Published in final edited form as:

Vet Pathol. 2012 January ; 49(1): 24–31. doi:10.1177/0300985811417247.

Designing Phenotyping Studies for Genetically Engineered Mice

C. J. Zeiss¹, J. M. Ward², and H. G. Allore³

¹Section of Comparative Medicine, Yale University School of Medicine, New Haven, CT, USA

²Global VetPathology, Montgomery Village, MD, USA

³Department of Internal Medicine, Yale University School of Medicine, New Haven, CT, USA

Abstract

A phenotyping study records physiologic or morphologic changes in an experimental animal resulting from an intervention. In mice, this intervention is most frequently genetic, but it may be any type of experimental manipulation. Accurate representation of the human condition under study is essential if the model is to yield useful conclusions. In this review, general approaches to the design of phenotyping studies are considered. These approaches take into account major sources of reduced model validity, such as unexpected phenotypic variation in mice, evolutionary divergence between mice and humans, unanticipated sources of variation, and common design errors. As poor design is the most common reason why studies fail to yield enduring results, emphasis is placed on reduction of bias, sampling, controlled study design, and appropriate statistical analysis.

Keywords

mouse; phenotyping; design; hypothesis

John Steinbeck could never have foreseen how often the title of his book *Of Mice and Men* would appear in comparative reviews of mouse and human disease. However, like the mouse that is the subject of the Robert Burns poem from which the title originates, the best laid plans may not yield the expected outcome. All too often, studies using rodents fail to accurately reflect the human disease they were intended to model, or they fail to translate a therapeutic intervention across species. Studies using mice fall mainly into one of two broad categories—those that model mechanisms of disease, and those that explore the capacity of an intervention to mitigate disease. For the latter to translate into effective clinical therapies, the former must accurately reflect the human disease.⁷³ The majority of phenotyping studies using genetically engineered mice (GEM) are hypothesis driven, and their design heavily influences the number and type of animals used, the data reported, and their analysis. All of these components should be planned, but like Burns's mouse, aspects that derail the process cannot always be foreseen. The purpose of this review is to explore the types of studies most commonly encountered, aspects of comparative biology or study design that may result in suboptimal results, and finally, to describe principles of good study design.

© The Author(s) 2012

Corresponding Author: Caroline Zeiss, Section of Comparative Medicine, Yale University School of Medicine, TAC N230, New Haven, CT 06520, USA, caroline.zeiss@yale.edu.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

What Constitutes a Phenotyping Study?

In its broadest sense, a phenotyping study records clinical, morphologic, physiologic, or cellular changes in mice resulting from an intervention. Today, this intervention is most frequently genetic but may be any type of experimental manipulation including dietary, pharmacologic, infectious, physical, or surgical.³⁶ Most commonly, phenotyping is only one aspect of a relatively narrowly focused hypothesis-driven study; however, it may take center stage in enormous in vivo phenotyping projects that lack a specific hypothesis.⁸⁰ In this section, approaches to major categories of hypothesis-driven and large-scale phenotyping are discussed.

Hypothesis-Driven Phenotyping

Our capacity to generate genetically defined rodents was revolutionized first by the random introduction of new genes into the mouse genome,^{7,57} and then by the ability inactivate specific genes in murine embryonic stem cells.⁷² Most commonly, GEM studies entail characterizing the phenotype induced by genetic alteration of a known gene that is often associated with a human disorder. In these studies, the only experimental variable is the introduced genetic alteration, although many unintended variables such as strain and concurrent disease may be present.^{6,82} Reports typically include methods of generating the mice, clinical and clinicopathologic data, salient histopathologic lesions, and molecular data illustrating a potential mechanism of disease.⁷⁷ In strongly mechanistic studies, clinical or pathologic abnormalities may be minimally reported (or reported in supplementary data) in favor of molecular or biochemical data.⁷⁷ In general, four to ten mice per sex, genotype, and age group are used.^{6,41,64,82} In some reports, animal numbers may be as small as two to four animals per experiment,⁷⁷ whereas in others, variation inherent in the technology (called measurement error) or biological outcome requires more animals.⁴³ In reality, the sample size depends on how variable the phenotype is and is discussed in detail below.

Large-Scale Phenotyping

Several large-scale, random mutagenesis efforts continue to generate mice in which phenotype must be assessed to infer gene function.²⁵ Phenotypic assessment of large numbers of mice must integrate the results of predetermined phenotyping screens and unified databases with comparable datasets and analytic methods.⁸⁰ The challenges posed by this “hypothesis-independent” approach lie more in large-scale logistics and implementation than in the principles of study design inherent to hypothesis-driven research. In Europe, several research organizations spanning the nations of the European Commission have developed programs and protocols for phenotyping genetically altered mice (<http://empress.har.mrc.ac.uk/viewempress>). The International Knockout Mouse Consortium, which includes the NIH knockout mouse program (KOMP) program (<http://www.komp.org/>), will soon attempt to phenotype numerous new lines of GEM using these European or other protocols (<http://www.knockoutmouse.org/>). Recently, Genentech has completed large-scale phenotyping on over 400 knockout lines for secreted and transmembrane proteins using some of the suggested protocols mentioned above.⁷⁰

How Good a Model Is the Mouse, Really?

