Introducing Therioepistemology: the study of how knowledge is gained from animal research

Joseph P Garner[1,2], Brianna N Gaskill[3], Elin M Weber[1], Jamie Ahloy-Dallaire[1] & Kathleen R Pritchett-Corning[4,5]

This focus issue of *Lab Animal* coincides with a tipping point in biomedical research. For the first time, the scale of the reproducibility and translatability crisis is widely understood beyond the small cadre of researchers who have been studying it and the pharmaceutical and biotech companies who have been living it. Here we argue that an emerging literature, including the papers in this focus issue, has begun to congeal around a set of recurring themes, which themselves represent a paradigm shift. This paradigm shift can be characterized at the micro level as a shift from asking "*what have we controlled for in this model?*" to asking "*what have we chosen to ignore in this model, and at what cost?*" At the macro level, it is a shift from viewing animals as tools (the furry test tube), to viewing them as patients in an equivalent human medical study. We feel that we are witnessing the birth of a new discipline, which we term *Therioepistemology*, or the study of how knowledge is gained from animal research. In this paper, we outline six questions that serve as a heuristic for critically evaluating animal-based biomedical research from a therioepistemological perspective. These six questions sketch out the broad reaches of this new discipline, though they may change or be added to as this field evolves. Ultimately, by formalizing therioepistemology as a discipline, we can begin to discuss best practices that will improve the reproducibility and translatability of animal-based research, with concomitant benefits in terms of human health and animal well-being.

We are honored to write the opening article of this focus issue of *Lab Animal*. This focus issue could not occur at a more important time for biomedical research and the use of animals in science in general. The progressive worsening of success rates in human trials (currently 1 in 9 drugs entering human trials will succeed)[1–4], combined with the explosion of interest in the reproducibility crisis[5–8] and the recognition that most drugs fail in human trials due to insufficient efficacy[1,2,4], has led to a growing suspicion that the failure of translation from animal work to human outcomes may in some way reflect issues in animal research itself[5–19]—after all, every drug that fails in humans "worked" in an animal model. Indeed pharmaceutical companies continue to disinvest in internal animal R&D, a trend begun in the last decade, passing on the cost and risk to academia and startups[5,20]. Even this approach is not foolproof as pharmaceutical companies often cannot replicate the results of published work from academia[5,8,13]. Accordingly, there is a growing trend to focus on human, not animal, work for basic discovery[17].

Discarding animal research entirely is not the answer. When properly used, animal models have incredible value, not least the ability to follow biomarkers from birth to disease onset in a year or less (in the case of mice), which is impossible in humans[17,18]. There are patterns and principles that can help us identify models and results that are more or less likely to translate, and there are also easily realized, simple changes in the execution of animal work that will inherently improve translation[16–18]. This isn't a new concept; looking back over the last 10–15 years we can see many authors have been candid about the merits, strengths, weaknesses, reproducibility, and translatability of various animal models[1,2,4–19,21–41]. Our goal with this article is to unite the common themes in this broader emerging literature and this special issue.

Thus, the central point is that we (*i.e.*, refs. 17,18,26–28,39,41) do not represent a voice in the wilderness, but one voice in a chorus and that this emerging literature[1,2,4–19,21–40,42,43] reflects a nascent discipline which can be codified as the study of how knowledge is gained from animal research. We propose the title "*Therioepistemology*"

[1]Stanford University, Department of Comparative Medicine, Stanford, California, USA. [2]Stanford University (by courtesy) Department of Psychiatry and Behavioural Sciences, Stanford, California, USA. [3]Purdue University Animal Science Department, West Lafayette, Indiana, USA. [4]Department of Comparative Medicine, University of Washington, Seattle, Washington, USA. [5]Harvard University Faculty of Arts and Sciences, Office of Animal Resources, Cambridge, Massachusetts, USA. Correspondence should be addressed to J.P.G. (jgarner@stanford.edu).

for this new discipline. "Epistemology" in philosophy is the study of the theory of knowledge, and the mechanisms by which rational inference is formed; while the prefix "therio" indicates of, or from, animals.

The papers that make up this emerging literature consistently address the same set of questions, which is all the more remarkable given that most of these authors have been thinking and writing in isolation. We believe that these are questions that every scientist should ask of themselves when they plan, interpret, and publish a study, and again when making translational, husbandry, or policy decisions on the basis of a study. We believe that by asking these questions, codifying them as a discipline, and educating our trainees to ask them, therioepistemology will guide biomedical research and the many disciplines that study animals to much more effective animal work in the future.

All models are imperfect: how that imperfection affects inference is what matters

At its core, therioepistemology is an applied exploration of validity. Validity is the degree to which a measure or an experimental result means what it is claimed to mean. This is fundamentally distinct from reliability (the degree to which a measure gives the same value under different circumstances) or reproducibility (the degree to which an experiment gives the same result when repeated)[17,18,44]. However these concepts are widely confused. To an extent this is understandable, as one does sometimes follow from the other, at least in a narrow sense. For instance, in molecular biology, finding the same bands on the same gel three times in a row (reliability) indicates an inherent truth about the macromolecules present in the sample (validity). However, the scope of this truth is extremely limited: the reliability of the measurement does not necessarily imply anything about the functions or roles of these macromolecules within a complex system such as an animal model, and generalizing these results in that way is problematic. Ten phrenologists might reliably find the same bumps on our heads but these reliable findings don't mean that the bumps have predictive power of our intelligence or criminality. As we discuss below, this failure to recognize the limits of inference is of particular importance to the over-interpretation of genetically modified mice as we and many others have argued[17,32,33,45].

