BIOLOGICAL REVIEWS
Cambridge Philosophical Society
Biol. Rev. (2015), pp. 000–000.
doi: 10.1111/brv.12237

Reducing sample size in experiments with animals: historical controls and related strategies

Matthew Kramer[1] and Enrique Font[2,*]

[1] *Statistics Group, Agricultural Research Service, United States Department of Agriculture, 10300 Baltimore Ave., Building 005, Beltsville, MD 20705, U.S.A.*
[2] *Laboratorio de Etología, Instituto Cavanilles de Biodiversidad y Biología Evolutiva, Universidad de Valencia, c/Catedrático José Beltrán 2, 46980, Paterna, Valencia, Spain*

ABSTRACT

Reducing the number of animal subjects used in biomedical experiments is desirable for ethical and practical reasons. Previous reviews of the benefits of reducing sample sizes have focused on improving experimental designs and methods of statistical analysis, but reducing the size of control groups has been considered rarely. We discuss how the number of current control animals can be reduced, without loss of statistical power, by incorporating information from historical controls, i.e. subjects used as controls in similar previous experiments. Using example data from published reports, we describe how to incorporate information from historical controls under a range of assumptions that might be made in biomedical experiments. Assuming more similarities between historical and current controls yields higher savings and allows the use of smaller current control groups. We conducted simulations, based on typical designs and sample sizes, to quantify how different assumptions about historical controls affect the power of statistical tests. We show that, under our simulation conditions, the number of current control subjects can be reduced by more than half by including historical controls in the analyses. In other experimental scenarios, control groups may be unnecessary. Paying attention to both the function and to the statistical requirements of control groups would result in reducing the total number of animals used in experiments, saving time, effort and money, and bringing research with animals within ethically acceptable bounds.

Key words: animal testing, animal welfare, borrowing information, control group, reduction, sample size, statistical power, three Rs.

CONTENTS

* Address for correspondence (Tel: +34 963543659; E-mail: enrique.font@uv.es).

I. INTRODUCTION

Treatment for severe spinal cord injuries is usually investigated with a rodent model. A typical experiment starts when a surgical procedure is used to produce a complete transection of the spinal cord on all subjects; experimental therapy is then applied to half, with the other half serving as controls. Scientists doing this kind of research know that mammals with a complete transection of the spinal cord do not recover spontaneously. However, the majority of these experiments using live animals still include one or more control groups, as large as or even larger than the treatment group (e.g. Ramón-Cueto *et al.*, 1998; Lu *et al.*, 2002; He *et al.*, 2013). Although rodents used in spinal cord research have their spinal cord crushed or transected under anesthesia and with the approval of institutional review boards, the end result is the same: control animals are rendered paraplegic or tetraplegic only to demonstrate that they do not recover spontaneously. Are controls in this kind of experiment superfluous? If not, can their number be reduced to minimize the adverse welfare impact of research with animals?

The focus of this review is to develop a comprehensive framework, largely in the mixed-models paradigm, to reduce the number of control animals needed without compromising the ability to detect a statistically significant treatment effect. We also explain circumstances when controls may not be necessary. We believe that the approaches and arguments outlined here will provide a valuable aid for researchers and contribute to alleviating the ethical burden of performing experiments with animals.

II. CONTROLS

Researchers may be uncertain about what the right controls are for a given experiment or whether they are really needed. This is not surprising; the conventional statistics textbooks used by biologists typically have little discussion on controls, neither their purposes nor how to integrate them into the experimental design. When mentioned, the advice is to consider them as another treatment group. The classic text by Cochran & Cox (1957) includes a short section on the use of controls, although their advice might be difficult to apply to current biomedical experiments due to advances in both technology and statistical design. Ruxton & Colegrave (2011) discuss ethics, animal welfare, and including control data from previous experiments, but lack specific instructions on how to redesign experiments to reduce current control group size.

The various types of controls used in experiments are explained clearly in Johnson & Besselsen (2002). They categorize controls as positive, negative, sham, vehicle (to test, e.g. the delivery system of a drug by itself), and comparative (positive control with a known treatment). In a negative control, control subjects remain in the 'normal' pre-experimental state; no change is expected from the pre-experimental to experimental condition. In a positive control, subjects receive some kind of pre-treatment (e.g. a toxin, a lesion) that is expected to cause a change from the pre-experimental state. The researcher can then see if a treatment is effective by comparing treated subjects with subjects that do not receive a treatment. This is the type of control used in the spinal cord experiment discussed above. The untreated control subjects may die or remain maimed; the treated ones may survive or improve. The positive control guards against miraculous recoveries, the negative control against spontaneous disease and death. Sham controls are subjected to a manipulation that mimics the procedure received by positive controls and treated animals, but nothing else. This could be as innocuous as a saline or a placebo injection, or as invasive as an operation where an organ is removed, then reattached. Placebo controls often are considered negative controls in some clinical trials, but are more correctly a type of sham control. In fact, labelling controls as 'positive', 'negative', etc, may not provide much clarity. Often a short explanation is better than a label. It is important to understand what the purpose of the control is and whether it is the right kind of control for the experiment; some experiments may require several kinds of controls. The main purpose of a control may be to show that animals do not recover on their own (as in the spinal cord example); in other cases the control can serve to make sure that animals have been exposed to the right dose of a toxin or a pathogenic agent. An alternative categorization of control group types, specific for clinical trials, can be found in ICH (2000).

How many control subjects are needed in an experiment? This should depend on the knowledge gained from the results of prior experiments. The cumulative knowledge in the field provides historical information that can be put to good use to reduce the number of animals in current control groups. The Bayesian statistical framework formally includes prior information when estimating a statistical model (e.g. French, Thomas & Wang, 2012). However, medical research typically does not take advantage of this: most researchers conducting biomedical experiments on animals continue to use 'classical' Fisherian statistics (Efron, 2013), and there is little published guidance on how to incorporate information from prior experiments (Pocock, 1976; Neuenschwander *et al.*, 2010; Viele *et al.*, 2014). Since controls are included

in most experiments, in the typical experiment we actually know far more about the control group than about the treatment group, either through experience with controls in similar experiments or through the literature (Schulz & Grimes, 2005). However, we ignore this prior knowledge when we analyze the results of an experiment carrying out a classical statistical analysis, like a *t*-test or analysis of variance (ANOVA) using only animals involved in the current experiment. A laboratory performing similar experiments over many years will be far more sensitive to controls responding in unanticipated ways than to peculiar responses of treatment groups. Why not incorporate this prior knowledge?

III. INCORPORATING INFORMATION FROM PREVIOUS EXPERIMENTS (HISTORICAL CONTROLS)

Controls can be current or historical. Current or concurrent controls are contemporaneous with the treatment group(s), whereas historical, retrospective or background controls are obtained from prior experiments where similar protocols were applied to controls. Although the use of historical controls has been widely discussed in the literature, many researchers are unaware of their potential usefulness or do not know how to incorporate them into their experimental designs. In fact, the use of historical controls in biomedical research appears restricted to some areas of toxicology (e.g. the micronucleus test, carcinogenicity studies in rodents; Yanagawa & Hoel, 1985; Hayashi *et al.*, 1989; Yoshimura & Matsumoto, 1994; Greim *et al.*, 2003; Elmore & Peddada, 2009; Hayes *et al.*, 2009; Keenan *et al.*, 2009; Dinse & Peddada, 2011), to estimate the incidence of problems in untreated animals and to detect changes in laboratory conditions, and to certain kinds of clinical trials involving humans (e.g. Chang, Shuster & Kepner, 2004; Korn & Freidlin, 2006; Zhang, Cao & Ahn, 2010; Gsteiger *et al.*, 2013), where experiments are costly and there may be ethical reasons to minimize the number of untreated subjects. In these types of research, the dependent variable is usually a proportion (dichotomous or binary outcomes), and historical controls may not come from the same laboratory, potentially introducing an additional source of variation, often termed 'laboratory bias'. A basic concern about using these historical controls is whether it is valid to test if the proportions from a current treated group and from historical controls differ. The answer depends on whether historical controls provide a good estimate for what one would have obtained had one used current controls, and how one handles the data statistically. These various possibilities have been laid out in a Bayesian context by Spiegelhalter, Abrams & Myles (2004) and Viele *et al.* (2014); they range from ignoring current controls (i.e. use only information from historical controls) to ignoring historical controls (traditional analysis). For quantitative data where normal distribution theory can be applied, there are additional options for using historical data, as discussed below.