We hear often that mice are not small, furry humans. Nevertheless, they are the most commonly used animal to model diseases that are often uniquely human. Their use is valid if the parameters of the hypothesis are clearly defined. For example, a mouse model cannot replicate the cognitive and emotional spectrum of Alzheimer’s disease in humans. However, a mouse can be used to explore the biochemical effects of amyloid precursor protein in nervous tissue.¹⁵ Nevertheless, the ultimate contribution of animal research to human health

has been questioned.^{24,38,58} Many of the causes for this issue are leveled at problems of study design,^{5,24,54} which will be addressed below. However, some less familiar sources of reduced model validity arise from physiologic and genetic differences between mice and humans, and these sources are described below.⁴⁶

Evolutionary Divergence May Create Unexpected Phenotypes

The use of animal models is based on the evidence that the physiology of divergent organisms is driven by homologous mechanisms. For example, loss of the transcription factor Pax 6 (PAX6) results in similar ocular defects in mice, humans, and drosophila.³¹ Ectopic expression of the mouse *Pax6* gene in flies induces formation of well-developed ectopic eyes, thus establishing Pax 6 as a master transcriptional regulator across species.²⁷ However, homologous proteins do not always remain functionally equivalent during evolution,^{2,3,9,26,46,59} which is particularly true at the systems level and reflects differences in target transcriptional networks accompanying speciation.⁴⁵

Functional differences in human and mouse genes can be difficult or impossible to predict and usually emerge during the course of the experiment.³ For example, the *OCRL* gene encodes a phosphatidylinositol 4,5-bisphosphate 5-phosphatase, and when it is mutated in humans, it results in a syndrome characterized by mental retardation and aminoaciduria known as X-linked Lowe syndrome.³⁵ However, the mouse *Ocrl* knockout appears normal.³⁵ The reason for this discrepancy lies with a related autosomal gene *Inpp5b* (inositol polyphosphate-5-phosphatase 5B). This gene is not associated with a human disease condition, and its mouse homologue is expressed at much higher steady-state levels than the human counterpart. When mouse *Inpp5b* is mutated, a phenotype unrelated to Lowe syndrome (testicular degeneration) ensues. Eliminating both genes in the mouse (*Ocrl-Inpp5b* double-knockout) is embryonic lethal,³⁵ thus requiring an alternate approach to such tissue-specific elimination of *Ocrl* on an *Inpp5b* knockout background to model the brain phenotype of Lowe syndrome.

Background Genetic Heterogeneity in Humans Exceeds That in Mice

The majority of prevalent human conditions such as diabetes; obesity; and immunologic, cardiovascular, and aging-associated diseases are complex phenotypes that result from interaction of multiple genetic and environmental factors.^{10,50,75} To study these conditions in mice, the predominant methodologic bias is toward simplification of a complex system into component subsystems that isolate a single intervention against a background of controlled variables.⁵¹ Although this approach may be used to establish causation of individual variables, conclusions reached may not apply when complex interactions are considered in the whole system. With a view to better modeling of the genomic diversity underlying complex traits, investigators at Oak Ridge National Laboratory have developed a new genetic reference population of mice (the Collaborative Cross) derived from eight inbred strains.³⁴ However, for most studies involving GEM, it is recommended that the mutation be back-crossed onto a single strain for 10 generations (congenic mice).⁶¹ The phenotype of animals that have been incompletely back-crossed to a single strain may be quite variable, and in some cases, the phenotype may be lost on one background and strengthened on another.⁶¹ A phenotype that is consistent regardless of background strain is likely to prove most useful over time.

Disease-Causing Mutations in Humans Encompass Broad Allelic Spectra

The majority of highly prevalent human diseases result from an allelic spectrum within a single gene, or multiple independent alleles that predispose the individual to disease.⁴⁸ This “geneticist’s nightmare” cannot be accurately modeled by the complete elimination of gene function seen in knockout mice. The goal of KOMP and other high-throughput knockout

projects is to observe the effect of loss of function of every gene in the mouse genome. Although these effects may be embryonic lethal and often do not reflect the corresponding human condition, this approach provides an essential platform upon which more subtle defects such as ENU point mutagenesis⁵³ may be assessed.

Gene–Environment Interaction

The interaction between genotype and factors such as diet, lifestyle, and/or dwelling situation result in a more diverse phenotype in humans than can be accurately modeled in mice. Further, in many human studies, measures of quality of life, mental health, and physical function may be as important as a primary disease-specific outcome that cannot be directly measured with animal models.⁶⁸ Genetic background may profoundly affect the result of an environmental intervention, as is commonly seen in the varying effect of mouse strain has on an intervention such as diet.^{12,23,30} Standardization of test conditions is generally regarded as a good practice, as it minimizes the effects of potentially confounding variables on the primary research question.⁷⁶ Recently, Richter et al raise an interesting point regarding behavioral studies, which are notorious for the degree to which the local environment influences outcome.^{11,60,76} They suggest that imposing rigid standardization is ultimately impossible to replicate across laboratories, that doing so may yield results specific to a particular laboratory, and that phenotypes that persist across some degree of heterogenization may be more reproducible.⁶⁰ Nevertheless, for most purposes, eliminating confounding variables from a study design is recommended.