In brief, validity can be thought of as three independent dimensions: face *vs* construct *vs* predictive validity; internal *vs* external validity; and convergent *vs* discriminant validity. **Table 1** provides definitions and examples. Note that while the poles of each dimension are conceptually distinct, they are not mutually exclusive in the sense that a given measure, model, or study could for instance show both internal *and* external validity. For in depth discussion of validity see refs. 17,18,24,44,46,47. Therioepistemology provides a framework to ask "*what type of validity is claimed, what type has been shown, and which is relevant to the research question at hand?*" It is remarkable how often the answers to these three questions are incongruous. Yet if the type of validity claimed and shown doesn't bear on the research question, the whole experiment can be worthless, if not actively misleading (see Table 2 in ref. 17 for a worked example).

With respect to the particular problem of attrition, we are concerned with predictive, external, discriminant validity (*i.e.* can the model predict a human outcome, and can it correctly avoid false positives). However we often rely on measures with, at best, face or construct, internal, convergent validity. For example, mouse measures of pain are largely based on reflexive or guarding responses, not the emotional experience of pain itself[48,49]. They show construct, internal, convergent validity. This is a fine model for studying the basic biology of nociception, but an incredibly poor choice for discovery of novel analgesics[48–50], precisely because an ideal analgesic in humans leaves reflexes intact but blunts the emotional experience of pain[49]. Thus, a good model for one question can be a terrible model for another, and therioepistemology helps us think this through.

Six questions: what do we choose to ignore?

We often think about experimental design in terms of which model organism or system we use and which variables we choose to manipulate, which we choose to measure and how, and which we choose to control or 'standardize' in order to minimize variability in measured outcomes. Our choices here are typically at least somewhat arbitrary, a matter of ease or convenience or what can be automated, and often implemented without knowledge of the effects on the animal and on the validity of the experiment.

Choosing a particular model species inherently involves choosing an animal with biological differences to humans. Ignoring these differences can be disastrous, if, for instance, normal species differences in anatomy are mistaken as pathological[51–53]. Similarly, the complexities of human diagnoses are ignored or glossed over at our peril[17,48,49,54,55]. A recurrent theme of this paper, and this special issue, is that the role that good animal well-being plays in good science also cannot be ignored[26,41–43,56–58]. Furthermore, factors that are controlled or standardized have also been widely shown to affect both the model and scientific outcomes, from the infectious agents we choose to exclude[59], to the choice of bedding materials[60–62] and enrichment[63–65], to cage changing practices[66], to handling technique[58], to the identity[31] or sex[67] of the experimenter. In some of these examples, animal health, animal well-being, and scientific quality are clearly improved by implementing a standard such as excluding pathogens that kill animals[68], or adding nesting enrichment[64,65]; in others they are clearly impaired, as when animals are subjected to "forced" handling techniques[58], or by the over-exclusion of infectious agents[59]. But, in most cases it is unclear whether any particular standard is the "right" thing to do.

Thus, the decision to conduct an experiment in a certain way is inherently a choice to ignore other aspects of the problem under study. Controlling or standardizing a particular variable also means that we choose to ignore the potential effects on study outcomes of doing so. Accordingly, therioepistemology asks us to make a frame shift, where we focus on acknowledging what we choose to ignore and our reasons for doing so, and on clearly understanding the effect this has on the model and the experiment. Therioepistemology is distinct from the formalization of reporting advocated in the ARRIVE guidelines[69]. The ARRIVE guidelines are valuable in setting a minimum standard for reporting what was controlled, not what was ignored or unknown, and focus on study reporting rather than study planning.

Table 1 | Three Dimensions of Validity (adapted from ref. 17)

Dimension	Subtype	Definition and examples
Face *vs.* Construct *vs.* Predictive	Face	Does the measure or model appear outwardly similar to what it is supposed to measure or model in terms of behavior, phenomenology, epidemiology *etc.*? (*e.g.*, Does a fear measure resemble fear responses for the species? Do the animal symptoms resemble the symptoms seen in human patients?)
	Construct	Does the measure or model involve the mechanism or processes that it is supposed to measure or model (at physiological, immunological, or neuropsychological levels *etc.*)? (*e.g.*, Can the measure actually access these processes? Is the methodology consistent with the theory behind the measure? Does an animal model involve the same physiology as the human measure or condition? Does an animal model show the same medical signs, or the same biomarkers, as the human condition?)
	Predictive	Does the measure or model actually predict outcomes it is supposed to? (*e.g.*, Does a behavioral stress measure predict stress hormone levels? Does an animal model predict human drug response? Does the animal model respond *only* to treatments that successfully treat human patients?)
Internal *vs.* External	Internal	Is the methodology and results of the measure or model consistent with *both* the theory and existing data from the model system? (*e.g.*, Is the methodology consistent with the mathematics describing the measured properties? Is the measure ecologically relevant to the test species? Does the measure agree with other measures of the same property in the same individuals?)
	External	Are results from the measure or model broadly applicable? (*e.g.*, Is the kind of fear measured in a fear test broadly applicable to the kind of fear being modeled in humans? Does the model give consistent results across a range of environmental conditions that accurately reflect the range of environmental conditions experienced by human patients?)
Convergent *vs.* Discriminant	Convergent	Does the measure or model show broad agreement with properties of the thing being measured, or properties of the human condition being modeled? (*e.g.*, Are different measures of fear correlated? Does the model show similar behaviors to the human condition? Do drugs that treat human patients also treat model symptoms? Is the gene knocked out in the model also downregulated in human patients? Do mechanisms in the model mirror those in humans?)
	Discriminant	Does the measure or model exclude alternative processes or differential diagnoses? (*e.g.*, Is a fear measure 'clean', or is it correlated with measures of other behavioral traits? Does the model show behaviors, physiology, signs, biomarkers or symptoms atypical of the human conditions, or typical of a differential diagnosis to the human condition? Do drugs that fail to treat humans also fail to treat the model? Do all human patients show downregulation of the gene knocked out in the model, or only a subset? Do mechanisms that distinguish human disorders or subtypes also distinguish the animal models?)