Historical controls can be used better to estimate parameters related to the current experiment under a variety of assumptions. Under the strongest assumption, if one has a large number of historical controls and one assumes that they are stable, i.e. neither the mean nor the variance of the historical controls changes over experiments, then one can consider the historical control mean and variance to be fixed and only the current control and treatment groups have uncertainty associated with them due to sampling error. In this case, the need for a current control is debatable (other than to monitor laboratory conditions or to augment the number of historical controls for future experiments). One has only to decide whether the treatment group mean and variance need to be estimated or just the mean, i.e. assume that the historical control groups provide a better estimate of the true within-group variance. This essentially is the framework used in the literature for conditional tests (conditional on the control group parameters considered as constants); the interest here is in differences of rates (proportions) (e.g. Yanagawa & Hoel, 1985; Yoshimura & Matsumoto, 1994). This use of historical controls can result in substantial reduction in experimental animal use. According to Browne (1976), for a given power and significance level, the estimated sample sizes are between one-quarter and one-half those needed if no historical controls are used (but see Lee & Tseng, 2001).

A less stringent assumption holds that there is only sampling variability in both historical and current controls, and that all controls vary about the same mean. In that case historical and current control data can be pooled, i.e. these observations are exchangeable, giving a larger sample size to estimate the control mean. If this model is not considered appropriate, models with even fewer assumptions can be used, e.g. to allow for random experiment-to-experiment variation by including a random effect for experiments, or for treatments and controls to have different variances; in these cases the observations are not exchangeable. The analysis is then done in the mixed-models framework, which makes it possible to 'borrow' information from previous experiments on both means and variances. This is a kind of 'dynamic' borrowing of information, the amount of borrowing dictated by the quantity and characteristics of historical control data (Viele *et al.*, 2014). A related approach, that we do not consider further, is subjectively discounting but not completely ignoring historical control data, as explained in Spiegelhalter *et al.* (2004).

A further relaxing of assumptions entails use of historical controls only to estimate variances since larger sample sizes are needed for that, and to use the current control group only for estimating the control mean, i.e. borrowing information across experiments only for variance estimation. This will reduce the number of animals compared to using data only from the current experiment but not as much as for the stronger assumptions discussed above. Here the gain one obtains is largely due to the increased residual degrees of

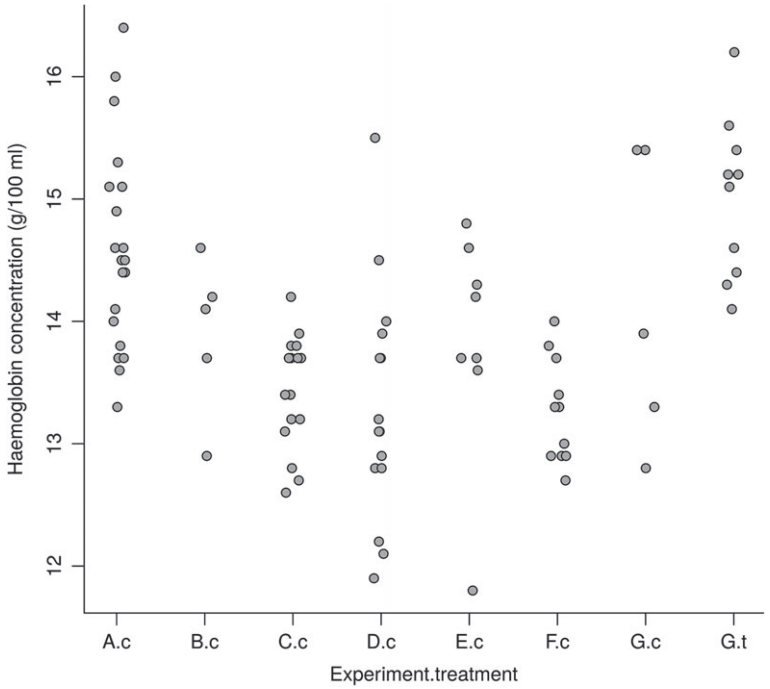

Fig. 1. Haemoglobin concentrations (g/100 ml) for control rats in eight nutrition experiments used to illustrate how different assumptions about incorporating historical controls in statistical testing affects outcomes. The historical controls are named A.c - F.c. In the current experiment, G.c is the control group and G.t the treatment group (rats in the G.t group were actually controls in the original experiment, but had the highest mean among the control groups, so we use them as the current 'treatment' group for illustrative purposes).

freedom involved in the contrast between the current control group and the treatment groups since the residual variance is better estimated by pooling over historical controls from prior experiments.

In this review we first describe how to incorporate (i.e. borrow) information from historical controls under different sets of assumptions using data from a series of nutrition experiments with rats. We then use simulations to demonstrate potential savings from using historical controls and to illustrate a few other points. We provide suggestions about the use of historical controls for quantitative and proportion data, when the variance of the control group differs from treatment groups, and we also discuss situations in which a control group may not be necessary at all.

IV. EXPERIMENTAL DATA AND SIMULATIONS ILLUSTRATING POTENTIAL IMPROVEMENTS USING HISTORICAL CONTROLS

(1) Experimental data

Our exemplary data are haemoglobin concentrations (g/100 ml), measured using an electronic cell counter, of Sprague–Dawley rats in a series of copper-deficiency experiments run over several years (Fig. 1; data from Reeves & DeMars, 2004; Reeves *et al.*, 2005; Saari *et al.*, 2006; Relling *et al.*, 2007; Johnson & Johnson, 2009). There were eight groups of control rats available for our analysis, with various numbers of individuals per group. As can be seen in Fig. 1, means of these controls varied among the groups (13.3–15.0); we chose the control group with the highest mean to act as the current 'treatment' group (G.t in Fig. 1) and one of the two groups with five observations to act as the current 'control' group (G.c in Fig. 1). While every data set is unique, these data were chosen because they appear to be fairly typical of control measurements collected over several years on a laboratory animal, thus are 'real data', even if group labelling is contrived.

We tested for a difference between a treatment group and controls using a variety of statistical models with assumptions corresponding to those outlined above (Table 1; see Table 2 for underlying statistical models and Fig. 2 for illustrative examples). Models, in decreasing order of borrowing information, are as follows.

(1) Many historical controls are available and we can consider controls to have a fixed mean; use a one-sample *t*-test to determine if the treatment mean differs from a constant (the control mean); our exemplary data are not consistent with these assumptions.

(2) Assume that controls are stable and pool current and historical controls; use a two-sample *t*-test or ANOVA to test if the treatment group mean differs from that of the pooled control mean; exemplary data are not consistent with these assumptions.

Table 1. Summary of assumption sets regarding historical controls

	Assumption set						
Assumptions	1	2	3	4	5	6	7
Control mean considered fixed	Yes	No	No	No	No	No	No
Borrow mean information from HC	NA	Yes	Yes	Yes	No	No	No
Borrow variance information from HC	No	Yes	Yes	Yes	Yes	Yes	No
Experiment-to-experiment random effects	No	No	Yes	Yes	Yes	No	No
Within-group variances	NA	All the same	All the same	T & C can differ	T & C can differ	All the same	All the same

See text for description of assumption sets.
HC, historical control groups; NA, not applicable; T & C, treatment and control groups.

(3) Assume that historical controls are relatively stable but allow for an experiment-to-experiment random effect, i.e. the current control and treatment group comprise one block; fit a linear mixed model, with experiment as a random effect, and two levels of the treatment factor (treated *versus* control), this contrasts the adjusted treatment mean to the overall adjusted control mean; exemplary data are consistent with these assumptions.

(4) Same as (3) but allow the treatment group to have a different within-group variance than control groups; fit a linear mixed model as above, with control groups all sharing one within-group variance and estimate a different within-treatment group variance; this model is over-parameterized for our exemplary data.

(5) Assume that historical control means are not sufficiently stable to use for comparison, and within-control group variances are stable but they differ from the treatment variance, allow for an experiment-to-experiment random effect; use the same linear mixed model as (4) but create a 1 d.f. contrast between the current treatment and control groups; over-parameterized for our exemplary data.

(6) Same as (5) except model as a typical ANOVA so all groups share a common within-group variance; fit a linear model with each group as a factor level, create a 1 d.f. contrast between the current treatment and current control groups; over-parameterized for exemplary data.

(7) Assume that historical control groups are not useful; test only current controls against the treatment group using a *t*-test or ANOVA. This is the typical assumption in most research laboratories where potentially useful information from historical controls is ignored; exemplary historical data ignored.

Results under the various assumptions follow, in reverse order, with assumption set abbreviated as AS.

AS 7: if only current groups are compared (G.c *versus* G.t), the *P* value from an ANOVA (1, 13 d.f.) on an estimate of the difference between means of 0.85 (S.E. = 0.47) is 0.094.

