etc. (e.g. He et al. 2010). It is important to be able to demonstrate that this intervention has produced a reliable change compared to the neutral baseline or, in some cases, to a sham control. As discussed above, these controls are no longer neutral, and only through appropriate preliminary experiments can you determine how much these interventions (sham or non-sham) interfere with the primary purpose of your study. Such preliminary experiments ultimately help to reduce animal numbers as otherwise it might be necessary to include neutral, sham and intervention controls in every experiment instead of just sham and intervention. The number of times you expect to use a particular experimental design should help guide you to the optimal solution for limiting animal use.

### 3.5 Controls for Mutant, Transgenic and Knockout Animals

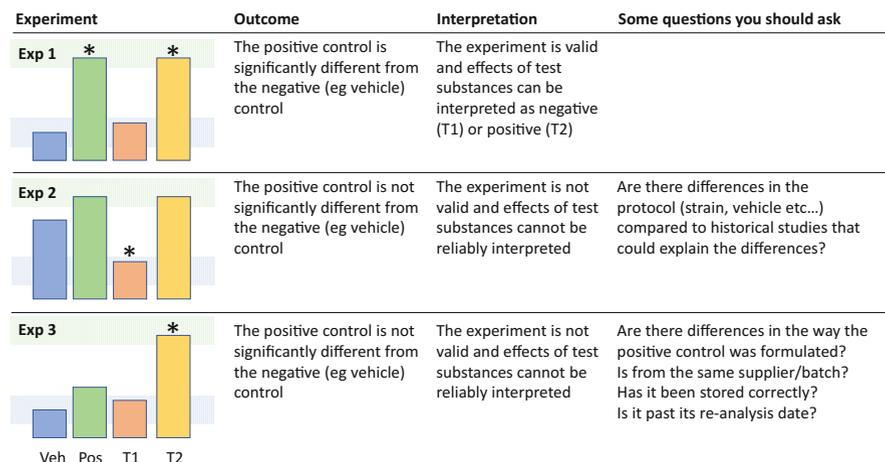
This is a vast topic that deserves a full review to itself (see chapter “Genetic Background and Sex: Impact of Generalizability of Research Findings in Pharmacology Studies”) and will only be covered superficially here. The use of mutant animals is increasing compared to non-genetically altered animals (e.g. UK Home Office statistics, 2017: <https://www.gov.uk/government/statistics/statistics-of-scientific-procedures-on-living-animals-great-britain-2017>), and they raise particular issues relating to the choice of control groups. If transgenic animals are maintained as a homozygous colony, they will gradually show genetic drift compared to the original non-modified founder line. The recommendation is to breed from heterozygotes and use wild-type littermates as controls for homozygous animals (Holmdahl and Malissen 2012). The alternative, using animals from the founder strain, is fraught with difficulty due to the multiplication of strains, as highlighted by Kelmensen (<https://www.jax.org/news-and-insights/jax-blog/2016/june/there-is-no-such-thing-as-a-b6-mouse>). Furthermore, background strain is known to have a major impact on the phenotype of transgenic animals (e.g. Brayton et al. 2012; Fontaine and Davis 2016). Jackson Laboratories have a page on their website dedicated to helping researchers choose an appropriate control for transgenic animals: <https://www.jax.org/jax-mice-and-services/customer-support/technical-support/breeding-and-husbandry-support/considerations-for-choosing-controls>. The use of

conditional expression systems, where the mutation can be expressed or not, can greatly improve the pertinence of control groups in transgenic animals (e.g. Justice et al. 2011).

Use of littermates is not always possible, particularly in the case of inbred mutant animals such as the spontaneously hypertensive rat (SHR). In the case of SHRs, the practice is to use the Wistar-Kyoto rat as a normotensive control as it is derived from the same ancestral Wistar line but its suitability as a control is questioned (Zhang-James et al. 2013).

## 4 Positive Controls

Whereas a negative control has the single job of providing a baseline from which to judge other interventions, positive controls can wear several hats at once. They can (a) demonstrate that the experimental conditions are sensitive to an active intervention, (b) provide a reference effect-size against which other interventions can be judged and (c) provide a reference effect in order to judge the conformity of that particular experiment to historical studies. If your positive control does not differ from the negative (e.g. vehicle) control group, then you need to study the data before making any interpretation of the results with novel test substances. Figure 3 illustrates some of the potential outcomes of an experiment with a negative and a positive control, suggesting how you should interpret the results.