Definitions are given, with example tests. Note that tests of validity often involve more than one of these dimensions (for example, when a drug works in a mouse but fails in humans, this is a failure of predictive, external, convergent validity). For additional discussion see refs. 17,18,24,44,46,47.

To formalize this process we provide a list of six questions that cover (in our experience at least) the vast majority of problematic animal work we have encountered. To a degree these questions overlap: for instance we singled out animal well-being as a distinct question in order to emphasize the central importance of good well-being to good science, but consideration of animal well-being overlaps with several of the other questions. We welcome changes and additions to these questions as the discipline of therioepistemology evolves.

What features of model biology are ignored?

An animal model, by definition, is not a perfect homolog to a human patient. This question focuses on ways in which the animal's biology, by virtue of its species, housing, or genetic manipulation, is unlike the human. We provide examples of three common issues.

Model biology is ignored through ignorance of species differences. Animals can differ physiologically, anatomically, and cognitively from humans. Being unaware of these differences can be ruinous to animal research. Many animals have anatomical features that humans do not, and consulting a veterinarian or veterinary pathologist when designing the experiment and before publication is wise. For example, there are several papers in which the paired subcutaneous preputial glands of mice were identified as various types of neoplasia including teratomas and squamous cell carcinomas[51–53]. The consequences for the validity of experiments in which these normal structures are seen as abnormal should be obvious. In fact the cancer literature often blames fundamental biological differences between rodents and humans for the particularly high attrition rate (at least of non-biologics)[9,19], but perhaps the real issue is not differences, but ignorance of these differences. For instance, we can contrast anti-neoplastics with cardiovascular drugs, which have the highest translation rate of any class[1]. Differences in heart physiology between humans and small rodents are just as profound—mass of the heart and electrical regulation of the heartbeat in mice and rats is such that fibrillation is difficult to induce. However, in this field the effect of this difference on the models has been intensively studied[70,71], suggesting that basic biological differences are not a major contributor to attrition, as long as we knowledgably compensate for them.

Model biology is ignored through managing only what we can monitor. Husbandry systems have an inherent tendency to manage

things that are either easily measured or matter to humans (light levels, air exchange rates, temperature, *etc.*), which may not be things that matter to model species. Concomitantly, we have a tendency not to manage or measure those things that do not matter to us. This is particularly important in the case of mice, which are highly adapted to living cryptically as human commensals—in fact, their stealth adaptations rely on using sensory ranges such as ultraviolet and ultrasound that we don't detect and being active when we aren't. Furthermore, as discussed elsewhere in this special issue[41], the critical issue for well-being and model quality is control, not of the animal, but by the animal. Through overengineering animal housing we take away an animal's control of its environment, which in turn makes it fundamentally abnormal[41,72]. This argument is traditionally a stress-psychology one, although it has recently been expanded to immunology. By intensively managing infectious agents because they can be measured, we have inadvertently created mice whose immune systems never develop beyond a neonatal naivety. Thus, "clean" lab mice have an immune system that may be a good model for human neonates, but is a terrible model for adult humans, whereas "dirty" mice have an immune system which does model adult humans[59]. This inadvertent immunological manipulation may be critical in explaining poor translation in cancer[59].

Model biology is ignored through ignoring experimentally induced changes in unintended aspects of model biology. Experimenters intending to manipulate one isolated aspect of an animal's biology may simultaneously and unintentionally also tweak others, an issue that manifests in two ways. Either investigators are unaware of this limitation of the model arising from its biology, from its housing, or from its methodology; or they choose to ignore it as a necessary evil. Genetically modified mice provide two excellent examples. First, the "linked gene problem"[32,33]—or the fact that the process of creating a congenic by backcrossing a chimera to a different strain reduces the proportion of the genome that is derived from the original embryonic stem cell donor strain, but that simple linkage ensures that the manipulated gene is flanked by linked DNA from the donor strain. There may be hundreds of linked genes in this flanking DNA, and although their number decreases each backcrossing generation, it never reaches zero. Thus, any phenotype, especially any unexpected phenotype, is much more likely to be due to linked genes than the manipulated gene[32,33]. For a long time this issue was ignored as a necessary evil, even though it potentially invalidates traditional genetically modified mouse models, and is actually relatively easy to work around with appropriate breeding schemes and experimental design[33].

Second, the "overlapping gene problem"—technologies like Cre-lox or CRISPR, which supposedly cleanly alter gene expression, often do not[73]. Roughly 10% of genes in the mouse genome have overlapping reading frames[74]. Typically this is in opposite strands, but genes can overlap on the same strand[74]. As a result, inserting *loxP* sequences on either side of a gene may introduce nonsense mutations in overlapping genes, and CRISPR deletion may delete portions of overlapping genes. This is particularly troubling because overlapping gene pairs are not well conserved between mice and humans[74], so an unintentional loss of function in the mouse is unlikely to also occur in humans. So once again, ignoring this problem as a necessary evil raises the risk that an observed phenotype, especially an unexpected phenotype, may not be truly due to the gene manipulated.

