Examples of justifications for experimental design and animal number in grant applications

Introduction

There is a wide range of designs and approaches to animal experimentation that are appropriate depending on the objectives of the research proposal. In all cases, the MRC expects that researchers provide well justified information in their applications concerning the experimental design and its suitability for robustly answering the research questions posed.

While we recognise that there is an ethical imperative to reduce the number of animals used, it is unethical to conduct a study that, because of its limited size, has inadequate statistical power to robustly answer a research question. As fully detailed in our guidance (p55-62), it is important that applicants provide adequate justification for their choice of design and numbers of animals and interventions.

A number of examples are included in the tables below to illustrate the level of detail and type of information we are looking for.

Table 1: Examples for avoidance of bias (randomisation and blinding)

Table 2: Examples covering breeding, pilot studies to determine effect size, justification of effect size, and sample size

Table 1: Avoidance of bias - randomisation and blinding

Randomisation and blinding	**Example 1**	Mice receiving the drug or sham treatment will be randomised using a random number generator. Staff administering the drug or sham treatment will be different from those assessing the effects. Mice and subsequent blood and tissue samples will be labelled such that staff assessing the effects of the treatment and analysing the results will be unaware which received the drug or sham treatment.
	Example 2	The treatment is administered in water and therefore each cage of mice is the experimental unit. Each cage will be assigned to treatment groups randomly. Researchers assessing the effect of the treatment on the whole mice will be different from those administering the drug. Staff assessing the mice will have no indication asto which cages receive the treatment and which do not e.g. cage cards and other records will not be marked with this information. Assessment of the outcome will be performed blind to treatment allocation.
	Example 3	Mice will be genotyped for the mutation but this information will not be written on the cage cards or be accessible to the staff phenotyping the mice so that the mice are assessed blind. Phenotyping and data scoring and analysis will be performed with the researcher blind to the genotype of the mice.
	Example 4	Animals will be randomly allocated to the treatment using a computer-generated sequence and researchers making measurements on the animals will be blind as to the allocation.
	Example 5	It is not possible to blind the whole mouse phenotyping experiments, due to the mutant mice having a coat colour phenotype different to that of the wild type. However, when the histology sections are cut, stained and analysed the samples will be blinded in order to reduce bias.

Table 2: Examples which cover breeding, pilot studies to determine effect size, justification of effect size, and sample size

Worked examples (values have been replaced with letters **A, B, C** and so on)

A) Reasoning behind choice of statistical methods		We have chosen to use the following statistical methods(s) for the following reasons…
B) Breeding	**Generation of experimental genetically modified (GM) animals for analysis**	To breed homozygous viable animals which are otherwise infertile, **A** breeding pairs of heterozygous parents will be required to generate cohorts of **B** homozygous same-sex mutant animals across at least **C** successive litters (amongst which homozygous mutants of one sex are expected at 1/8, and their heterozygous sibling controls at 1/4) = **D** animals total.
	Creation of transgenic lines	Eggs are provided by e.g. our industrial partner/collaborator/obtained commercially. For each construct, it is typically necessary to follow at least **A** transgenic founders (because of …), and to screen offspring of at least **B** successive litters. This requires **C** animals.
	Maintenance and cryo-preservation of GM lines:	Genetically modified lines will be maintained to provide reference animals for breeding or experimental controls, or fertilised eggs cryo-preserved for future work. Typically, **A** breeding pairs will be required to generate sufficient fertilized eggs per line for these purposes, but difficult backgrounds for cryo-preservation may require additional breeding.
		For maintenance only, **A** breeding pairs per year with **B** litters each are required = **C** animals.
		For cryo-preservation of a homozygous line, **A** breeding pairs are

		required to produce **B** females, for subsequent production of sufficient numbers of fertilized = **C** animals.
C) Pilot studies to determine effect size		For pilot experiment **A** an effect size of **B** will be sufficient to accept the result as worth investigating because …. We need **C** number of animals to determine the standard deviation with a precision of **D**% to enable us to calculate sample size requirements for further experiments.
D) Justification of effect size		An effect size of **A** or greater would be considered of sufficient interest to be worth taking forward in further research because...
E) Sample size – general examples	**Example 1**	We have used data from preliminary experiments to obtain an estimate of the standard deviation of **A** for our primary outcome and then performed power calculations in order to calculate the sample size necessary per treatment group to be **B**% confident that we will be able to detect a difference of at least **C** at the **D**% level.
	Example 2	For Experiment **A**, based on a standard deviation of **C** in the measurement of **D** (effect size of **E** from e.g. previous experience or publication reference, etc. – see justification of effect size), we will require **F** or more animals per group to detect an effect size of **E** or greater at a significance level of **G%** with power of **H%** using statistical test **I**.
	Example 3	For each experimental situation the following number of observations, mean, standard deviation and standard error was calculated for each genotype….. We tested for homogeneity of variances **A**1 and **A**2 for genotypes 1 and 2. In each comparison the null hypothesis of equality of