Gene–Gene Interaction

Interaction between non-allelic genes, or epistasis, is a well-established phenomenon that underlies the variation of phenotype induced by a single genetic alteration in differing mouse strains.^{32,44,55,65} More recently, the capacity to assess genome-wide gene expression has fostered our ability to assess functional interactions at the transcriptional level. Using a variety of computational methods,⁴² expression data may be aggregated into functional gene networks. Although in its infancy, a systems approach to understanding disease is rapidly evolving in parallel with whole-genome mutagenesis.^{66,79}

General Principles of Good Study Design

Guidelines for experimental design, analysis, and reporting are available in the literature,^{19–21,37,67,78} and of course through consultation with a statistician. A checklist of design and statistical parameters to consider when performing phenotyping experiments is given in Table 1 and described in more detail below.

Hypothesis

A hypothesis is a statement of a statistically testable outcome. It sets the framework for the experimental design and thus forms the backbone of the experiment. The hypothesis is often framed as two sided, with a null and alternate hypothesis (eg, the null hypothesis states that there is no difference between the experimental genotype and wild type for levels of interleukin-6). Alternately, it may be presented as one sided (eg, the experimental genotype has a higher level of interleukin-6 than the wild-type control group).

Description of Variables and Experiments

Next, the ability to replicate a study rests heavily on good reporting of the animal variables used. This information should ideally be presented in the methods, but it may be inadequately described or buried in the results section, figure legends, or tables.³⁶ For each experiment, animal variables include number of animals used, as well as their age (or weight), sex, background strain, and source, if purchased. A general statement about

management practices such as type of housing, light–dark cycle, diet, and microbial status is usually sufficient. However, if the study is likely to be particularly influenced by variability in these parameters (eg, the presence of *Helicobacter spp* in a colitis study), additional clarification of management practices may be indicated. Common genetic, environmental, and microbial variables that may confound a study have been previously reviewed.^{6,82} The experimental intervention should be clearly described and groups of control and experimental animals defined. Implementing the three Rs (replacement, reduction, and refinement) introduced by Russell and Burch in 1959 is a mainstay of experimental animal use,⁶³ and well-designed, humane studies on animals are consistent with good science.^{19,21,56,63,78} A good example of the principles listed above is given in the methods section of Yuan et al.⁸¹

Controls

Controls are selected relative to the intervention. Each time an additional variable (eg, diet or age in addition to genotype) is introduced, the number of animals increases. In many GEM studies, the genetic intervention is the only variable, and in this case, equal numbers of male and female age-matched litter-mates of each genotype (wild-type, homozygous mutant, and heterozygous) are used. In genetic studies, control littermates are typically available. In the event that they are not, controls of a similar background strain (and raised under similar conditions) may be used. However, greater variance within the control group may result in a corresponding increase in sample size to detect a difference between control and experimental groups. Similar numbers of control and experimental animals should be used to account for variation in normal background pathology such as tumor incidence.^{28,29,71} An alternative method is to refer to previously published wild-type data (historical controls). This method is very likely to result in bias as diet, environmental conditions including the number of mice per cage, breeding methods, strain subline differences, and other factors can play an important role in the incidence of lesions.

Reducing Bias

Systematic bias may arise in several ways,^{47,73} and not uncommonly, methods to reduce systemic bias are not reported in published papers.^{36,67} There are many randomization schemes ranging from assignment to treatment arms using a table of random numbers or stratified randomization to provide balance on some characteristic such as sex or age to adaptive randomization methods. Observational bias is introduced when the observer is aware of the genotype or intervention status of the subject. Ideally, the observer (pathologist or investigator) should be unaware of which genotype the animal belongs to, or to which arm a subject is assigned throughout the study. Thus, we suggest that whenever within-genotype random allocation is feasible, it should be performed in a blinded manner, such as by a technician who will not perform the outcome measures.

Sample Size Calculation to Demonstrate Clinically Meaningful Differences

The actual numbers of mice needed depends on variability of the outcome in both control and experimental groups. Sample size calculations typically require several pieces of information, such as the difference between groups to be detected (or frequency of the outcome for each treatment arm), the variation in the outcome for each group (these may be different), how many follow-up time points will be included, and expected survival of each treatment arm. Many of these inputs are from pilot experiments or from the literature. If one has not found significance based on a priori sample size calculations, then the initial inputs can be revisited. Was there more variation than expected? Is the difference observed a clinically meaningful difference? Adding mice to an experiment that is not demonstrating clinically meaningful differences is unlikely to achieve a robust outcome. Finally, the reality

of intrastrain variation should not be overlooked.⁵² Wild-type mice of the same strain may be used as controls for experiments,¹⁶ and if these mice are not litter-mates, intrastrain variation resulting from genetic drift may influence outcome.