Besides introducing confounds that may drive experimental results, such perceived necessary evils can additionally interfere with measurements. For example, in a macaque model of diabetes, physical restraint of non-human primates for blood collection activates a stress response that causes rapid changes in blood glucose levels[75,76]. Habituation to the procedure or use of voluntary handling techniques based on positive reinforcement training can allow researchers to obtain blood with a comparatively minimal stress response; these measures will more closely reflect the animal's biology, rather than the intensity of the aversive experience itself[75–77].

What features of human biology are ignored?

This question focuses attention on essential features of the human disease that are ignored such that the model or the measure can't actually be a meaningful homolog to the human. At best the animal work lacks specificity, but often because the wrong thing is being modeled, any results are of little relevance to the human disease.

Human biology is ignored through ignorance of human diagnostic criteria. This is a particular issue in psychiatric models. For instance, a diagnosis of obsessive-compulsive disorder (OCD) requires the exclusion of differential diagnoses such as stereotypies or trichotillomania, yet we are unaware of any proposed mouse model of OCD where the "OCD-like" behavior would not be an exclusionary differential (like hair pulling, stereotypies, or self-injury) in humans[17]. Non-clinical researchers are often ignorant of the specific inclusionary diagnostic criteria as well—in the case of OCD, those criteria state that compulsions are performed to relieve the anxiety of experiencing the obsession. So if we can't measure the obsession in a mouse, we can't meaningfully call a repetitive behavior a compulsion. This leads to the second reason for ignoring human biology.

Human biology is ignored because human symptoms are considered unmeasurable. Again this is a common issue in psychiatric models, where part of the human diagnosis is based on patient reports of internal experiences; but is technically true for all diseases (as symptoms are patient reported, whereas signs are objectively observed). Obsessions in OCD are a clear example, as is catastrophic thinking in depression[78]. In both cases these are clinically relevant symptoms because treating them can be key to treating the disorder as a whole. This problem might seem intractable, but it is not[79]. For instance, in psychiatry we know of many neuropsychological biomarkers that are uniquely correlated with these private symptoms. The animal well-being literature in particular has reverse-translated many of these biomarkers to measure subjective states in animals[54,78–81]. Indeed biomarkers—or necessary control points in disease development, such as insulin resistance in metabolic syndrome—provide a general solution to many of the issues we raise here. If the same biomarker can be measured in the same way then there is no need for the "-like" measures that are so prone to false discovery[17,18,82].

Human biology is ignored because doing something is perceived as better than doing nothing. This argument is usually

posited by researchers who are highly motivated to help the patient population, and as such, acknowledge the limitations of the model. The thinking is that if this is the only model or measure that we have to work with, then we have no choice but to use it and hope that it will tell us something. Criticizing or abandoning a model or measure is often characterized as throwing the baby out with the bathwater. For instance, autistic children show impairments in making eye contact, social play, and in theory of mind; but so do mice as a species[55]. It is nonsensical to attempt to model the pathological absence of a cognitive function in an animal that does not possess it in the first place. When researchers ignore this issue, they are ignoring the fact that there never was a baby in the bathwater in the first place, and thus any result is a false discovery. Autism also shows us how to resolve this problem—sometimes we may have to pick different species for different components of the diagnosis. Developmental delay and repetitive behaviors can be meaningfully modeled in mice[83], but we will have to turn to species that do possess complex social cognition in order to model the social dimension of the disorder[55].

As we have argued before[17], the suffix "-like" does not resolve this or any of the problems discussed so far: either the measure or model is or is not homologous to the human symptom or condition. "-like" merely indicates that the behavior is known to not be homologous, or it is being used without validation, and neither would be acceptable in other disciplines[17]. "-like" is also inherently dangerous because it encourages the ongoing use of a model or measure that is *un*like the human disease, and thus will generate nothing but false positive results in terms of translation.

Human biology is ignored through reductionism. Humans are complex and messy, as are our diseases and their treatment. Reductionism attempts to understand the world by ignoring this complexity, which works for basic science but is a bad idea for translational research. Disease genetics illustrates the fallacy inherent in translational research reductionism. Few diseases are the result of the malfunction of a single gene. Instead, multiple genes confer risk, and more importantly, it is the activation or inactivation of particular genes or the interaction of those risky genes with the environment that allows a disease to develop. These differentially regulated genes are the interesting targets therapeutically, because they not only control disease development but they can be used to both detect a disease before it is fully developed, and to stop it. They are also the hardest genes to find with a typical genome-wide association study approach. The reductionist answer is to seek out rare or *de novo* mutations which confer Mendelian inheritance, in which case only a few individuals are needed to find the culprit. For instance, a handful of families have been identified with loss of function mutations in *SLITRK1*, and near Mendelian inheritance of Tourette's Syndrome and trichotillomania[84,85]; but the resulting knockout mice[86] are at best models of only these few families, not the disorders as a whole[17]. Furthermore, such highly penetrant rare *de novo* mutations may be the easiest to find, but they are of the least interest clinically. Precisely because they are penetrant, they are unlikely to be malleable. As with our other examples, there is a way out of this trap. Animals like mice, which go from conception to middle age in under a year, are incredibly valuable if the model is hypothesis-driven and biomarker-based. Following biomarkers through disease development serves the same purpose as looking for rare individuals in humans—it narrows down the number of candidates to the point where it is viable to test for them in a broader clinical population. But instead of identifying rare genes of little broader relevance, biomarker-based animal models allow us to identify the differentially regulated genes that are viable targets for intervention[17,87].