**Fig. 3** Interpreting results when the positive and negative controls are not significantly different. The first case, Experiment 1, has a vehicle group (Veh) and a positive control (Pos) that are within their historical ranges (horizontal light blue and green shaded boxes, respectively). The asterisk indicates significant difference compared to the vehicle. In Experiments 2 and 3, the positive and negative controls are outside their expected range. How the effects of the test substances T1 and T2 are interpreted, as well as the questions that should be asked, are different between these scenarios. This figure is inspired by, and adapted from, an article from PAASP (<https://paasp.net/heads-i-win-tails-you-lose/>)

Proving that the experimental conditions are sensitive to an active intervention (typically a clinically active substance) is the main reason for using a positive control. Just as we do not necessarily know all the variables that can affect the response in negative control groups, similarly we cannot know all the variables that affect the response to an intervention we expect to work. Sometimes experiments just don't seem to work as we expect. A classic example of this is the effect of noise on the development of tolerance to benzodiazepines. Normally, chronic treatment with benzodiazepines results in tolerance developing to their anxiolytic and sedative effects. However, a series of studies carried out during a period of laboratory renovation failed to show this effect, and subsequent, more controlled studies demonstrated that under conditions of noise stress, tolerance developed to the sedative effects of benzodiazepines but not to their anxiolytic effects (File and Fernandes 1994).

The second and third uses of positive controls have an element of quality control about them. During drug development it is often necessary to compare a novel substance with one that has already demonstrated an effect in the clinic or with the lead substance of a competitor. This is particularly true when substances with similar mechanisms of action are being studied. However, some caution should be used with novel mechanisms of action. Experimental conditions are often 'optimised' to detect known compounds, and this may reduce their ability to detect compounds with novel mechanisms of activity. Models of anxiety provide some examples of this: all anxiolytic tests detect benzodiazepines, for many years the only available treatment for anxiety, but they are typically less responsive to putative anxiolytics with other mechanisms of action such as 5-HT<sub>1A</sub> partial agonists (e.g. buspirone), which have shown positive effects in the clinic but are inconsistent at best in animal models (e.g. Moser 1989; Moser et al. 1990).

When the positive control does not work as expected, it is important to check historical values to make sure that it is not the negative control that has produced an aberrant response. In any such situation, it is important to be clear what we mean by 'work'. Many drugs can have side-effects which are rate-dependent (e.g. increase behaviour when activity is low but decrease it when activity is high) which could mask or enhance another effect (e.g. Sanger and Blackman 1976). Being aware of such possibilities in a particular experimental set-up can help you better understand what has really happened in your experiment.

Finally, it is important that all the groups in the experiment are treated equally. By that I mean they are an integral part of the initial experimental design, subject to the same blinding and randomisation as other groups and included in the analysis of the experiment. There can be a temptation to first analyse the positive and negative controls using a simple two-group test and, if significant, to declare that the experiment has worked. A subsequent multigroup comparison would then analyse the effects of the test substance against the negative control. Such an analysis is not comparing the effects of the positive control and the test substance under the same conditions. This introduces a bias towards declaring experiments to have 'worked' (i.e. the positive control is significantly different compared to the negative control) when a multigroup comparison including all experimental groups might be non-significant. In some cases where the positive effect might be a unilateral lesion, the

bias is aggravated by the use of a within-subject test to validate the experiment but a between subject test to evaluate the test substance. Any such analysis that does not include all experimental groups together should be viewed with caution.

---

## 5 Secondary Controls

### 5.1 Can Baseline Values Be Used as Control?

Baseline values can be very useful points of comparison, but they are not the same as controls for a subsequent treatment or intervention. If we consider that controls are designed to measure the impact of extraneous variables in an experiment, then the baseline, by definition, cannot control for subsequent changes over time. Many physiological parameters change over time, such as locomotor activity and body temperature which show marked circadian variation (e.g. Moser and Redfern 1985). There are also many other factors which can modify drug effects and which vary over time, such as activity of metabolising enzymes (Redfern and Moser 1988).