Factorial Designs

Factorial experimental designs allow combinations of two or more design factors to be evaluated in one experiment, these are referred to as factorial designs.¹ These types of experimental designs may be 10 times as efficient as a series of two-armed (treatment and control) experiments, and they will likely reduce the number of animals used, allow for estimation of factor interactions (eg, genotype and diet), and increase the strength of the scientific findings.²¹

Appropriate Analytic Techniques

The appropriate analytic method should be selected to test the hypothesis, and the data should meet the assumptions of that test. Small sample sizes may not meet the assumptions of methods that assume a normal distribution (parametric methods). These are called means models and include the *t* test and analysis of variance. Sample sizes smaller than 30 that do not meet normality may result in biased results if a means model is used. Therefore, in these cases, nonparametric methods that do not assume normality and can be used to test median differences in an unbiased way are preferred.¹⁸ In a 2009 *Nature Cell Biology* editorial,¹⁴ guidelines for the analysis and presentation of data for articles submitted to *Nature* are presented. These articles specifically address reporting of results from data collected from small sample sizes.

Control for Correlated Outcomes Within a Subject

When several outcomes are measured at the same time (eg, metabolic compounds) or a single outcome is measured over time (eg, DHEA through a day), then the correlation among the measures within a subject needs to be accounted for, because common analytic methods assume all the observations are independent. When there are correlated measures, the variance will be biased, which typically inflates the type I error.⁷⁴ Working with a statistician will ensure proper methods for correlated outcomes.

Intent-to-Treat Analysis

Intent-to-treat analysis denotes analyzing the data according to the randomized groups, such as drug or diet, regardless of whether the treatment was adhered to by each subject. Bias can be induced by dropping those that cannot adhere because of intolerance of the treatment or even death, as well as by removing outliers. Mistakes include replacing mice that become too ill on a treatment or die and reporting only the results of mice who survived the experimental period. Exceptions to intent-to-treat analysis occur when there are protocol or equipment failures, such as accidental administration of incorrect dose of an infectious agent or drug, miscalibration of equipment, outbreak of disease in the colony, or losing animal identification. In situations such as these, the animal should be removed from the analytic dataset.

Interpretation and Presentation of the Results

Descriptive statistics such as sample size, distribution of data, mean (or median), and measures of variability are essential. The unit of measurement for all analyses should be clearly stated, and calculated *P* values rather than a range (eg, $P < .05$) given. When presenting data in graph form, error bars should be included, and if truncation of graphs along the y axis is used, it should be clearly indicated. Investigators should include mention of the type II error rate (false negative) when reporting nonsignificant results, as there may

have been insufficient power to detect a difference. Finally, correct nomenclature with regard to mouse strain is very important so that results from published GEM may be accurately referenced and compared.⁶⁹

External Validity

External validity means that the effect estimates from an initial study have been replicated in a separate cohort. In a 2006 Editorial in *Nature Cell Biology*,¹³ it was noted that the competitive publishing and funding climate has reduced the frequency of reproducing results within the same lab and that reproducibility is based on articles published within the same time frame. External validity supports whether observed changes can be attributed to the phenotype or intervention (ie, the cause) and not to other possible causes (sometimes described as “alternative explanations” for the outcome). In the context of this paper, alternative explanations most commonly include variations arising from mixed background strain or from the introduction of systematic bias.

Analyzing Causes of Death and Survival

An important aspect of evaluating aging in GEM studies is survival and cause of death analysis (CODA).^{39,81} In order to better characterize the GEM and to understand the biology of the condition induced and comparison to humans, it is of great importance to determine the cause of death (COD) of the GEM and compare the COD in GEM versus wild-type mice. For aging studies, this analysis assumes great importance. Why should new lines of aging mice have shorter or longer life spans than those of wild-type mice? What are the mechanisms for prolonging life? Can it include inhibition or prevention of tumor development, decreased degenerative aging diseases of major organs, or other causes? Few publications take on these important issues.^{4,17,40,81} Ladiges et al⁴⁰ suggest methods for evaluating mice in aging studies and validating findings in one study, yet no method for evaluating or comparing causes of death are noted. We suggest that for all aging mouse studies, a CODA analysis should be performed that includes methods of evaluating survival, clinical and anatomic pathology work-ups to assess COD, and inclusion of potential mechanistic indicators of aging (eg, insulin-like growth factor 1 levels).^{39,81} Statistical methods should be used for CODA for comparing wild-type versus GEM lines. Only then will meaningful publications on effects of aging by gene modification occur.

Including Relevant Expertise

Expertise in pathology and data interpretation, especially mouse pathology, is imperative for the conduct and publication of studies involving mice and mouse pathology.^{8,33} Examples of erroneous pathology²² or lack of statistical evaluation of the data⁶² appear in leading high-impact journals. Since the publication of these initial articles on two new lines of GEM, no subsequent articles have been published on research with these GEM, despite their initial publication in leading journals. The reviewers of manuscripts involving GEM are often experts in molecular biology or genetics, not pathology. This deficiency appears to be a major cause of publication of erroneous pathology diagnoses, especially in molecular biology and genetics journals. Pathologists familiar with mouse pathology should be reviewers of manuscripts that include mouse pathology.