What features of the measures are ignored?

With the notable exception of reverse translated biomarkers[17,18,54,78,82], measures in animal models, like models themselves, are often only approximations of the signs and symptoms of human disease. This question overlaps with the previous questions, but serves to focus attention on the measures taken in the animal model, because however good the model, if the measures taken are flawed, then the experiment is bound to produce spurious results.

Features of the measure are ignored through ignorance of the discipline from which a measure is borrowed. In modern biomedical research the range of disciplines involved in any project, basic or translational, is too broad for any one investigator to be an expert. Problems arise when there is an implicit bias or asymmetry in the perceived skill required for different components of a project. Nowhere is this more true than behavioral phenotyping—an ethologist attempting to do genetic work without genetics expertise would never survive peer review, but the behavioral genetics literature is characterized by geneticists assessing behavioral phenotypes with no behavioral training[88]. Accordingly, the majority of behavioral phenotyping tasks have been thoroughly discredited by ethologists and experimental psychologists[29,30,36,38,45] (for a comprehensive review, see refs. 17,18). Yet almost every medical school has a behavior core churning out data from the same discredited methods. Ignorance of basic behavioral methodology has led to essential quality controls being sacrificed in the interest of throughput and automation, the result being garbage in, garbage out[17,18,29,30,36,38,45].

Ignoring evidence that the measures are flawed. Aside from theoretical considerations, the previous example is neatly illustrated by a lack of internal convergent construct validity (*i.e.*, that often, measures of the same supposed quantity do not agree[89]); internal discriminant construct validity (*i.e.*, that alternative deficits affect the measure[36]); external construct validity (for example, measures interpreted as being those of a trait are actually being determined by state[29,30,90]); or internal predictive validity (for example, that the ability of a measure to detect different classes of drugs depends on how the animals were handled[30,67]). Again, such choices are often justified by the "this is the best we have" argument, but this is a fallacy of argument from ignorance—it is simply not true. Behavioral neuroscience has developed far better measures which could have been used, if only ethologists or experimental psychologists had been consulted[17,18,54,78,91,92].

Ignoring sensitivity *vs* specificity with respect to the research question. Sensitivity is the proportion of truly positive individuals identified by a measure. Specificity is the proportion of truly negative individuals identified by a measure. If hypothesis-free screening is the purpose of the work (as is the case for phenotyping in general), then there is nothing wrong with using overly sensitive

methods, as long as appropriate corrections are employed for multiplicity and false discovery rates, which is rarely the case[17]. Indeed, almost all other areas of biomedical research where hypothesis-free discovery is the goal have come to understand the need for false-discovery correction[93], yet this is strangely lacking for phenotyping in general in animal models[17]. Indeed attrition can be boiled down to an imbalance between sensitivity and specificity—in animal work we are seduced by the possibility of a result (sensitivity) and do not attempt to rule out a false discovery (specificity). Again, the OCD example is illustrative—all that is repetitive is "OCD-like" (sensitivity), but little repetitive behavior is truly OCD (specificity). Thus, if our research question is truly "do I have a model of X"?, then we should be employing highly specific measures instead of, or as a follow-up to, highly sensitive phenotyping measures[17,18]. This problem is more of a challenge of changing the business-as-usual mindset than anything, as it is readily solved by adopting highly specific biomarkers, which may require reverse translation, but can also be as simple as adapting a human assay kit for use in animals[54,78,82,88].

What features of background methodology and husbandry are ignored?

Experiments do not occur in a vacuum. We tend to focus on the experiment itself, and not the supporting scaffolding or "experimental background" that surrounds it. Reporting standards[69], while valuable, do not answer the question of what was ignored when the experiment was designed. This question focuses attention on what aspects of experimental background are being ignored at the cost of experimental validity.

Experimental background is ignored because this is how it's always been done, and through fear that the model will stop "working". Experimenters are often very resistant to changes in best practice that affect the immediate experimental background, such as changes in analgesia regimen, post-surgical recovery procedures, group housing, or enrichment. There is an inherent temptation to keep methodology and experimental background identical over time, for fear of some small change leading to the model "not working"; but in reality the experimental background of our animals is changing constantly in ways we cannot control. Furthermore, if a model only works under very specific conditions[30], then it lacks external validity: it doesn't generalize across circumstances even within the same species studied using the same measures, and is therefore unlikely to translate to humans. Such examples are probably not very good models, and it is therefore crucial that we attempt to identify and weed out these cases with low external validity by deliberately running experiments against a variety of experimental backgrounds[26–28].

Often though, changes in models may recapitulate clinically relevant phenomena in humans. For instance, all other things being equal, the major predictors of breast cancer survival in humans are measures of social support—contributing an astonishing 50-fold change in risk when combined[94]. The same is true of spontaneous breast cancer in rats: singly housed rats show an 84–fold increase in tumor burden over those housed in groups, and increased risk for more invasive forms of breast cancer, predicted by elevated anxiety and hypothalamic–pituitary–adrenal (HPA) axis sensitization[95]. Similar effects are seen in mice[96]. Far from being a source of unwanted noise, socially housing rats and mice allows us to study the underlying psychobiology of social support in humans.