AS 6 & 5: if an ANOVA is applied to the whole data set and an *a priori* contrast between G.c and G.t made, with 82 d.f., the estimate is still 0.85, but the S.E. is now 0.42 and the *P* value 0.047. We have borrowed information from other control groups, so have a better estimate of the 'true' S.E. for the contrast, and have more degrees of freedom to test it. The same results are obtained if we model the factor, experiment, as a random effect, with the experiment-to-experiment variance estimated to be 0.328.

AS 4 & 3: if we assume for these exemplary data that all control groups share an underlying true common mean but there is a random experiment-to-experiment effect, the treatment–control difference is estimated as 0.99 (S.E. = 0.37) with a *P* value of 0.010; we get similar results assuming unequal variances.

AS 2: if we had simply pooled over controls, the treatment–control difference is 1.22 and, using ANOVA, the *P* value is 0.0001 on 1 and 88 d.f.

AS 1: if we test the treatment group against a constant control mean (= 13.79 over all controls, difference with treatment group = 1.22) using a *t*-test, on 9 d.f., the *P* value = 0.0002.

As stronger assumptions are made, *P* values decrease and therefore, our ability to detect a statistically significant treatment effect increases. However, even for the mildest AS, there is a benefit to including the historical controls; fewer concurrent controls are needed to achieve the same power if historical controls are included.

(2) Simulations

To quantify how the various assumptions affect statistical test results, we ran simulated data sets ($N = 5000$) testing for differences between controls and the treatment group through standard statistical models in R [lm and t.test in base R (R Core Team, 2013), lme in the nlme package (Pinheiro *et al.*, 2013)]. We picked a set of characteristics that we felt would be typical for biomedical studies involving animals, and similar to those from our exemplary data in the extent to which control group means differed from each other. We used the same variance for all groups; models allowing for different variances are over-parameterized. Thus, AS 3–7

Table 2. Models and tests for assumption sets

AS	Model for mean	Model for variance structure	test
1	$y_i = \mu_t$	$\widehat{\Sigma} = \sigma^2 \mathrm{I}$	$\frac{\mu_t - \mu_c}{\sigma/\sqrt{n}}$, μ_c fixed
2	$y_i = \mu_j; j = t, c$	$\widehat{\Sigma} = \sigma^2 \mathrm{I}$	$\frac{\mu_t - \mu_c}{\sigma_{tc}\sqrt{\frac{1}{n_t}+\frac{1}{n_c}}}$, where $\sigma_{tc} = \sqrt{\frac{(n_1-1)\sigma_t^2 + (n_2-1)\sigma_c^2}{n_1+n_2-2}}$
3	$y_i = \mu_j + \gamma_l$; $j = t, c_1, c_2, \ldots, c_k$; $l = 1, 2, \ldots, k$	Each block (treatment or control group, here written for $n = 3$) has the form $\begin{bmatrix} \sigma^2 & \rho & \rho \\ \rho & \sigma^2 & \rho \\ \rho & \rho & \sigma^2 \end{bmatrix}$, where current observations from $j = t$, c_1 are in block $l = 1$; each historical control group forms its own block, ρ is an estimated within-block covariance, it is 0 for observations from different blocks.	$\frac{\mu_t - (\mu_{c1} + \ldots + \mu_{ck})/k}{\sqrt{\boldsymbol{a}'\left(\boldsymbol{X}'\widehat{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{a}}}$, where $\boldsymbol{a}$ is the contrast vector, here $\left(1, \frac{-1}{k}, \frac{-1}{k}, \ldots\right)'$, $\boldsymbol{X}$ is the design matrix, and $\widehat{\Sigma}$ is the estimated variance-covariance matrix for the y_i observations.
4	$y_i = \mu_j + \gamma_l$; $j = t, c_1, c_2, \ldots, c_k$; $l = 1, 2, \ldots, k$	Historical control groups as for AS 3, current treatment and control groups as $\begin{bmatrix} \sigma_t^2 & \rho & \rho & & & & \\ \rho & \sigma_t^2 & \rho & & 0 & & \\ \rho & \rho & \sigma_t^2 & & & & \\ & & & \ddots & & & \\ & & & & \sigma^2 & \rho & \rho \\ & 0 & & & \rho & \sigma^2 & \rho \\ & & & & \rho & \rho & \sigma^2 \end{bmatrix}$, where variances of observations of the treatment group differ from those of current and historical controls.	Same as AS 3 except $\widehat{\Sigma}$ differs (see variance structure).
5	$y_i = \mu_j + \gamma_l$; $j = t, c_1, c_2, \ldots, c_k$; $l = 1, 2, \ldots, k$	Same as AS 4	$\frac{\mu_t - \mu_{c1}}{\sqrt{\boldsymbol{a}'\left(\boldsymbol{X}'\widehat{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{a}}}$, where $\boldsymbol{a}$ is the contrast vector, here $(1, -1, 0, \ldots, 0)'$
6	$y_i = \mu_j$; $j = t, c_1, c_2, \ldots, c_k$	$\sigma^2 \mathrm{I}$	Same as AS 5 except $\widehat{\Sigma}$ differs (see variance structure).
7	$y_i = \mu_j; j = t, c_1$	$\sigma^2 \mathrm{I}$	Same as AS 2 but only the current control group is used.

These are not meant to be complete descriptions since the expressions are mostly written out for equal sample sizes to avoid including weightings, but should be sufficient to understand how models for the assumption sets (see Table 1) differ. Also, to avoid making the notation more complicated and to connect better to the other columns, the expressions in the test column are given using parameters rather than estimates of the parameters (i.e. σ [population standard deviation] rather than s [sample standard deviation], μ[population mean] rather than $\overline{X}$ [sample mean]). The random block (group) effect is notated as γ, with the number of observations in a group represented by n. The subscript i indexes observations (y), j indexes means of treatment (μ_t) or control groups (μ_{c1}, μ_{c2}, ...), and there are k blocks (groups), indexed with l. The symbol I represents the identity matrix, the symbol $'$ indicates the matrix transpose operation. Other symbols are defined in the table. Complete formulae can be found in statistics books covering linear mixed models, e.g. Milliken & Johnson (2009).