These variations can modify the effects of drugs. This can be demonstrated by a simple example of measuring body temperature. Rodent body temperature varies over 24 h, with a trough during the light period and a peak during the dark period. Depending on when we administered a drug during the day, we might conclude that it either increased or decreased body temperature when compared against baseline when, in reality, it has no effect (e.g. Redfern and Moser 1988). The opposite is also possible: a real effect of the test substance could be masked or enhanced by the circadian variation when compared against baseline.

Measurement of baseline values is often a first step in making experimental groups homogeneous. It is essential that the groups be made as identical as possible using suitable randomisation and the spread of baseline values should not be too great. Stratified randomisation (e.g. Altman 1999) could be used if there is a wide range of baseline values but it is very rare that experimental effects will be independent of the baseline. Many drug effects are baseline- or rate-dependent such as drug effects on locomotor activity (e.g. Glick and Milloy 1973), and there is always the risk of effects being due to regression towards the mean (e.g. Morton and Torgerson 2005).

One common practice that uses baseline as a type of control value is the use of 'change from baseline' as the main variable. This can be useful when the underlying variable has low variance (such as body temperature in untreated animals at a given time of day) but dangerously confounding when the measure shows large variance and regression towards the mean (such as locomotor activity in untreated animals at a given time of day). Analysing original data or the change from baseline measure can result in different statistical outcomes (e.g. Le Cudennec and Castagné 2014). Change from baseline might be useful as an exploratory analysis but still needs very careful interpretation: a repeat measures analysis would probably be an altogether more satisfactory approach in most circumstances.

Thus, under very specific conditions, the baseline can be used as a control value. However, those conditions require that the baseline values be homogenous and with low variance and that the experiment be carried out over a very short period of time. This may be very restrictive for many types of study, but, when appropriate, the use of baseline as control can be a good way to reduce the number of animals used in a study and increase statistical power (in part due to the necessity for low variation but also because of the use of within-subject statistics), both ethically desirable.

## 5.2 Historical Control Values

Like any group in an experiment, the control groups can give a spurious result simply as a result of randomly selecting data from a population. Differences relative to this control group could then be interpreted as an effect of treatment in the other groups when in fact it is the control group that has created the difference. The frequency of this occurrence can be limited by powering your study correctly, but it may still happen from time to time. This can be controlled for in experiments that are carried out frequently, such as screening studies, by comparing your control group against historical controls.

As the objective of control groups is to provide a contemporaneous point of comparison for your experimental treatment, it is clear that historical values cannot fulfill all the functions of a control group. In particular, historical controls cannot control for the possible impact of those unknown unknowns specific to a given experiment. Indeed, it has been suggested that historical controls may be responsible for the unreliable results obtained with some test substances in the SOD-1 model of ALS (Scott et al. 2008) and Papageorgiou et al. (2017) have also demonstrated how using historical controls instead of concurrent controls in clinical studies can introduce bias.

At the same time, historical controls can provide an additional point of comparison to provide some validity for the experimental conduct. If your control values deviate from historical reference values, it allows you to determine if the data may be unreliable and potentially identify additional, previously unknown, variables that help you to improve an experimental protocol. Contemporaneous controls are not designed to provide a 'true' value that is somehow independent of the current study but are there to allow you to evaluate your current study correctly.

However, under certain conditions *or* if you are prepared to relax some of the requirements of a true control group *and* if you have a procedure that is regularly used, there are situations where historical control values can be better than concurrent controls and integrated into your experiment.

One of these situations, which has a relatively long history, is in the context of toxicology studies. For example, there are rare tumours that can occur spontaneously in rats at a rate below that which is reliably detectable in any reasonably sized control group. If such tumours are observed in a treated group, it is possibly only by chance, but comparison with the control group, in which no such tumours occurred, will not help determine if this is, or is not, the case. This is a serious issue in toxicology

testing, and the Society for Toxicologic Pathology set up a Historical Control Data Working Group to examine the issue in some depth. Their full discussion of the issues and recommendations for best practice in the case of proliferative lesions in rodents has been presented by Keenan et al. (2009). It is a similar story for in vitro toxicity studies, such as the in vitro micronucleus test where historical controls are regarded as important for evaluating data quality and interpreting potential positive results, an approach enshrined in the OECD guidelines on genetic toxicology testing (Lovell et al. 2018).

It has been suggested that for often-repeated studies historical controls could potentially replace contemporaneous controls (e.g. Festing and Altman 2002). Kramer and Font (2017) make a strong case for considering historical controls a replacement for contemporaneous controls and present a number of simulations of typical conditions showing how they could be effectively used to reduce animal usage.