Experimental background is ignored because of the one-size-fits-all industrialization of animal facilities. The last 20 years have seen wholesale changes in animal facility infrastructure, for instance from static caging to individually ventilated (IVC); from wood-based bedding to paper bedding to corncob; and from the growing introduction of enrichment. Once a change like investing in IVC caging is made, there's no going back, and all too often we discover that there are unintended consequences. For instance, mice find IVCs and high ventilation rates aversive[97], which results in a general state of heightened fear and anxiety[98] as well as HPA axis sensitization and immune supression[99] compared to mice in conventional caging. IVCs are typically adopted to extend cage-change intervals, yet they are not able to limit ammonia levels sufficiently to avoid nasal lesions in breeding[100] and stock cages[101]. Similarly, recent years have seen a broad adoption of corncob bedding, primarily because this material is easily dispensed by automatic systems. However, not only is corncob also a sandblasting material (which should raise concerns about comfort), but it also contains a number of potent estrogen disruptors. Accordingly, rats housed on corncob show less slow-wave sleep[60], reduced reproductive output, and female acyclicity[102]; prostate-cancer xenografts grow at an accelerated rate in mice housed on corncob[102]; and *Peromyscus californicus* housed on corncob show twofold increases in aggression, on the same scale as those due to treatment with the aromatase inhibitor fadrazole[62]. Similar examples exist for many other aspects of animal housing.

The point here is not that one particular system is the best, but that it might be wise to ask the animals about the impact a change in housing will have before we implement it. Furthermore, there simply is no such thing as a "historical control", either experimentally (one would never exclude controls from an experiment and use data from control animals in the past), or conceptually—animal environments change, and as a result, so do animals.

Experimental background is ignored through seeing animals as tools, not as patients. When we are used to seeing mice in barren standardized environments, with standardized chow, and standardized genetics, referring to them as models not mice or animals, it is easy to fall into the trap of thinking of them more as little furry test tubes than animal patients. But the other examples in this question boil down to this simple change in perspective: if we think of animals as patients, not tools, it forces us to think about all the aspects of the experimental background that differ from humans that we might otherwise ignore.

What animal well-being issues are ignored?

Although animal well-being considerations are spread across the other questions, asking this as a separate question emphasizes the central importance of good well-being to good science.

Animal well-being issues are ignored through ignorance of human well-being effects on health outcomes. In humans we know that many aspects of general well-being impact health outcomes[95]. The animal well-being literature focuses on the underlying biology of these effects to argue that "good wellbeing is good science"

(see this special issue refs. 41–43, and more generally refs. 26,56–58). Regardless, investigators may be unaware of the role well-being plays in human health, and as such be unaware of the potential impacts of animal well-being issues on their animal models. As discussed above, improving well-being often adds to the model, not just because we are normalizing the animal's biology[41], but because we can now model the effects of clinically relevant variables like social support in cancer[96,97], or educational level and cognitive demand, as modeled by enrichment in mice, in preserving function in Alzheimer's disease[103].

Animal well-being issues are ignored through mistaking absence of evidence for evidence of absence. The catch-22 of laboratory animal well-being science is that it is, for all intents and purposes, completely unfunded in the US. As a result, we are constantly responding to negative consequences of cost or engineering-driven changes in practice, rather than, with notable exceptions[58,104], providing evidence that drive changes in practice. This vacuum is often filled by the fallacious argument that absence of evidence is evidence of absence—that because nobody has shown that there is an issue, it means that no issue exists. For instance, there was great resistance to the idea that fish feel pain, in part because it took so long for animal well-being science to pay attention to the issue. In fact, fish have homologous neurophysiology to mammals, and they show complex cognitive responses to pain consistent with a central subjective experience of it[105].

Animal well-being issues are ignored by treating the animal as a tool, not a patient. The examples in this question, again, boil down to the observation that by seeing mice as tools rather than patients, we are more likely to overlook a well-being problem and its impact on an experiment. The worst outcome though, is falling into the trap of thinking not that well-being doesn't matter, but that animals can't feel pain, or can't be fearful, or can't be depressed. This position is best challenged as an issue of validity—if mice can't feel pain like we do, then they are of little use as a means to find new analgesics, an extension of the argument for the need to measure central experience in developing analgesics[48,50]; and if mice can't be depressed then they can't be useful models of depression. If we claim that mice are valid models of pain, depression, and a myriad of other disorders, then we have to recognize that they can experience these states in other experiments and we need to mitigate them for the sake of both the animal and the experiment.

What principles of experimental design and statistics are ignored?

In many ways, experimental design and statistics are the easiest fix in terms of addressing translatability. The differences between animal and human experimental designs are shocking, and go a long way to explain the failure of translation[17,18]. Similarly, basic errors in experimental design and analysis are easy to spot and do predict the likelihood of a result being robust and replicable[7,17,19,106]. This question, once again, challenges us to ask, are we treating animals as tools or patients?

Principles of experimental design and statistics are ignored through ignorance of advanced designs. The incongruity between a highly standardized animal experiment and a human trial that embraces variation is astonishing. What human trial would propose studying the effect of a drug only in 43 year old males who are all twin brothers living in one small town in California, with identical studio apartments, identical educations, identical monotonous jobs, identical furniture, identical monotonous diets, identical locked thermostats set to uncomfortably cold temperatures, where the house is cleaned by a grizzly bear that erases all of their social media every two weeks? But this bizarre "Stepford Experiment" is exactly what we aspire to in an animal study. In human work, not only do we recognize the richness of individual diversity, but we actively study it[17,18,40,107–109]. More advanced experimental designs and analyses that either study individual variation or spontaneous disease within the animal population[54], or which deliberately introduce variability in a controlled manner[26–28] as we do in real human trials, offer several advantages. They allow us to understand variability and thus find biomarkers; they test the generality of the result across different experimental backgrounds and are thus much less prone to false positives; they are a match to human clinical study design further increasing the chance that they will translate; and they are more powerful, reducing sample size sometimes by orders of magnitude[17,26]. Indeed these basic points were made by R.A. Fisher[110], the father of biostatistics, in 1935. For details of such designs, and simulations demonstrating these points, see refs. 17,26–28.