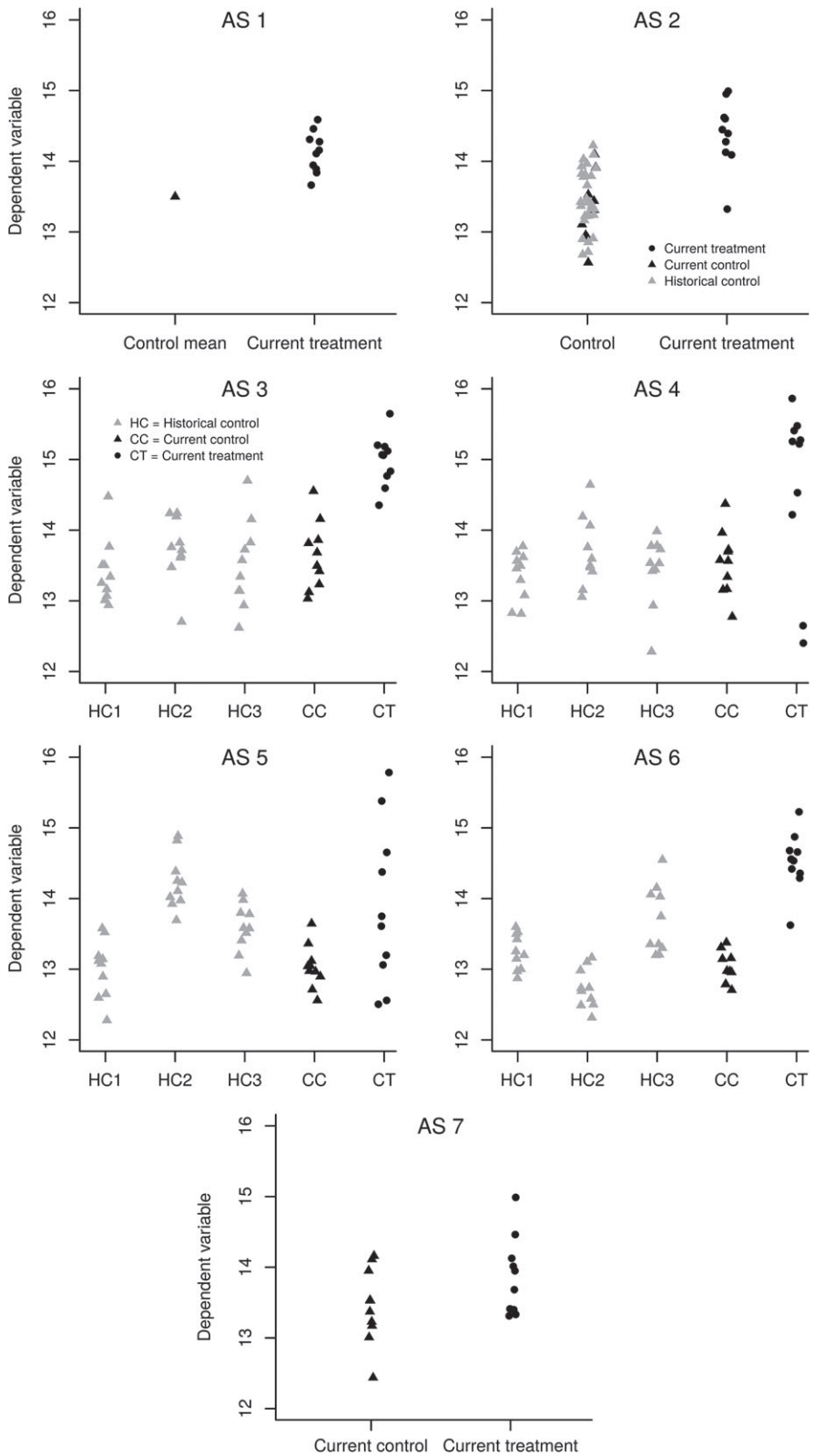

Fig. 2. Illustrations of underlying assumptions for statistical models discussed in the text and detailed in Table 2 (data simulated). In all models, samples are drawn from a normal distribution. In assumption set 1 (AS 1), the control group estimate is considered fixed and measured without error. In AS 2, AS 3, AS 6, and AS 7, all groups have the same variance; in AS 4 and AS 5 the treatment group has a larger variance than control groups.

are satisfied, while AS 1 and 2 are not. However in AS 1, the control is considered fixed; the only control information used in the test is a fixed mean, so simulation results are not affected by the violation. AS 2 is not met because control group means are affected by introduced experiment-to-experiment variation. We do not give simulation results from AS 2; power for two-sample *t*-tests can easily be obtained using available software.

The simulations were set up with all groups having the same S.D. of 3, but with group means differing. Means of historical controls were samples drawn from a normal distribution with mean 10, S.D. = 1, current control group with mean 10, current treatment group with mean 14 (effect size, Cohen's $d = 4/3$). We varied the sample sizes of the current treatment and control groups, the number of historical control groups included when testing, and the model assumptions as described above (see Table 1). Power was near 50%, i.e. significance at $\alpha = 0.05$ was detected for about half the simulated data sets, which allowed us easily to see effects of changing assumptions. Note that these simulations are purposely under-powered; for actual experiments with similar true differences in means, larger sample sizes are necessary to reach 80–90% power. We also ran the same set of simulations, but setting the current treatment mean to 10, the same as the control means, to see if the nominal 5% Type I error rate was respected by the software when making tests. That is, approximately 5% of the tests for differences between controls and treated should be flagged as significant, even though the true means do not differ. We do not give results for this. For the most part, as expected, approximately 5% of tests were significant, the exception being tests made under AS 4 above, where values ranged from 6 to 9%. These liberal values suggest to us that variances are being slightly under-estimated (downward bias) by the software. This does not affect the conclusions drawn from the simulation results that we do report.

We provide results from the simulations in Table 3. While these results apply directly only to the simulated data sets we created, the main conclusions drawn should be generalizable to other data sets with similar characteristics, like our exemplary haemoglobin concentration data.

Increasing the number of total subjects in control groups beyond 18 does little to improve power (first four rows, with four, three, two and one historical control groups) under any of the assumptions. Considering historical controls fixed (at mean = 10, AS 1) always had high power (column 5), and was surpassed only when there were many residual degrees of freedom involved in the statistical tests due to having large numbers of historical control subjects available [columns 6 and 7 (AS 3 and 4)]. Allowing for unequal variances had little effect on power [column 6 *versus* 7 (AS 3 and 4), and column 8 *versus* 9 (AS 5 and 6); note that the true variances of all groups were the same, so the small differences seen are due only to the extra parameter estimated]. Large gains in power can be achieved by using historical controls; including historical controls always increased power unless a large number of current controls was used, compare column 10

Table 3. Simulation results ($N = 5000$) giving power to detect a difference between treatment and control groups under various assumptions

Number of current controls	Number of current treated	Number of historical controls per group	Number of historical control groups	Power for a fixed control mean	Power assuming stable control means and equal variances	Power assuming stable control means and unequal variances	Power assuming unstable control means and unequal variances	Power assuming unstable control means and equal variances	Power ignoring historical controls
3	5	5	4	0.62	0.65	0.65	0.42	0.42	0.34
3	5	5	3	0.61	0.64	0.62	0.41	0.41	0.34
3	5	5	2	0.61	0.60	0.61	0.40	0.40	0.34
3	5	5	1	0.62	0.55	0.54	0.39	0.39	0.34
6	5	5	3	0.61	0.67	0.67	0.56	0.55	0.50
9	5	5	3	0.61	0.71	0.71	0.63	0.63	0.59
12	5	5	3	0.61	0.73	0.72	0.67	0.67	0.65
15	5	5	3	0.61	0.75	0.75	0.71	0.71	0.69
3	10	5	3	0.96	0.80	0.79	0.49	0.49	0.44
3	15	5	3	1.00	0.86	0.86	0.53	0.53	0.50
4	5	5	3	0.60	0.65	0.64	0.47	0.47	0.40
3	5	3	3	0.61	0.59	0.59	0.40	0.40	0.35

The true within-group S.D. was 3 for all groups. For all simulated data sets, historical control means were samples from a normal distribution (mean = 10, S.D. = 1), and the current control mean = 10, the current treatment mean = 14. Models for columns 5–10 are as follows (assumption set = AS): column 5, test against a fixed control mean of 10 (AS 1); column 6, random experiment effect (i.e. experiment 1 had current controls and treated, experiment 2 had one group of historical controls, experiment 3 had one group of historical controls, etc.), within-group variances all assumed equal, test (adjusted) current treated mean against the (adjusted) mean of all controls (AS 3); column 7, same as column 6 except variance of treated group estimated separately (AS 4); column 8, same as column 7 except test current treated against the current control as a contrast, with historical data included in parameter estimation (AS 5); column 9, same as column 8 except using ANOVA (no random effects, equal variances) (AS 6); column 10, test current treated against current control (AS 7).

(AS 7) with columns 6–9 (AS 3–6), for rows 2 and 5–8. Using current controls only [column 10 (AS 7), the usual situation in most experiments] had consistently the lowest power, followed by borrowing information only for variance estimation [columns 8 and 9 (AS 5 and 6)]. If one can assume that historical controls are relatively stable, as could be done for these simulated data, tests had high power [columns 6 and 7 (AS 3 and 4)]. Increasing the number of current controls (rows 5–8) had the largest impact on tests with the weakest assumptions [columns 8–10 (AS 5–6)]. Increasing the sample size of the treatment group (rows 9–10) had its largest effect when testing against a fixed control mean [column 5 (AS 1)].

As an example of how many animals could be eliminated by designing an experiment using historical controls under the assumption that control means are stable (AS 3 and 4), using 3 current controls and 20 historical controls yielded the same power (0.65) as using 12 current and no historical controls, thus one could eliminate nine animals (i.e. use eight animals in total instead of 17) from the current experiment without sacrificing any power. Fewer current controls could be eliminated if one only borrows variance information from historical controls, but nonetheless some savings are possible.

V. WHEN CAN HISTORICAL CONTROLS BE USED?

Compared to other kinds of research, biomedical experiments are centred on a narrow range of model organisms, often in highly controlled settings. In many research laboratories, an experiment is not an isolated event but one of a sequence of similar experiments using similar procedures and experimental animals. If treatments vary but control conditions stay the same for long series of experiments, there should be no controversy using historical controls (Festing & Altman, 2002). In less clear cases, one can rely on a combination of familiarity and expertise with the system, perhaps with a small pilot study, to make the decision. The key for designing an appropriate control group is that it is identical to the treatment group in every way except for the treatment being tested. To the extent that historical controls meet this requirement their use – in conjunction with current controls – will be justified. Generally, it should be safe to use historical controls from previous experiments conducted in the same laboratory using the same species and procedures (Greim *et al.*, 2003; Keenan *et al.*, 2009; Hayashi *et al.*, 2011). This recommendation pertains to all the various types of controls.