Consequences of Poorly Designed Animal Studies

The study design and statistical analyses determine whether the hypothesis advanced in the introduction was adequately tested. Many GEM publications report studies with deficiencies in experimental design.^{5,24,54,67,8} In addition, data interpretation and conclusions can be based on limited evidence.^{8,33} Unfortunately, many published studies provide poor documentation of the findings in the study, especially involving pathology. These findings

can negate the conclusions and lead to errata, retractions, and at least, poorly reported publications.^{8,33} On the other hand, the lack of optimal design may not detract from the conclusions and significance of the studies, if obvious positive results are found. For example, consider a study in which knockout mice develop a high incidence of lymphoma at 6 months of age and in which no wild-type mice were used as controls. Lymphomas in 6-month-old mice are rare in most mouse lines and hopefully in the mice of the background used. More common, however, is a report that shows a few tumors of various types in 10 knockout mice at 12 months of age, with both sexes combined, and 10 wild-type controls have no or few tumors, and no statistical evaluations are used. These types of studies have been published in high-impact and other journals.^{8,33} Since these studies are already published, one can only use their publications as references even if the design and interpretation are not necessarily accurate or would require further studies to prove their hypotheses. Publication of a study does not necessarily mean that what it concludes is true or even that the study was well done.⁴⁹

Conclusions

Designing an animal experiment so that its conclusions can be trusted is often more complicated than one would expect. Failure to conclusively address a hypothesis may stem from a number of design errors, as well as the inevitable unexpected findings that accompany much research. Adequate study design requires considerable care, and is best done in consultation with various members of the team prior to initiating the project. A proposed study would benefit from statistical or pathologic review, as would completed articles submitted to journals. Finally, well-designed, humane studies on animals that implement the three Rs are consistent with good science.

Acknowledgments

The work for this report was funded in part by grants from the National Institute on Aging R21EY018719 and P30AG21342 at the Yale Claude D. Pepper Older Americans Independence Center.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

References

1. Allore HG, Murphy TE. An examination of effect estimation in factorial and standardly-tailored designs. *Clin Trials*. 2008; 5:121–130. [PubMed: 18375650]
2. Anan K, Yoshida N, Kataoka Y, et al. Morphological change caused by loss of the taxon-specific polyalanine tract in Hoxd-13. *Mol Biol Evol*. 2007; 24:281–287. [PubMed: 17065594]
3. Barbaric I, Miller G, Dear TN. Appearances can be deceiving: phenotypes of knockout mice. *Brief Funct Genomic Proteomic*. 2007; 6:91–103. [PubMed: 17584761]
4. Blackwell BN, Bucci TJ, Hart RW, et al. Longevity, body weight, and neoplasia in ad libitum-fed and diet-restricted C57BL6 mice fed NIH-31 open formula diet. *Toxicol Pathol*. 1995; 23:570–582. [PubMed: 8578100]
5. Bracken MB. Why are so many epidemiology associations inflated or wrong? Does poorly conducted animal research suggest implausible hypotheses? *Ann Epidemiol*. 2009; 19:220–224. [PubMed: 19217006]
6. Brayton C, Justice M, Montgomery CA. Evaluating mutant mice: anatomic pathology. *Vet Pathol*. 2001; 38:1–19. [PubMed: 11199155]
7. Brinster RL, Chen HY, Trumbauer M, et al. Somatic expression of herpes thymidine kinase in mice following injection of a fusion gene into eggs. *Cell*. 1981; 27:223–231. [PubMed: 6276022]