Principles of experimental design and statistics are ignored because we assume everything important is controlled, and those controls have worked. An implicit assumption of the Stepford Experiment is that we have controlled everything we need to control, but this cannot be the case—animals see colors we do not, hear sounds we do not, have electrical and magnetic senses that we do not, respond to odors and pheromones that we can't detect and are fundamentally affected by things we are unaware of. For instance, mice are generally more stressed the higher from the floor they are housed, thus rack position affects abnormal behavior[111], and anxiety and immune function[112], so much so that when NOD mice are housed at the top of a rack, they are sufficiently immunosuppressed that the time to onset and the proportion of animals developing Type I diabetes is significantly affected[112]. Even if we know about such effects we might not be able to control for them—for instance, the identity of the experimenter has a far greater impact on measures of pain in mice than the genetics of the animal[31], due in part to a stress-induced analgesia in mice in response to pheromones produced by male experimenters[67]. It is clearly impossible to have one experimenter perform all of the pain assays in an institution, let alone the whole world, so this confound simply cannot be controlled. And even if this was possible, who should we use? Which row of the rack should we perform our NOD mouse Type I diabetes experiments on? This choice could in theory be informed by the characteristics of the human patient population we wish to model. In reality, though, patient populations are heterogeneous in many important respects, further underscoring the need for experimental designs that explicitly take this variation into account rather than ignoring it.

If we cannot standardize or control major experimental confounds such as these, or if it would be ill-advised to do so, what should we do? Again the answer is simple: adopt the experimental designs and analyses used in humans that are specifically designed to deal with these problems. As most unseen variables cluster at the

cage level, while many experimental treatments may be applied to individual animals within a cage, the simplest such version is to adopt a randomized block design where cage is included as blocking factor in the analyses—this is equivalent to a human study where mice from the same cage are matched-pair controls. For more detailed discussion, see refs. 17,26–28,113.

Principles of experimental design and statistics are ignored because there is not a primary hypothesis and primary outcome measure. In human trials it is considered best practice to register an analysis plan, to specify primary outcome measures, and in general to formally state both a null hypothesis and how it will be tested. There is a clear understanding that the more rigorously defined an analysis is before its performance, the more likely a significant result is to be true, for a variety of reasons[22], not least the avoidance of *p*-hacking[107]. Contrast this with animal trials that often have multiple stages of confirmation, which is not hypothesis testing. Multiple outcome measures may be taken, and those that are significant are believed even when they disagree with other measures of the same property[90]. Perhaps a fishing expedition is dressed up in a pseudo hypothesis: "*knocking out gene X will cause a change in the mouse*" is not a falsifiable hypothesis, because it lacks any specificity—clearly it will cause *some* change or other—but this is the essence of phenotyping. Aside from egregious examples of bias and cherry-picking, these experiments work best when they are also analyzed correctly.

To assign a probability to something unknown, we have to state something else as known as a reference: the logic and math of a p-value rests on the assumption that the null hypothesis is correct (i.e. p = the chance of seeing a result this unusual given that the null hypothesis is correct). Therefore an experiment without a hard null hypothesis can't be analyzed with a *p*-value. Instead in all of the examples above it is far more appropriate to calculate the false discovery rate (FDR)[17].

Principles of experimental design and statistics are ignored because false discovery rate is not considered, especially for unexpected results. The probability of false discovery can be stated as the *q*-value, by taking the positive result as the known entity. Thus, q = given that the observed result is significant, what is the chance that it is a false positive? In the situations described above where the null hypothesis is lacking, or many tests are used to ask the same question, as when many different measures of anxiety are taken in a phenotyping screen, each with multiple variables, and any significant result would be believed, the *q*-value is the correct test statistic[17,94]. This is particularly true for an unexpected result. Imagine a phenotyping screen, with potentially tens or hundreds of readouts. If I perform just five tests, and the null hypothesis is true in each case, the chance that at least one will be significant at $p < 0.05$ by chance alone is 23%. For this reason alone, one can virtually guarantee that when hundreds of tests are performed, an unexpected result is a false discovery[17,22]. This is even more true when we consider all of the examples above where an apparent difference has no broader biological relevance, such as a phenotype caused by a gene linked to a mutation, not the mutation itself. This isn't to say that we should ignore serendipity in science, but that serendipitous discoveries need to be followed up not by confirmatory experiments but by aggressive attempts to prove them untrue.

Principles of experimental design and statistics are ignored when we misinterpret confirmatory experiments and technical replicates as bolstering evidence. Confirmatory experiments are an essential part of science when technical errors could produce a false result—this is particularly true in molecular biology for example. However it is a mistake to approach them as an attempt to confirm an original unexpected result; they should always be an attempt to prove it untrue. Consider a phenotype due to a linked gene: rederiving the mutation on a different background would replicate the original technical error. If we wanted to truly try to invalidate the causal relationship we might derive the mutation by a different means, or use RNAi, or pharmacologically interfere with the gene product in some way. It is not difficult to find examples of researchers suggesting invasive therapies in humans on the basis of a series of confirmatory experiments that did not attempt to disprove the original unexpected result, even when the conclusion is implausible on its face (for example, that some kind of selective silencing of *HOXB8* in bone marrow is the cause of trichotillomania[114], which is simply implausible for a disorder that affects 3–5% of women)[17].