Consistency between historical and current experimental conditions should be given paramount consideration in the decision to incorporate historical controls as part of the experimental design. In the absence of strong evidence for heterogeneity, the use of historical data enables one to reduce the number of subjects allocated to the current control group. But if historical controls are inconsistent with current controls, there is a potential for bias and increased type I error (i.e. concluding there is a treatment effect when in fact there is none). Drift in control values over time, e.g. cure rates affected by evolving antibiotic resistance, are a clear case where historical and current controls are not consistent. Lack of stability in controls may also reflect changes in study design-related parameters such as species/strain, eligibility criteria, route of administration, vehicle, reagents, chemical and equipment suppliers, feeding and housing practices, or may be the result of sampling error as when the measurements of interest have inherently high variance. Unexplained fluctuations in the frequency of spontaneous tumours in control rats have been reported in some research facilities (Ando *et al.*, 2008; Kuroiwa *et al.*, 2013). However, what little information is available suggests that, when laboratories are well managed, controls are repeatable. For example, Hayes *et al.* (2009) report that the data for 116 control groups (each group with $N = 7$ rats) run over several years in the same laboratory were highly repeatable; the rats served as vehicle controls for the bone marrow micronucleus test. The authors conclude that 'no significant experimental variability was seen within or between control animals' (p. 423). This conclusion holds despite the fact that the source of the rats changed midway through the study. In this example, all control groups originated from the same laboratory. This differs from clinical studies in which historical controls are obtained from sources external to the laboratory conducting the research (Thomas, 2008). The latter use of historical controls is controversial because unaccounted differences between laboratories, researchers, patients or experimental protocols may render them unsuitable as a reference against which current treatment groups can be compared (Diehl & Perry, 1986). Thus, even though it is generally agreed that designs using historical controls are highly desirable for ethical and economic reasons (e.g. Gehan & Freireich, 1974; Cranberg, 1979), the evidence they provide is often considered weaker than that afforded by alternative designs, e.g. randomized clinical trials (Doll & Peto, 1980; Pocock, 1983). Our proposed use of historical controls is more conservative, as it is based on the use of control data from previous experiments conducted in the same laboratory rather than on external data sources.

The decision whether to use historical controls entails a trade-off: the potential introduction of some bias in the current study if experimental conditions have changed (i.e. 'historical bias') needs to be balanced against the reduction of current controls allowed by adding historical information.

Unfortunately, there is no automatic, foolproof method to help the researcher decide whether or not to include historical controls when only one or two historical control groups are available. In fact, to make the determination based on statistics alone would require the large sample sizes we advocate eliminating. If many historical control groups are available, we recommend using statistical quality-control methods. These methods for determining the consistency of a process are well developed and can be used to monitor

control animal performance over a series of experiments (for examples see Festing & Altman, 2002; Hayes *et al.*, 2009); they may already be instituted in well-run laboratories. There are control methodologies for both means (X-bar charts) and standard deviations (S-charts) (Wheeler & Chambers, 1992). The X-bar chart alerts the user that a group mean is excessively high or low compared to previous means, indicating that the process is not in control. The S-chart alerts the user that a group standard deviation is excessively large or small compared to previous standard deviations, indicating that the process is not in control. For example, we created standard quality-control charts for the eight control data sets of rat haemoglobin concentrations and found that three of the groups (A.c, D.c, G.t; see Fig. 1) exceeded the X-bar chart confidence limits (i.e. the assumption of a common mean is not supported). However, none exceeded the S-chart confidence limits (i.e. assuming a common standard deviation or variance is reasonable). These data would have satisfied AS 6, and also AS 3 if the kind of variation seen in the means was typical experiment-to-experiment variability, i.e. could be modelled as a random effect, or had additional information explaining why these control groups differed, e.g. due to different age or sex compositions.

What assumptions are applicable in a particular case depends largely on the type of historical controls available. If the experimental protocol remains unchanged throughout the period when historical control data were collected and the results suggest little variation among controls over time, one may assume that historical controls have a fixed mean (and variance) and use them to estimate parameters relating to the current experiment. With less-stable historical controls the researcher has to decide whether it makes sense to incorporate information from their mean and/or variance. Assuming that historical control groups can be used to estimate a common control mean and variance allows for a far greater reduction in animals than assuming that controls can only be used to estimate a common variance.

Randomization and blinding are essential aspects of experimental design, but their application to historical controls is not straightforward. Blinding may be hard to implement but randomization should not be an issue in most cases. Subjects should be randomly allocated to current control and treatment groups, but what about historical controls? One has to assume that animals will have been suitably randomized in the experiments from which the historical controls are taken, but randomizing across experiments has been constrained, i.e. each experiment can be considered a constraint on randomization. This is the classic situation producing a randomized block design. The mixed models take that into account by considering each experiment as a block (see Table 2).

(1) Determining sample sizes when using historical controls

Currently, there is no off-the-shelf software for determining how to substitute historical controls for current controls for many of our assumption sets. We suggest determining sample size using the usual tools for current controls only (AS 7), then determine how many historical controls can be substituted for each current control using simulations or 'rules of thumb', which will depend on which AS is used. For AS 2, each historical control is equivalent to one current control, i.e. there is a 1:1 substitution. AS 6 uses a contrast from a linear model, so sample sizes can be calculated using standard statistical software that estimates power, as long as contrasts can be specified. Based on our simulations for AS 3, that is, based on characteristics we dictated when creating data, substitution was conservatively 2:1 (historical:current). This can be seen in Table 3 by comparing columns 6 and 10 where power is approximately the same. In one case, the ratio is 10:6 (historical:current), in another it is 15:9 (this is after adjusting for the three current controls under AS 3). This result is only loosely generalizable: a different set of characteristics would yield different ratios. AS 3–6 are built on mixed models, where assessing power is not trivial since models can vary in many dimensions; simulation currently provides the most accurate estimate of power under different allocations of controls (Johnson *et al.*, 2015). Software for calculating power and sample size for experiments like these is readily available for AS 1, 2, and 7. It can also be found for AS 3 (blocks may be labeled 'clusters'), for example, in the software MLPowSim (which can write R scripts or call MLwiN for estimation; downloadable from http://www.bristol.ac.uk/cmm/software/mlpowsim/), OD (hlmsoft.net/od/), and GLIMMPSE (glimmpse.samplesizeshop.org, which uses the general linear multivariate model parameterization). Another on-line tool, SMEEACT (research.mdacc.tmc.edu/SmeeactWeb/Default.aspx) uses a Bayesian approach and can provide the weight that should be given to historical controls, based on their stability, for a set of anticipated or realized data. We do not know of software packages that can be used off-the-shelf for these purposes for AS 4, 5, 6 and 8; the R code we used is available from the authors.

(2) Proportion data

We do not advocate borrowing information from historical controls for analyses using proportion data, e.g. proportion that improved after receiving a treatment, unless one can make the assumption that control proportions are constant over time. This is the basic strategy taken by Korn & Freidlin (2006) and others dealing with proportion data (e.g. Yanagawa & Hoel, 1985; Hayashi *et al.*, 1989; Ryan, 1993; Yoshimura & Matsumoto, 1994). The reason for this lies in the difference between the normal distribution and other members of the exponential family of distributions (e.g. binomial, Poisson). The normal distribution has the property that the mean and variance are independent. Because of this property, if one assumes that within-control group variances are relatively constant, it makes sense to estimate this variance from a large number of observations: by borrowing information from historical data we increase the number of observations. However, the binomial and

Poisson distribution are one-parameter distributions, with a single parameter and the known sample size determining both the mean and variance. Unless the control proportion rate is stable, borrowing information from historical controls in a proportion data set which is assumed to be generated by a binomial process would not increase the precision of the current control proportion since there is no independent variance estimate. We should use historical control data to assess whether variability over time in controls is greater than that expected from sampling error alone (i.e. resulting in over-dispersion) as a means of monitoring stability in laboratory conditions. Models for binomial data that include historical control groups as random effects (Maringwa *et al.*, 2007) can improve the estimate for the true response proportion of controls (this is in the generalized linear mixed-models framework). These models can also allow for over-dispersion, a common feature of binomial data, and the additional historical controls will improve the estimate of over-dispersion.

VI. SHOULD CONTROL GROUPS BE AS LARGE AS TREATMENT GROUPS?

(1) Balanced *versus* unbalanced experimental designs

There is a sizeable literature on the use of animals in experiments, and many articles deal with sample size and a related issue, power to detect differences between treatments or conditions (e.g. Lenth, 2001; Dell, Holleran & Ramakrishnan, 2002; Devane, Begley & Clarke, 2004; Lewis, 2006; McCrum-Gardner, 2010). In general, for sample-size calculations, these articles treat controls as just another treatment group since a balanced design is usually optimal for detecting a significant treatment effect. For many experiments, that is sensible advice, but there is no statistical theory requiring that control groups be the same size as treatment groups. In fact, designs with unequal allocation to treatment and control groups can in some circumstances be more efficient than traditional balanced designs (Gail *et al.*, 1976; Bate & Karp, 2014). Most suggestions found in the literature consist of strategies for reducing the number of subjects in the treatment group. The reasoning behind recommending a smaller treatment group is that subjects in the treatment group may be more costly to produce or may undergo more invasive manipulations than controls (e.g. Ruxton & Colegrave, 2011). However, if the overall goal is to minimize the amount of suffering caused by experiments there is no reason why the same logic could not be applied to reducing the number of subjects in any other group, including the controls. Reducing the number of controls is also desirable if injury, pain or discomfort is caused to the animals in the control groups. Controls may be 'intact' animals that are spared many of the experimental manipulations applied to the animals in the treatment group, especially in the case of negative controls. But in many other cases controls undergo painful or stressful procedures, often without the potential benefits of an experimental therapy reserved for animals allocated to the treatment groups. The use of an unduly large number of controls is problematic even when control animals are relatively unharmed at the conclusion of the experiment, because most controls end up being killed as surplus.