8. Cardiff RD, Ward JM, Barthold SW. "One medicine—one pathology": are veterinary and human pathology prepared? *Lab Invest.* 2008; 88:18–26. [PubMed: 18040269]
9. Chen M, Abele R, Tampé R. Functional non-equivalence of ATP-binding cassette signature motifs in the transporter associated with antigen processing (TAP). *J Biol Chem.* 2004; 279:46073–46081. [PubMed: 15322097]
10. Chen YF, Wu CY, Kao CH, et al. Longevity and lifespan control in mammals: lessons from the mouse. *Ageing Res Rev.* 2010; 9(Suppl 1):S28–S35. [PubMed: 20667513]
11. Crabbe JC, Wahlsten D, Dudek BC. Genetics of mouse behavior: interactions with laboratory environment. *Science.* 1999; 284:1670–1672. [PubMed: 10356397]
12. Dansky HM, Charlton SA, Sikes JL, et al. Genetic background determines the extent of atherosclerosis in ApoE-deficient mice. *Arterioscler Thromb Vasc Biol.* 1999; 19:1960–1968. [PubMed: 10446078]
13. Reproducing data. *Nature Cell Biol.* 2006; 8:541. Editorial.
14. How robust are your data? *Nature Cell Biol.* 2009; 11:667. Editorial.
15. Elder GA, Gama Sosa MA, De Gasperi R. Transgenic mouse models of Alzheimer's disease. *Mt Sinai J Med.* 2010; 77:69–81. [PubMed: 20101721]
16. Elmore SA, Peddada SD. Points to consider on the statistical analysis of rodent cancer bioassay data when incorporating historical control data. *Toxicol Pathol.* 2009; 37:672–676. [PubMed: 19516052]
17. Enns LC, Morton JF, Treuting PR, et al. Disruption of protein kinase A in mice enhances healthy aging. *PLoS One.* 2009; 4:e5963. [PubMed: 19536287]
18. Fay MP, Proschan MA. Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Stat Surv.* 2010; 4:1–39. [PubMed: 20414472]
19. Festing MFW. The scope for improving the design of laboratory animal experiments. *Lab Anim.* 1992; 26:256–267. [PubMed: 1447904]
20. Festing MF. The choice of animal model and reduction. *Altern Lab Anim.* 2004; 32(Suppl 2):59–64. [PubMed: 15601228]
21. Festing MF. Improving toxicity screening and drug development by using genetically defined strains. *Methods Mol Biol.* 2010; 602:1–21. [PubMed: 20012389]
22. Fu L, Pelicano H, Liu J, Huang P, Lee C. The circadian gene *Period2* plays an important role in tumor suppression and DNA damage response in vivo. *Cell.* 2002; 111:41–50. Erratum in: *Cell* 2002;111:1055. [PubMed: 12372299]
23. Funkat A, Massa CM, Jovanovska V, et al. Metabolic adaptations of three inbred strains of mice (C57BL/6, DBA/2, and 129T2) in response to a high-fat diet. *J Nutr.* 2004; 134:3264–3269. [PubMed: 15570023]
24. Giles J. Animal experiments under fire for poor design. *Nature.* 2006; 444:981. [PubMed: 17183281]
25. Gondo Y. Trends in large-scale mouse mutagenesis: from genetics to functional genomics. *Nat Rev Genet.* 2008; 9:803–810. [PubMed: 18781157]
26. Hanks MC, Loomis CA, Harris E, et al. *Drosophila engrailed* can substitute for mouse *Engrailed1* function in mid-hindbrain, but not limb development. *Development.* 1998; 125:4521–4530. [PubMed: 9778510]
27. Halder G, Callaerts P, Gehring WJ. Induction of ectopic eyes by targeted expression of the *eyeless* gene in *Drosophila*. *Science.* 1995; 267:1788–1792. [PubMed: 7892602]
28. Haseman JK, Hailey JR, Morris RW. Spontaneous neoplasm incidences in Fischer 344 rats and B6C3F1 mice in two-year carcinogenicity studies: a National Toxicology Program update. *Toxicol Pathol.* 1998; 26:428–441. [PubMed: 9608650]
29. Haseman, JK.; Elwell, MR.; Hailey, JR. Neoplasm incidence in B6C3F1 mice: NTP historical data. In: Maronpot, RR., editor. *Pathology of the Mouse: Reference and Atlas.* Vienna, IL: Cache River Press; 1999. p. 679–689.
30. Hempenstall S, Picchio L, Mitchell SE, et al. The impact of acute caloric restriction on the metabolic phenotype in male C57BL/6 and DBA/2 mice. *Mech Ageing Dev.* 2010; 131:111–118. [PubMed: 20064544]