		variances was rejected/accepted because… Power calculations for each genotype, and test (using the different standard deviations in the different groups accounting for where the variances differed) were performed, for statistical power of **C**% to detect a difference of **D** or greater (see effect size justification) the number of animals required for genotype 1 is **E** and genotype 2 is **F**.
F) Sample size – specific examples	**Normally-distributed data, comparison of two groups at one time point**	Our primary outcome is blood pressure measured at **A** weeks after the start of the experiment. In our previous work (give details or references), blood pressure measurements have been approximately normally distributed with a standard deviation of **B** and we will test the difference between the two groups at this time point using a t-test. With **C**% power, at a **D**% level of significance we would be able to detect a difference of **E** (see effect size justification) or greater with **F** animals in each group.
	Normally-distributed data, comparison of more than two (B) groups at one time point	Our primary outcome is liver weight measured when the animals are culled at 6 weeks. In our previous work (give details or references), liver weight measurements were approximately normally distributed with a standard deviation of **A** and we will test the difference between the **B** groups using one-way analysis of variance. With **C**% power, at a **D**% level of significance we would be able to detect a difference between the highest and lowest groups of **E** (see effect size justification) or greater with **F** animals in each group.
	Normally-distributed data, with the treatment administered to dams and the offspring being assessed	Our primary outcome is offspring weight 3 weeks after birth. High fat diet is fed to the dams and we need to take account of clustering of offspring within dams. We assume an intra-cluster

		(i.e. intra-litter) correlation coefficient of **A,** which we have found in previous work using this strain of mouse (give details or references). Weight in the offspring is approximately normally distributed with a standard deviation of **B** and we will analyse the data using random effects models to take account of the clustering of the litters. To estimate the number of dams required, we assume litter sizes of **C**. With a power of **D**% and a level of significance of **E**% we would be able to detect a difference of **F** (see effect size justification) between the offspring of high fat diet fed dams and those of control dams fed normal chow.
	Binary outcome, comparison of two groups	Our primary outcome is presence or absence (clearance) of middle ear inflammation. In our control group we anticipate that we will identify clearance of middle ear inflammation in **A**% of animals as found in our previous work (give details or references). We will test the difference between the treated and control groups using a chi-squared test. With **B**% power, at a **C**% level of significance we would be able to detect a difference of **D**% or greater between the two groups (i.e. if inflammation were cleared in more than (**A+D**)% of the treatment group (see effect size justification) with F animals in each group.
	Skewed data that can be transformed to approximate a normal distribution, comparison of two groups	Our primary outcome is vaccine response in International Units at 3 weeks. In our experience this variable is positively skewed (give details or references), so a log transformation will be used to normalise the data before using a t-test in the analysis. The mean of the logged values divided by their standard deviation gives us our standardised measure of the outcome. With **A**% power, at a **B**% level of significance we would be able to detect an effect size of **C** (see effect size justification) or greater with **D** animals in each group.

	Non-normally distributed data requiring a non-parametric test, comparison of two groups	Our primary outcome is observed counts of abnormal glomeruli in **N** sections from a single kidney taken from each animal. This variable is not normally distributed, so we will use a Mann-Whitney test of the difference between our two groups, this being a non-parametric test. This test is less powerful than a test of normally-distributed data so we will need to inflate our sample size above that required for normally-distributed data. From our previous work (give details or references) we estimate the standard deviation of our data (albeit not-normally distributed) to be **A**, and aim to detect an effect size of **B**. If the data were normally distributed, and assuming a power of **C**% and a level of significance of **D**%, we would require **E** animals in each of our mutant and wildtype groups to detect a difference of **B** or greater. To allow for the non-normality of the data we will increase the sample size in each group by **F**%.
	Repeated measures, comparison of two groups	Our primary outcome is blood pressure measured over the course of the study. We will measure the animals at **A** time points. Repeated measures analyses will be used. Assuming a standard deviation of **B** in the blood pressure measurements, we have conducted 1,000 simulations of the data to estimate the number of animals required in each group to detect a difference of **C** between the treated and untreated animals over the six time points. With **D**% power and at **E**% level of significance our simulations indicate that we will be able to detect a difference of **C** (see effect size justification) or greater with **F** animals per group
	Repeated measures using averages of measurements, comparison of two groups	We are measuring blood pressure at **A** time points after treatment. Our primary outcome is the average of these **A** measurements for each animal. In our experience the averages of

		A measurements are normally distributed with a standard deviation of **B** (give references or pilot data). With **C**% power and at **D**% level of significance, **E** animals per group would allow us to detect a difference between the two groups of **F** or greater in the means of the average blood pressures for each individual animal (see effect size justification).