(2) Heterogeneous variances

In general, balanced experimental designs, with the same number of individuals in the control and treatment groups, are more powerful than unbalanced designs. However, this is only true if all groups have the same variance. In many experiments, controls are not subjected to as many procedures as individuals in treatment groups, each of which can add variability to their response. As a result, there is often less variability among controls, so equal treatment and control group sizes may not maximize power.

In the spinal cord example, experimental animals need to be anaesthetized and operated on to expose the spinal cord for lesioning. To avoid possible confounding effects of anaesthesia or surgery on the variables being studied, some animals are allocated to a sham surgery control group. Animals in the sham group are anaesthetized and operated on just like the animals in the treatment group but their spinal cords are left intact. Usually, sham control groups comprise as many subjects as the treatment group. But, if one argues that, based on previous work (i.e. historical controls), the surgical procedures themselves are unlikely to affect any of the dependent variables (i.e. sham-operated controls will have low variability), the size of the sham group can be reduced.

In fact, by including a control group with little or no variability, a traditional analysis is inappropriate because the underlying assumption of homogeneity of variance for ordinary *t*-tests or ANOVA will have been violated, with the average residual variance used for calculating *P* values below that of any of the 'real' treatment groups. This can be remedied by explicitly modelling the variance or allowing for different groups to have different variances, as we did in our simulations; for discussion on *t*-tests for samples with unequal variances see Ruxton (2006). As stated above, if the variance of controls is small, then few are needed to estimate their mean accurately, allowing the number of control subjects to be reduced. However, that has to be balanced by estimating a separate variance parameter for controls and another for treatment groups, rather than estimating one variance parameter common to all groups. This is another benefit for having historical data: one can estimate necessary sample sizes for control and treatment groups without running an experiment only to discover that half the controls used were unnecessary. For binomial data (e.g. proportion that improved), a control response of zero (with no variability) will result in estimation problems with current computer software.

VII. ARE CONTROLS ALWAYS NECESSARY?

(1) No controls recover

What if prior information indicates that no subjects will recover if their spines are fully transected, the example used for opening this review? How many controls do we need? We argue that, in this type of experiment, none are necessary, since any improvement by animals with lesioned spines given a treatment differs from what we know to be historically true: that a rat that suffers a complete transection of its spinal cord does not spontaneously recover. If we have extensive data for a single species, such as some laboratory rodents, and the event of interest, i.e. spontaneous recovery from spinal cord transection, has never been reported, there is no reason to believe that conditions in a given research laboratory are so unusual that rodents raised in a standard way would differ from the norm in any basic way. The same applies to a host of well-tried experimental procedures that have a highly predictable endpoint, such as sciatic nerve crush, coronary obstruction, pancreactectomy, adrenalectomy, gonadectomy, etc. A similar rationale is used in clinical research with humans. For example, Byar (1990) states that one criterion for obviating controls in clinical AIDS trials is that 'there is sufficient experience with untreated disease to permit unambiguous evaluation of the trial result' (p. S17). Maybe it is worth checking a few animals to confirm that surgical and other procedures are working as expected, but why as many as in treatment groups if it inflicts pain and suffering on the subjects?

(2) Controls for rare events

Say that one is interested in estimating the incidence of an allergic reaction to a medication. What is the control? If subjects do not take the medication, they cannot have the allergic reaction. This type of situation is actually quite common, especially in toxicity experiments (Morton, 1998). Imagine that one is interested in convulsion rates following a treatment. A normal, healthy animal does not spontaneously convulse, so how does one compare rates of convulsions between a treatment and a control group? Note that this is not the same question as asking what sample size is needed to determine if the rate or incidence of convulsions is lower than a certain proportion, say 1/100, in which case standard formulae can be used (Dell *et al.*, 2002). Again, the answer lies in prior knowledge. If normal animals do not convulse, any animals convulsing after receiving a treatment are not exhibiting the control (normal) response, and there is an effect of the treatment, even if it occurs in only one out of 100 animals. Formally, the control has zero variance. Using any of the acceptable generalized linear model links for binomial data (e.g. logit link, probit link), if any of the treatment subjects has a response, the variance of the treatment group will be greater than 0. Estimated 95 or 99% confidence intervals (on the link scale) of the treatment groups can approach $-\infty$ (0%) or $+\infty$ (100%) but can never attain them, so will not overlap with the control group, hence they differ statistically.

VIII. RECOMMENDATIONS

The three Rs of replacement, reduction and refinement originally proposed by Russell & Burch (1959) provide a widely accepted framework for conducting animal experiments. Progress has been made in the replacement, reduction and refinement of procedures involving animals, but the use of animals in research remains a highly controversial issue and much more needs to be done. One area where progress has been somewhat limited is in the reduction of the number of animals included in an experiment. Reduction entails seeking ways of obtaining comparable levels of information from the use of fewer experimental animals, or of obtaining more information from a given number of animals (Festing *et al.*, 1998). Compared to refinement and replacement, which often require technical advances, implementing reduction strategies has an immediate impact on animal welfare. Reducing the number of animals will also result in a reduction of the resources and workload required to run an experiment. Therefore, for both ethical and economic reasons, experiments should be designed such that they use the minimum number of animals necessary to achieve meaningful scientific results. Reduction of the number of current controls provides a relatively straightforward means to this end.

Advances in the reduction of numbers of research animals have been made mainly in the fields of toxicology and vaccine testing (e.g. Hutchinson *et al.*, 2003; Jeram *et al.*, 2005). Of wider applicability are several strategies based on the implementation of more sophisticated experimental designs and statistical analyses (e.g. Mann, Crouse & Prentice, 1991; Engeman & Shumake, 1993; Festing, 1997; Festing *et al.*, 1998; Shaw *et al.*, 2002; Puopolo, 2004). However, few of the available proposals explicitly consider the issue of control group size (see Morton, 1998; Festing & Altman, 2002). We believe that, by implementing the procedures described herein, researchers will be able to reduce the number of animals they use, thereby saving time, effort and money, and bringing their research within ethically acceptable bounds.

We are not advocating that researchers eliminate or arbitrarily reduce control groups. Quite to the contrary, our aim here is to convince the reader that control groups deserve far greater attention than is current practice. A thorough understanding of the role of controls is crucial to conducting research that is both effective and ethically acceptable. Experiments in which controls are subjected to painful or stressful procedures with a highly predictable endpoint are relatively common and should be given special consideration. Recent surveys have shown that much animal experimentation is grossly underpowered (e.g. Button *et al.*, 2013), and as a consequence a typical researcher may be hesitant to reduce the number of subjects used in an

experiment. But if the aim of the controls is simply to guarantee that experimental procedures work adequately, a reduction may be fully justified on ethical grounds.

Sometimes the power loss resulting from a reduction in the number of current controls may be offset by borrowing information from historical controls. This allows the researcher to minimize the amount of suffering while at the same time preserving statistical power. One common misconception about the use of historical information is that the researcher has to choose between historical and current controls, and use only one of the two. But the use of historical information does not imply dispensing with current controls and replacing them entirely with historical controls. Provided that historical controls are available and can be used (i.e. they are consistent with current controls), a sensible combination of current and historical controls may give the best results (see Pocock, 1976). Our simulations show that there is always a benefit to including historical controls as fewer current controls are needed to achieve the desired power. Interestingly, the largest gains in power (up to 50%) in our simulations are obtained when the number of animals in the treatment group is increased and historical controls are used. Thus, a researcher that has discretion over the number of subjects allocated to different groups may achieve a more powerful design by incorporating historical controls and relocating subjects from the current control to the treatment group.

The current discussion of the use of historical controls is targeted to biomedical experimentation, which arguably accounts for the largest number of procedures involving animals. However, the focus can be fruitfully broadened to other areas of research using animals. Circumstances will likely vary among different disciplines but drawing on historical control data will generally allow a reduction of the number of animals allocated to current control groups and therefore contribute to the goal of saving time, resources and animal lives. Although satisfying the assumptions for one of the models proposed might be difficult, historical controls would still be very useful to understand how much variability there is under 'normal' situations.