---

## 6 When Are Control Groups Not Necessary?

The short answer is almost never. However, there may be circumstances where rare or anecdotal events can be reported as preliminary evidence for an effect. Two well-known and somewhat tongue-in-cheek examples include the protective effects of umbrellas against lion attacks (Anderson 1991) and parachute use to prevent injury when falling (Smith and Pell 2003). Despite their humorous approach, both papers make serious points, and the latter has been extensively cited as presenting a situation analogous to some medical practices, in which the benefits of a treatment are so obvious that they do not need testing. However, a recent analysis has found that this analogy is frequently false: when actually evaluated in randomised controlled clinical trials, only a modest proportion (27%) of such ‘obvious’ findings showed a significant benefit (Hayes et al. 2018). Thus, even if you think the outcome is so obvious that a proper control group is not necessary, be aware that you are more likely than not to be wrong. Furthermore, if this is frequently true of clinical hypotheses, I suspect that for non-clinical studies, it is likely to be true in almost all cases.

---

## 7 Conclusion

Good control groups are the ground upon which your study stands. Without them, who knows where you will fall.

---

## References

- Aban IB, George B (2015) Statistical considerations for preclinical studies. *Exp Neurol* 270:82–87  
Altman DG (1999) How to randomize. *BMJ* 319:703–704

- Ameringer S, Serlin RC, Ward S (2009) Simpson's paradox and experimental research. *Nurs Res* 58:123–127
- Anderson DR (1991) Umbrellas and lions. *J Clin Epidemiol* 44:335–337
- Andrade C (2017) Offspring outcomes in studies of antidepressant-treated pregnancies depend on the choice of control group. *J Clin Psychiatry* 78:e294–e297
- Bate S, Karp NA (2014) Common control group – optimising the experiment design to maximise sensitivity. *PLoS One* 9(12):e114872. <https://doi.org/10.1371/journal.pone.0114872>
- Bourson A, Moser PC (1989) The effect of pre- and post-operative procedures on physostigmine- and apomorphine-induced yawning in rats. *Pharmacol Biochem Behav* 34:915–917
- Brayton CF, Treuting PM, Ward JM (2012) Pathobiology of aging mice and GEM: background strains and experimental design. *Vet Pathol* 49:85–105
- Castro CA, Hogan JB, Benson KA, Shehata CW, Landauer MR (1995) Behavioral effects of vehicles: DMSO, ethanol, Tween-20, Tween-80, and emulphor-620. *Pharmacol Biochem Behav* 50:521–526
- Cavas M, Beltran D, Navarro JF (2005) Behavioural effects of dimethyl sulfoxide (DMSO): changes in sleep architecture in rats. *Toxicol Lett* 157:221–232
- Cole JT, Yarnell A, Kean WS, Gold E, Lewis B, Ren M, McMullen DC, Jacobowitz DM, Pollard HB, O'Neill JT, Grunberg NE, Dalgard CL, Frank JA, Watson WD (2011) Craniotomy: true sham for traumatic brain injury, or a sham of a sham? *J Neurotrauma* 28:359–369
- Culley DJ, Baxter M, Yukhananov R, Crosby G (2003) The memory effects of general anesthesia persist for weeks in young and aged rats. *Anesth Analg* 96:1004–1009
- Erlich L, Yu D, Pallister DA, Levinson RS, Gole DG, Wilkinson PA, Erlich RE, Reeve LE, Viegas TX (1999) Relative bioavailability of danazol in dogs from liquid-filled hard gelatin capsules. *Int J Pharm* 179:49–53
- Festing MF (2001) Guidelines for the design and statistical analysis of experiments in papers submitted to ATLA. *Altern Lab Anim* 29:427–446
- Festing MF, Altman DG (2002) Guidelines for the design and statistical analysis of experiments using laboratory animals. *ILAR J* 43:244–258
- File SE, Fernandes C (1994) Noise stress and the development of benzodiazepine dependence in the rat. *Anxiety* 1:8–12
- Fontaine DA, Davis DB (2016) Attention to background strain is essential for metabolic research: C57BL/6 and the International Knockout Mouse Consortium. *Diabetes* 65:25–33
- Gad SC, Cassidy CD, Aubert N, Spainhour B, Robbe H (2006) Nonclinical vehicle use in studies by multiple routes in multiple species. *Int J Toxicol* 25:499–521
- Glick SD, Milloy S (1973) Rate-dependent effects of d-amphetamine on locomotor activity in mice: possible relationship to paradoxical amphetamine sedation in minimal brain dysfunction. *Eur J Pharmacol* 24:266–268
- Haimez C (2002) How much for a star? Elements for a rational choice of sample size in preclinical trials. *Trends Pharmacol Sci* 23:221–225
- Hayes MJ, Kaestner V, Mailankody S, Prasad V (2018) Most medical practices are not parachutes: a citation analysis of practices felt by biomedical authors to be analogous to parachutes. *CMAJ Open* 6:E31–E38
- He M, Su H, Gao W, Johansson SM, Liu Q, Wu X, Liao J, Young AA, Bartfai T, Wang M-W (2010) Reversal of obesity and insulin resistance by a non-peptidic glucagon-like peptide-1 receptor agonist in diet-induced obese mice. *PLoS One* 5:e14205
- Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD (2015) The extent and consequences of p-hacking in science. *PLoS Biol* 13:e1002106
- Hogg S, Sanger DJ, Moser PC (1998) Mild traumatic lesion of the right parietal cortex in the rat: characterisation of a conditioned freezing deficit and its reversal by dizocilpine. *Behav Brain Res* 93:157–165
- Holman L, Head ML, Lanfear R, Jennions MD (2015) Evidence of experimental bias in the life sciences: why we need blind data recording. *PLoS Biol* 13:e1002190
- Holmdahl R, Malissen B (2012) The need for littermate controls. *Eur J Immunol* 42:45–47
- Hüske C, Sander SE, Hamann M, Kershaw O, Richter F, Richter A (2016) Towards optimized anesthesia protocols for stereotactic surgery in rats: analgesic, stress and general health effects of