31. Hill RE, Favor J, Hogan BL, et al. Mouse small eye results from mutations in a paired-like homeobox-containing gene. *Nature*. 1991; 354:522–525. [PubMed: 1684639]
32. Hosoda Y, Sasaki N, Agui T. Hypothyroid phenotype of the *Tpst2* mutant mouse is dependent upon genetic background. *Biomed Res*. 2010; 31:207–211. [PubMed: 20622471]
33. Ince TA, Ward JM, Valli VE, et al. Do-it-yourself (DIY) pathology. *Nat Biotechnol*. 2008; 26:978–979. [PubMed: 18779800]
34. Iraqi FA, Churchill G, Mott R. The Collaborative Cross, developing a resource for mammalian systems genetics: a status report of the Wellcome Trust cohort. *Mamm Genome*. 2008; 19:379–381. [PubMed: 18521666]
35. Jänne PA, Suchy SF, Bernard D, et al. Functional overlap between murine *Inpp5b* and *Ocr1l* may explain why deficiency of the murine ortholog for *OCRL1* does not cause Lowe syndrome in mice. *J Clin Invest*. 1998; 101:2042–2053. [PubMed: 9593760]
36. Kilkenny C, Parsons N, Kadyszewski E, et al. Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PLoS One*. 2009; 4:e7824. [PubMed: 19956596]
37. Kilkenny C, Browne WJ, Cuthill IC, et al. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLoS Biol*. 2010; 8:e1000412. [PubMed: 20613859]
38. Knight A. Systematic reviews of animal experiments demonstrate poor contributions toward human healthcare. *Rev Recent Clin Trials*. 2008; 3:89–96. [PubMed: 18474018]
39. Kodell RL, Farmer JH, Gaylor DW, et al. Influence of cause-of-death assignment on time-to-tumor analyses in animal carcinogenesis studies. *J Natl Cancer Inst*. 1982; 69:659–664. [PubMed: 7050501]
40. Ladiges W, Van Remmen H, Strong R, et al. Lifespan extension in genetically modified mice. *Aging Cell*. 2009; 8:346–352. [PubMed: 19485964]
41. Lavigne A, Maltby V, Mock D, et al. High incidence of lung, bone, and lymphoid tumors in transgenic mice overexpressing mutant alleles of the *p53* oncogene. *Mol Cell Biol*. 1989; 9:3982–3991. [PubMed: 2476668]
42. Lee WP, Tzou WS. Computational methods for discovering gene networks from expression data. *Brief Bioinform*. 2009; 10:408–423. [PubMed: 19505889]
43. Lincecum JM, Vieira FG, Wang MZ, et al. From transcriptome analysis to therapeutic anti-CD40L treatment in the *SOD1* model of amyotrophic lateral sclerosis. *Nat Genet*. 2010; 42:392–399. [PubMed: 20348957]
44. Liu Q, Saveliev A, Pierce EA. The severity of retinal degeneration in *Rplh* gene-targeted mice is dependent on genetic background. *Invest Ophthalmol Vis Sci*. 2009; 50:1566–1574. [PubMed: 19060274]
45. Lynch VJ, Wagner GP. Resurrecting the role of transcription factor change in developmental evolution. *Evolution*. 2008; 62:2131–2154. [PubMed: 18564379]
46. Lynch VJ. Use with caution: developmental systems divergence and potential pitfalls of animal models. *Yale J Biol Med*. 2009; 82:53–66. [PubMed: 19562005]
47. Macleod MR, Fisher M, O'Collins V, et al. Good laboratory practice: preventing introduction of bias at the bench. *Stroke*. 2009; 40:e50–e52. [PubMed: 18703798]
48. Manolio TA. Genomewide association studies and assessment of the risk of disease. *N Engl J Med*. 2010; 363:166–176. [PubMed: 20647212]
49. Martinson BC, Anderson MS, de Vries R. Scientists behaving badly. *Nature*. 2005; 435:737–738. [PubMed: 15944677]
50. Mathieu P, Lemieux I, Després JP. Obesity, inflammation, and cardiovascular risk. *Clin Pharmacol Ther*. 2010; 87:407–416. [PubMed: 20200516]
51. McClearn GE. Contextual genetics. *Trends Genet*. 2006; 22:314–319. [PubMed: 16678298]
52. Meyer CW, Elvert R, Scherag A, et al. Power matters in closing the phenotyping gap. *Naturwissenschaften*. 2007; 94:401–406. [PubMed: 17216184]
53. Michaud EJ, Culiat CT, Klebig ML, et al. Efficient gene-driven germ-line point mutagenesis of *C57BL/6J* mice. *BMC Genomics*. 2005; 6:164. [PubMed: 16300676]

54. Mignini LE, Khan KS. Methodological quality of systematic reviews of animal studies: a survey of reviews of basic research. *BMC Med Res Methodol*. 2006; 6:10. [PubMed: 16533396]
55. Nishino T, Sasaki N, Nagasaki KI, et al. Genetic background strongly influences the severity of glomerulosclerosis in mice. *J Vet Med Sci*. 2010; 72:1313–1318. [PubMed: 20484839]
56. Osborne NJ, Payne D, Newman ML. Journal editorial policies, animal welfare, and the 3Rs. *Am J Bioeth*. 2009; 9:55–59. [PubMed: 20013503]
57. Palmiter RD, Brinster RL. Transgenic mice. *Cell*. 1985; 41:343–345. [PubMed: 2985274]
58. Perel P, Roberts I, Sena E, et al. Comparison of treatment effects between animal experiments and clinical trials: systematic review. *BMJ*. 2007; 334:197. [PubMed: 17175568]
59. Ranganayakulu G, Elliott DA, Harvey RP, et al. Divergent roles for NK-2 class homeobox genes in cardiogenesis in flies and mice. *Development*. 1998; 125:3037–3048. [PubMed: 9671578]
60. Richter SH, Garner JP, Würbel H. Environmental standardization: cure or cause of poor reproducibility in animal experiments? *Nat Methods*. 2009; 6:257–261. [PubMed: 19333241]
61. Rivera J, Tessarollo L. Genetic background and the dilemma of translating mouse studies to humans. *Immunity*. 2008; 28:1–4. [PubMed: 18199409]
62. Ruggero D, Grisendi S, Piazza F, et al. Dyskeratosis congenita and cancer in mice deficient in ribosomal RNA modification. *Science*. 2003; 299:259–262. [PubMed: 12522253]
63. Russell, WMS.; Burch, RL. *The Principles of Humane Experimental Technique*. London: Methuen; 1959.
64. Sandgren EP, Luetkeke NC, Palmiter RD, et al. Overexpression of TGF alpha in transgenic mice: induction of epithelial hyperplasia, pancreatic metaplasia, and carcinoma of the breast. *Cell*. 1990; 61:1121–1135. [PubMed: 1693546]
65. Schlosser G. Epistasis, constraints, and coevolution. *Evol Dev*. 2009; 11:459–461. [PubMed: 19754702]
66. Sieberts SK, Schadt EE. Moving toward a system genetics view of disease. *Mamm Genome*. 2007; 18:389–401. [PubMed: 17653589]
67. Strasak AM, Zaman Q, Marinell G. The use of statistical methods in medical research: a comparison of the New England Journal of Medicine and Nature Medicine. *Am Stat*. 2007; 61:47–55.
68. Sundberg KK, Doukkali E, Lampic C, et al. Long-term survivors of childhood cancer report quality of life and health status in parity with a comparison group. *Pediatr Blood Cancer*. 2010; 55:337–343. [PubMed: 20582940]
69. Sundberg JP, Schofield PN. Commentary: Mouse genetic nomenclature: standardization of strain, gene, and protein symbols. *Vet Pathol*. 2010; 47:1100–1104. [PubMed: 20685919]
70. Tang T, Li L, Tang J, et al. A mouse knockout library for secreted and transmembrane proteins. *Nat Biotechnol*. 2010; 28:749–755. [PubMed: 20562862]
71. Tarone RE, Chu KC, Ward JM. Variability in the rates of some common naturally occurring tumors in Fischer 344 rats and (C57BL/6N x C3H/HeN)F1 (B6C3F1) mice. *J Natl Cancer Inst*. 1981; 66:1175–1181. [PubMed: 6941047]
72. Thomas KR, Capecchi MR. Site-directed mutagenesis by gene targeting in mouse embryo-derived stem cells. *Cell*. 1987; 51:503–512. [PubMed: 2822260]
73. van der Worp HB, Howells DW, Sena ES, et al. Can animal models of disease reliably inform human studies? *PLoS Med*. 2010; 7:e1000245. [PubMed: 20361020]
74. Verbeke, G.; Molenberghs, G. *Linear mixed models in practice*. In: Verbeke, G.; Molenberghs, G., editors. *A SAS-Oriented Approach*. New York: Springer; 1997.
75. Yu H, Huang J, Qiao N, Green CD, Han JD. Evaluating diabetes and hypertension disease causality using mouse phenotypes. *BMC Syst Biol*. 2010; 4:97. [PubMed: 20642857]
76. Wahlsten D, Metten P, Phillips TJ, et al. Different data from different labs: lessons from studies of gene-environment interaction. *J Neurobiol*. 2003; 54:283–311. [PubMed: 12486710]
77. Wang J, Farr GW, Zeiss CJ, et al. Progressive aggregation despite chaperone associations of a mutant SOD1-YFP in transgenic mice that develop ALS. *Proc Natl Acad Sci U S A*. 2009; 106:1392–1397. [PubMed: 19171884]