We make the following practical recommendations for researchers.

(1) Verify that the experiment requires controls. There are situations for which a control group does not make sense. An experiment designed to compare different doses of the same drug, for example, may not require a control (see also Ruxton & Colegrave, 2011). Other situations may require more than one control group. Consider the type of control needed (positive, negative, sham, etc.). Lack of adequate controls may result in an inconclusive experiment, which is hard to justify for ethical reasons.

(2) If controls are necessary, consider strategies that reduce the number or the size of control groups. For example, a design that allows experimental animals to act as their own control may obviate the need for an independent control group. An incomplete block design, where controls do not appear in every block, can also reduce the number of animals used and still allow for comparisons of treatments and controls, as well as treatments with each other.

(3) If control groups are less variable than treatment groups, variances should be estimated separately for the control group, which will provide a more accurate test of mean differences. If the variance of the control group is very small relative to that of treatment groups, and this can be established using historical controls, then fewer current controls are needed since there will be little uncertainty about their responses.

(4) If available and it is reasonable to do so (i.e. no obvious systematic differences exist between current and historical control groups), consider making use of historical controls in the data analysis.

(5) If historical controls are used, one needs to decide how much information to borrow from them (i.e. which set of assumptions apply in each particular case).

IX. CONCLUSIONS

(1) The three Rs tenet is a widely accepted cornerstone that provides guidelines for improving the welfare of animals used in research. The R of reduction seeks to minimize the number of animals used in an experiment. Control groups are included in most experiments, and they can offer a relatively untapped potential to reduce sample size without sacrificing statistical power. Controls can be categorized as current or historical. Historical controls are controls from past experiments that used the same protocols as the current experiment.

(2) Historical information is part of the design of most experiments (i.e. to calculate sample sizes based on previously observed variabilities and effect sizes). However, use of historical controls is restricted to a few areas of biomedical research and is regarded with skepticism by many researchers who are unaware of their potential usefulness or do not know how to incorporate them into their experimental designs.

(3) We show, using both real data extracted from the primary literature and computer simulations, how to use historical controls to improve parameter estimates of the current experiment (i.e. means and standard deviations) under various sets of assumptions. In general, use of historical controls reduces the number of current controls necessary in an experiment and improves the researcher's ability to detect treatment effects.

(4) Borrowing information from historical controls entails a trade-off between the potential introduction of bias (if historical controls do not adequately reflect current experimental conditions) and the reduction of current control subjects. Consistency between historical and current experimental conditions should be the most important consideration in the decision to incorporate historical controls as part of the experimental design. When similar experiments are performed repeatedly in the same laboratory, using the same standard research and husbandry

protocols and the same animals (e.g. a single strain and sex of mice), there will often be scope for the incorporation of historical controls.

(5) Some experiments do not require a control group or a control group as large as the treatment group. A smaller current control group provides an easy and straightforward route to reducing the number of animals used in an experiment. This applies especially to well-tried experimental procedures that have a highly predictable endpoint, such as spinal cord transection, coronary obstruction, pancreactectomy, etc.

(6) When variability among controls is lower than among treatment groups, a control group should not be treated as just another treatment group. Instead, some thought must be given to the purpose of the controls and how to handle them appropriately from a statistical perspective, perhaps by allowing for heterogeneous group variances or not including them at all in the analysis (e.g. for zero-variance cases).

(7) Understanding the role of controls is crucial to conducting research that is both effective and ethically acceptable. An efficient use of controls can reduce the number of animals required and maximize the information obtained per experiment. Given the sheer number of animals that are routinely used as part of control groups, the procedures we outline can make a substantial contribution towards the goal of reducing the number of animals used in biomedical research.

X. ACKNOWLEDGEMENTS

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors. We thank Pau Carazo, Marian Dawkins, Hal Herzog, Benjamin Rosenthal, Sonja von Aulock, Paul Weldon, and two anonymous reviewers for providing critical feedback on earlier versions of the manuscript. We thank LuAnn Johnson (ARS statistician, Northern Plains Area), who graciously assembled the control data used in our example, and Philip G. Reeves (deceased), W. Thomas Johnson and David P. Relling for agreeing to our use of their data.

XI. REFERENCES

Ando, R., Nakamura, A., Nagatani, M., Yamakawa, S., Ohira, T., Takagi, M., Matsushima, K., Aoki, A., Fujita, Y. & Tamura, K. (2008). Comparison of past and recent historical control data in relation to spontaneous tumors during carcinogenicity testing in Fischer 344 rats. *Journal of Toxicologic Pathology* **21**, 53–60.

Bate, S. & Karp, N. A. (2014). A common control group: optimising the experimental design to maximise sensitivity. *PLoS One* **9**, e114872.

Browne, R. H. (1976). Reducing sample sizes when comparing experimental and control groups. *Archives of Environmental Health* **31**, 169–170.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J. & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* **14**, 365–376.

Byar, D. (1990). Design considerations for AIDS trials. *Journal of Acquired Immune Deficiency Syndromes* **3** (Suppl. 2), S16–S19.

Chang, M. N., Shuster, J. J. & Kepner, J. L. (2004). Sample sizes based on exact unconditional tests for phase II clinical trials with historical controls. *Journal of Biopharmaceutical Statistics* **14**, 189–200.

Cochran, W. G. & Cox, G. M. (1957). *Experimental Designs*. Second Edition. John Wiley & Sons, Hoboken.

Cranberg, L. (1979). Do retrospective controls make clinical trials "inherently fallacious"? *British Medical Journal* **2**, 1265–1266.

Dell, R. B., Holleran, S. & Ramakrishnan, R. (2002). Sample size determination. *ILAR Journal* **43**, 207–213.

Devane, D., Begley, C. M. & Clarke, M. (2004). How many do I need? Basic principles of sample size estimation. *Journal of Advanced Nursing* **47**, 297–302.

Diehl, L. F. & Perry, D. J. (1986). A comparison of randomized concurrent control groups with matched historical control groups: are historical controls valid? *Journal of Clinical Oncology* **4**, 1114–1120.

Dinse, G. E. & Peddada, S. D. (2011). Comparing tumor rates in current and historical control groups in rodent cancer bioassays. *Statistics in Biopharmaceutical Research* **3**, 97–105.

Doll, R. & Peto, R. (1980). Randomized controlled trials and retrospective controls. *British Medical Journal* **280**, 44.

Efron, B. (2013). Bayes' theorem in the 21st century. *Science* **340**, 1177–1178.

Elmore, S. A. & Peddada, S. D. (2009). Points to consider on the statistical analysis of rodent cancer bioassay data when incorporating historical control data. *Toxicologic Pathology* **37**, 672–676.

Engeman, R. M. & Shumake, S. A. (1993). Animal welfare and the statistical consultant. *American Statistician* **47**, 229–233.

Festing, M. F. W. (1997). Experimental design and husbandry. *Experimental Gerontology* **32**, 39–47.

Festing, M. F. W. & Altman, D. G. (2002). Guidelines for the design and statistical analysis of experiments using laboratory animals. *ILAR Journal* **43**, 244–257.

Festing, M. F. W., Baumans, V., Combes, R. D., Halder, M., Hendriksen, C. F. M., Howard, B. R., Lovell, D. P., Moore, G. J., Overend, P. & Wilson, M. S. (1998). Reducing the use of laboratory animals in biomedical research: problems and possible solutions. The report and recommendations of ECVAM workshop 29. *Alternatives to Laboratory Animals* **26**, 283–301.

French, J. L., Thomas, N. & Wang, C. (2012). Using historical data with Bayesian methods in early clinical trial monitoring. *Statistics in Biopharmaceutical Research* **4**, 384–394.

Gail, M., Williams, R., Byar, D. P. & Brown, C. (1976). How many controls? *Journal of Chronic Diseases* **29**, 723–731.

Gehan, E. A. & Freireich, E. J. (1974). Non-randomized controls in cancer clinical trials. *New England Journal of Medicine* **290**, 198–203.

Greim, H., Gelbke, H. P., Reuter, U., Thielmann, H. W. & Edler, L. (2003). Evaluation of historical control data in carcinogenicity studies. *Human & Experimental Toxicology* **22**, 541–549.

Gsteiger, S., Neuenschwander, B., Mercier, F. & Schmidli, H. (2013). Using historical control information for the design and analysis of clinical trials with overdispersed count data. *Statistics in Medicine* **32**, 3609–3622.

Hayashi, M., Dearfield, K., Kasper, P., Lovell, D., Martus, H.-J. & Thybaud, V. (2011). Compilation and use of genetic toxicity historical control data. *Mutation Research* **723**, 87–90.

Hayashi, M., Yoshimura, I., Sofuni, T. & Ishidate, M. (1989). A procedure for data analysis of the rodent micronucleus test involving a historical control. *Environmental and Molecular Mutagenesis* **13**, 347–356.