- injectable anesthetics. A comparison of a recommended complete reversal anesthesia with traditional chloral hydrate monoanesthesia. *Brain Res* 1642:364–375
- Johnson PD, Besselsen DG (2002) Practical aspects of experimental design in animal research. *ILAR J* 43:202–206
- Julius SA, Mullee MA (1994) Confounding and Simpson's paradox. *BMJ* 309:1480
- Justice MJ, Siracusa LD, Stewart AF (2011) Technical approaches for mouse models of human disease. *Dis Model Mech* 4:305–310
- Keenan C, Elmore S, Francke-Carroll S, Kemp R, Kerlin R, Peddada S, Pletcher J, Rinke M, Schmidt SP, Taylor I, Wolf DC (2009) Best practices for use of historical control data of proliferative rodent lesions. *Toxicol Pathol* 37:679–693
- Kim KB, Anand SS, Muralidhara S, Kum HJ, Bruckner JV (2007) Formulation-dependent toxicokinetics explains differences in the GI absorption, bioavailability and acute neurotoxicity of deltamethrin in rats. *Toxicology* 234:194–202
- Kramer M, Font E (2017) Reducing sample size in experiments with animals: historical controls and related strategies. *Biol Rev* 92:431–445
- Le Cudennec C, Castagné V (2014) Face-to-face comparison of the predictive validity of two models of neuropathic pain in the rat: analgesic activity of pregabalin, tramadol and duloxetine. *Eur J Pharmacol* 735:17–25
- Lew MJ (2008) On contemporaneous controls, unlikely outcomes, boxes and replacing the 'Student': good statistical practice in pharmacology, problem 3. *Br J Pharmacol* 155:797–803
- Loos WJ, Szebeni J, ten Tije AJ, Verweij J, van Zomeren DM, Chung KN, Nooter K, Stoter G, Sparreboom A (2002) Preclinical evaluation of alternative pharmaceutical delivery vehicles for paclitaxel. *Anti-Cancer Drugs* 13:767–775
- Lotfizadeh AD, Redner R, Edwards TL, Quisenberry AJ, Baker LE, Poling A (2012) Effects of altering motivation for food in rats trained with food reinforcement to discriminate between d-amphetamine and saline injections. *Pharmacol Biochem Behav* 103:168–173
- Lovell DP, Fellows M, Marchetti F, Christiansen J, Elhajouji A, Hashimoto K, Kasamoto S, Li Y, Masayasu O, Moore MM, Schuler M, Smith R, Stankowski LF Jr, Tanaka J, Tanir JY, Thybaud V, Van Goethem F, Whitwell J (2018) Analysis of negative historical control group data from the in vitro micronucleus assay using TK6 cells. *Mutat Res* 825:40–50
- Machin I, Gurrel R, Corradini L (2009) Impact of study blinding on outcome of behavioural studies in pain research. Proceedings of the British Pharmacological Society. <http://www.pa2online.org/abstracts/1Vol7Issue3abst002P.pdf>
- Matheus MG, de-Lacerda JC, Guimarães FS (1997) Behavioral effects of "vehicle" microinjected into the dorsal periaqueductal grey of rats tested in the elevated plus maze. *Braz J Med Biol Res* 30:61–64
- Morton V, Torgerson DJ (2005) Regression to the mean: treatment effect without the intervention. *J Eval Clin Pract* 11:59–65
- Moser PC (1989) An evaluation of the elevated plus-maze test using the novel anxiolytic buspirone. *Psychopharmacology* 99:48–53
- Moser PC, Redfern PH (1985) Circadian variation in behavioural responses to 5-HT receptor stimulation. *Psychopharmacology* 86:223–227
- Moser PC, Sanger DJ (1999) 5-HT<sub>1A</sub> receptor antagonists neither potentiate nor inhibit the effects of fluoxetine and beflouxatone in the forced swim test in rats. *Eur J Pharmacol* 372:127–134
- Moser PC, Moran PM, Frank RA, Kehne JH (1995) Reversal of amphetamine-induced behaviours by MDL 100,907, a selective 5-HT<sub>2A</sub> antagonist. *Behav Brain Res* 73:163–167
- Moser PC, Tricklebank MD, Middlemiss DN, Mir AK, Hibert MF, Fozard JR (1990) Characterization of MDL 73005EF as a 5-HT<sub>1A</sub> selective ligand and its effects in animal models of anxiety: comparison with buspirone, 8-OH-DPAT and diazepam. *Br J Pharmacol* 99:343–349
- Papageorgiou SN, Koretsi V, Jäger A (2017) Bias from historical control groups used in orthodontic research: a meta-epidemiological study. *Eur J Orthod* 39:98–105
- Redfern PH, Moser PC (1988) Factors affecting circadian variation in responses to psychotropic drugs. *Ann Rev Chronopharmacol* 4:107–136
- Rucker G, Schumacher M (2008) Simpson's paradox visualized: the example of the rosiglitazone meta-analysis. *BMC Med Res Methodol* 8:34