78. Workman P, Aboagye EO, Balkwill F, et al. Guidelines for the welfare and use of animals in cancer research. *Br J Cancer*. 2010; 102:1555–1577. [PubMed: 20502460]
79. Wu S, Lusis AJ, Drake TA. A systems-based framework for understanding complex metabolic and cardiovascular disorders. *J Lipid Res*. 2009; 50(Suppl):S358–363. [PubMed: 19033210]
80. Wurst W, de Angelis MH. Systematic phenotyping of mouse mutants. *Nat Biotechnol*. 2010; 28:684–685. [PubMed: 20622838]
81. Yuan R, Tsaih SW, Petkova SB, et al. Aging in inbred strains of mice: study design and interim report on median lifespans and circulating IGF1 levels. *Aging Cell*. 2009; 8:277–287. [PubMed: 19627267]
82. Zeiss CJ. Mutant mouse pathology: an exercise in integration. *Lab Anim (NY)*. 2002; 31:34–39. [PubMed: 11910406]

Table 1**Checklist of Design Parameters to Consider When Performing Phenotyping Studies**

I.	Hypothesis
	Is the intervention and its relationship to the outcome stated?
	Can the outcome be statistically tested?
II.	Description of animal variables
	Are numbers of animals, age, sex, background strain, and source stated?
	Are management variables such as housing, light–dark cycle, diet, microbial status, and veterinary care stated?
III.	Experimental design
	Is the experimental intervention defined?
	Are control and experimental groups defined relative to the intervention?
	Are control and experimental groups similar except for the intervention?
	Are numbers of controls appropriate to two-armed (more controls) or factorial (fewer controls) designs?
	Are sample size calculations or justification provided?
	Is the observer blinded (ie, unaware of genotype or treatment status)?
	If relevant, is randomization described?
	Are technologies and approaches used to record outcome described?
	Are humane end points for aging studies or those resulting in illness defined?
IV.	Analytic methods
	Are the statistical methods used to test the hypothesis appropriate?
	Do the data meet the assumptions of the test?
	Are nonparametric methods used for small or non-normally distributed datasets?
	Are measured outcomes independent? If not, correlated outcomes must be controlled for.
	Are all animals included in the analysis? If not, is there a justification?
V.	Reporting
	Are descriptive statistics such as sample size, distribution of data, and measures of variability given?
	Is the unit of measurement in all analyses clearly stated?
	If relevant, is data transformation explained?
	Include calculated <i>P</i> values as opposed to ranges (eg, < .05) when reporting a significant difference between experimental and control groups.
	Do graphs reflect effect sizes correctly, and are error bars present?
	Is rodent nomenclature correct?
	Is reference made to external validity (ie, independent replication of experiment by same or other labs)?