Hayes, J., Doherty, A. T., Adkins, D. J., Oldman, K. & O'Donovan, M. R. (2009). The rat bone marrow micronucleus test — study design and statistical power. *Mutagenesis* **24**, 419–424.

He, B.-L., Ba, Y.-C., Wang, X.-Y., Liu, S.-J., Liu, G.-D., Ou, S., Gu, Y.-L., Pan, X.-H. & Wang, T.-H. (2013). BDNF expression with functional improvement in transected spinal cord treated with neural stem cells in adult rats. *Neuropeptides* **47**, 1–7.

Hutchinson, T. H., Barrett, S., Buzby, M., Constable, D., Hartmann, A., Hayes, E., Huggett, D., Laenge, R., Lillicrap, A. D., Straub, J. O. & Thompson, R. S. (2003). A strategy to reduce the numbers of fish used in acute ecotoxicity testing of pharmaceuticals. *Environmental Toxicology and Chemistry* **22**, 3031–3036.

ICH [International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use] (2000). ICH Harmonised tripartite guideline: E10 choice of control group and related issues in clinical trials. Current Step 4 version. Available at http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E10/Step4/E10_Guideline.pdf. Accessed 1.9.2015.

Jeram, S., Riego Sintes, J. M., Halder, M., Baraibar Fentanes, J., Sokull-Klüttgen, B. & Hutchinson, T. H. (2005). A strategy to reduce the use of fish in acute ecotoxicity testing of new chemical substances notified in the European Union. *Regulatory Toxicology and Pharmacology* **42**, 218–224.

Johnson, P. C. D., Barry, S. J. E., Ferguson, H. M. & Müller, P. (2015). Power analysis for generalized linear mixed models in ecology and evolution. *Methods in Ecology & Evolution* **6**, 133–142.

Johnson, P. D. & Besselsen, D. G. (2002). Practical aspects of experimental design in animal research. *ILAR Journal* **43**, 202–206.

Johnson, W. T. & Johnson, L. K. (2009). Copper deficiency inhibits Ca2 + -induced swelling in rat cardiac mitochondria. *Journal of Nutritional Biochemistry* **20**, 248–253.
Keenan, C., Elmore, S., Carroll-Francke, S., Kemp, R., Kerlin, R., Peddada, S., Pletcher, J., Rinke, M., Schmidt, S. P., Taylor, I. & Wolf, D. C. (2009). Best practices for use of historical control data of proliferative rodent lesions. *Toxicologic Pathology* **37**, 679–693.
Korn, E. L. & Freidlin, B. (2006). Conditional power calculations for clinical trials with historical controls. *Statistics in Medicine* **25**, 2922–2931.
Kuroiwa, Y., Ando, R., Kasahara, K., Nagatani, M., Yamakawa, S. & Okazaki, S. (2013). Transition of historical control data for high incidence tumors in F344 rats. *Journal of Toxicologic Pathology* **26**, 227–230.
Lee, J. J. & Tseng, C.-H. (2001). Uniform power method for sample size calculation in historical control studies with binary response. *Controlled Clinical Trials* **22**, 390–400.
Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *American Statistician* **55**, 187–193.
Lewis, K. P. (2006). Statistical power, sample sizes, and the software to calculate them easily. *BioScience* **56**, 607–612.
Lu, J., Féron, F., Mackay-Sim, A. & Waite, P. M. E. (2002). Olfactory ensheathing cells promote locomotor recovery after delayed transplantation into transected spinal cord. *Brain* **125**, 14–21.
Mann, M. D., Crouse, D. A. & Prentice, E. D. (1991). Appropriate animal numbers in biomedical research in light of animal welfare considerations. *Laboratory Animal Science* **41**, 6–14.
Maringwa, J. T., Faes, C., Aerts, M., Geys, H., Teuns, G., Poel, B. V. D. & Bijnens, L. (2007). On the use of historical control data in pre-clinical safety studies. *Journal of Biopharmaceutical Statistics* **17**, 493–509.
McCrum-Gardner, E. (2010). Sample size and power calculations made simple. *International Journal of Therapy and Rehabilitation* **17**, 10–14.
Milliken, G. A. & Johnson, D. E. (2009). *Analysis of Messy Data, Designed Experiments*. Second Edition (Volume **1**). CRC Press, Boca Raton.
Morton, D. B. (1998). The importance of non-statistical design in refining animal experiments. *ANZCCART News* **11**, 1–17.
Neuenschwander, B., Capkun-Niggli, G., Branson, M. & Spiegelhalter, D. J. (2010). Summarizing historical information on controls in clinical trials. *Clinical Trials* **7**, 5–18.
Pinheiro, J., Bates, D., DebRoy, S. & Sarkar, D., R Development Core Team (2013). nlme: linear and nonlinear mixed effects models. R package version 3.1-109.
Pocock, S. J. (1976). The combination of randomized and historical controls in clinical trials. *Journal of Chronic Diseases* **29**, 175–188.
Pocock, S. J. (1983). *Clinical Trials – A Practical Approach*. John Wiley & Sons, Chichester.
Puopolo, M. (2004). Biostatistical approaches to reducing the number of animals used in biomedical research. *Annali dell'Istituto Superiore di Sanità* **40**, 157–163.
R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna ISBN 3-900051-07-0. Available at http://wwwR-projectorg/. Accessed 4.12.2013.
Ramón-Cueto, A., Plant, G. W., Avila, J. & Bartlett Bunge, M. (1998). Long-distance axonal regeneration in the transected adult rat spinal cord is promoted by olfactory ensheathing glia transplants. *Journal of Neuroscience* **18**, 3803–3815.
Reeves, P. G. & DeMars, L. C. (2004). Copper deficiency reduces iron absorption and biological half-life in male rats. *Journal of Nutrition* **134**, 1953–1957.
Reeves, P. G., DeMars, L. C., Johnson, W. T. & Lukaski, H. C. (2005). Dietary copper deficiency reduces iron absorption and duodenal enterocyte hephaestin protein in male and female rats. *Journal of Nutrition* **135**, 92–98.
Relling, D. P., Esberg, L. B., Johnson, W. T., Murphy, E. J., Carlson, E. C., Lukaski, H. C., Saari, J. T. & Ren, J. (2007). Dietary interaction of high fat and marginal copper deficiency on cardiac contractile function. *Obesity* **15**, 1242–1257.
Russell, W. M. S. & Burch, R. L. (1959). *The Principles of Humane Experimental Technique*. Methuen, London.
Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to student's t-test and the Mann-Whitney U test. *Behavioral Ecology* **17**, 688–690.
Ruxton, G. D. & Colegrave, N. (2011). *Experimental Design for the Life Sciences*. Third Edition. Oxford University Press, Oxford.
Ryan, L. (1993). Using historical control data in trend tests for clustered binary data. *Biometrics* **49**, 1126–1135.
Saari, J. T., Reeves, P. G., Johnson, W. T. & Johnson, L. K. (2006). Pinto beans are a source of highly bioavailable copper in rats. *Journal of Nutrition* **136**, 2999–3004.
Schulz, K. F. & Grimes, D. A. (2005). Sample size calculations in randomised trials: mandatory and mystical. *Lancet* **365**, 1348–1353.
Shaw, R., Festing, M. F. W., Peers, I. & Furlong, L. (2002). Use of factorial designs to optimize animal experiments and reduce animal use. *ILAR Journal* **43**, 223–232.
Spiegelhalter, D. J., Abrams, K. R. & Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. John Wiley & Sons Ltd., Chichester.
Thomas, N. (2008). Historical control. In *Wiley Encyclopedia of Clinical Trials*, pp. 1–3. John Wiley & Sons, Hoboken.
Viele, K., Berry, S., Neuenschwander, B., Amzal, B., Chen, F., Enas, N., Hobbs, B., Ibrahim, J. G., Kinnersley, N., Lindborg, S., Micallef, S., Roychoudhury, S. & Thompson, L. (2014). Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical Statistics* **13**, 41–54.
Wheeler, D. J. & Chambers, D. S. (1992). *Understanding Statistical Process Control*. Second Edition. SPC Press, Knoxville.
Yanagawa, T. & Hoel, D. G. (1985). Use of historical controls for animal experiments. *Environmental Health Perspectives* **63**, 217–224.
Yoshimura, I. & Matsumoto, K. (1994). Notes on the use of historical controls. *Environmental Health Perspectives* **102** (Suppl. 1), 19–23.
Zhang, S., Cao, J. & Ahn, C. (2010). Calculating sample size in trials using historical controls. *Clinical Trials* **7**, 343–353.

(*Received 2 March 2015; revised 9 October 2015; accepted 12 October 2015*)