- Sanger DJ, Blackman DE (1976) Rate-dependent effects of drugs: a review of the literature. *Pharmacol Biochem Behav* 4:73–83
- Scott S, Kranz JE, Cole J, Lincecum JM, Thompson K, Kelly N, Bostrom A, Theodoss J, Al-Nakhala BM, Vieira FG, Ramasubbu J, Heywood JA (2008) Design, power, and interpretation of studies in the standard murine model of ALS. *Amyotroph Lateral Scler* 9:4–15
- Simmons JP et al (2011) False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci* 22:1359–1366
- Singh VP, Pratap K, Sinha J, Desiraju K, Bahal D, Kukreti R (2016) Critical evaluation of challenges and future use of animals in experimentation for biomedical research. *Int J Immunopathol Pharmacol* 29:551–561
- Smith GCS, Pell JP (2003) Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. *BMJ* 327:1459
- Szekér S, Fogarassy G, Vathy-Fogarassy A (2017) Comparison of control group generating methods. *Stud Health Technol Inform* 236:311–318
- ten Tije AJ, Verweij J, Loos WJ, Sparreboom A (2003) Pharmacological effects of formulation vehicles. *Clin Pharmacokinet* 42:665–685
- Werner W, Sallmon H, Leder A, Lippert S, Reutzel-Selke A, Morgül MH, Jonas S, Dame C, Neuhaus P, Iacomini J, Tullius SG, Sauer IM, Raschzok N (2014) Independent effects of sham laparotomy and anesthesia on hepatic microRNA expression in rats. *BMC Res Notes* 7:702
- Zhang-James Y, Middleton FA, Faraone SV (2013) Genetic architecture of Wistar-Kyoto rat and spontaneously hypertensive rat substrains from different sources. *Physiol Genomics* 45: 528–538

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