Principles of experimental design and statistics are ignored when randomization, blinding, and other cornerstones of good practice are disregarded. A recent survey of published, peer-reviewed animal work found that 86% did not blind observer to treatment when subjective assessment was involved, and 87% did not randomly allocate animals to treatment[115]. In fact, a lack of blinding, a lack of proper control, lack of measure or method validation, selective reporting of results and other cornerstones of good practice systematically predicted which studies were more or less likely to be replicated when repeated in-house by pharmaceutical companies[7,19]. It is depressing that poor practice is so widespread, but heartening that we do in fact know what to do to produce robust, replicable work[7].

Therioepistemology: a change in perspective

It may seem audacious to propose a new discipline, but we believe marking this moment as such is warranted. New disciplines are born in paradigm shifts, where the prevailing world view can no longer bear the weight of evidence against it, and a shift in perspective is required to conceptualize the evidience[116]. Therioepistemology formalizes two related shifts in thinking that are sufficiently tectonic to count as a paradigm shift. First, the shift from asking "*what have we controlled?*" to asking "*what have we chosen to ignore, and at what cost?*" And second, the shift from viewing research animals as little furry test tubes, to viewing them as animal patients. As we argue throughout, these two changes in perspective are inherently linked.

As we change our mindset, we can begin to more formally assess what we've chosen to ignore, or have disregarded due to ignorance of its importance, in our own work, the work we read, and the work we review. To aid in assessment, we have asked six overarching questions to focus attention on what has been ignored and ways in which the animal patient may be critically unlike the human patient. These are by no means exhaustive and we welcome improvements to this heuristic structure. We hope that we have at least formalized the process enough that therioepistemology

provides a framework to help investigators and reviewers ask *"what type of validity is claimed, what type has been shown, and which is relevant to the research question at hand?", and most importantly "are the validities properly aligned?"* Given the existing evidence surveyed here, we firmly believe that getting this alignment right will radically improve the reproducibility and translatability of animal work, with great benefits to the animals used in research and human health in all its facets.

ACKNOWLEDGMENTS

This paper represents the culmination of years of discussion between the authors and many of our colleagues, of particular note: Doctors Hanno Würbel, Mark Tricklebank, Joy Mench, Charles Clifford, Guy Mulder, Bernard Rollin, Daniel Weary, Jeffrey Alberts, Michael Festing, Karen Parker, Amy Lossie, and Edmond Pajor. This work was supported in part by the National Institute of Neurological Disorders And Stroke of the National Institutes of Health under Award Number R21NS088841. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests

Received 15 January 2017; accepted 17 February 2017
Published online at www.nature.com/laban

1. Kola, I. & Landis, J. Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov.* **3,** 711–716 (2004).
2. Paul, S.M. *et al.* How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discov.* **9,** 203–214 (2010).
3. Scannell, J.W., Blanckley, A., Boldon, H. & Warrington, B. Diagnosing the decline in pharmaceutical R&D efficiency. *Nat. Rev. Drug Discov.* **11,** 191–200 (2012).
4. Hay, M., Thomas, D.W., Craighead, J.L., Economides, C. & Rosenthal, J. Clinical development success rates for investigational drugs. *Nat. Biotechnol.* **32,** 40–51 (2014).
5. Rosenblatt, M. An incentive-based approach for improving data reproducibility. *Sci. Transl. Med.* **8,** 336ed5 (2016).
6. Pusztai, L., Hatzis, C. & Andre, F. Reproducibility of research and preclinical validation: problems and solutions. *Nat. Rev. Clin. Oncol.* **10,** 720–724 (2013).
7. Begley, C.G. Six red flags for suspect work. *Nature* **497,** 433–434 (2013).
8. Prinz, F., Schlange, T. & Asadullah, K. Believe it or not: how much can we rely on published data on potential drug targets? *Nat. Rev. Drug Discov.* **10,** 712 (2011).
9. Mak, I.W., Evaniew, N. & Ghert, M. Lost in translation: animal models and clinical trials in cancer treatment. *Am. J. Transl. Res.* **6,** 114–118 (2014).
10. McManus, R. Ex-director Zerhouni surveys value of NIH research. *NIH Record* **LXV** (2013).
11. Cummings, J., Morstorf, T. & Zhong, K. Alzheimer's disease drug-development pipeline: few candidates, frequent failures. *Alzheimer's Res. Ther.* **6,** 37 (2014).
12. Zahs, K.R. & Ashe, K.H. 'Too much good news' - are Alzheimer mouse models trying to tell us how to prevent, not cure, Alzheimer's disease? *Trends Neurosci.* **33,** 381–389 (2010).
13. Sena, E.S., van der Worp, H.B., Bath, P.M.W., Howells, D.W. & Macleod, M.R. Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. *PLoS Biol.* **8,** e1000344 (2010).
14. van der Worp, H.B. *et al.* Can animal models of disease reliably inform human studies? *PLoS Med.* **7,** e1000245 (2010).
15. Peers, I.S., Ceuppens, P.R. & Harbron, C. In search of preclinical robustness. *Nat. Rev. Drug Discov.* **11,** 733–734 (2012).
16. Macleod, M.R. *et al.* Evidence for the efficacy of NXY-059 in experimental focal cerebral ischaemia is confounded by study quality. *Stroke* **39,** 2824–2829 (2008).
17. Garner, J.P. The significance of meaning: why do over 90% of behavioral neuroscience results fail to translate to humans, and what can we do to fix it? *ILAR J.* **55,** 438–456 (2014).
18. Tricklebank, M.D. & Garner, J.P. in *Drug Discovery for Psychiatric Disorders* Vol. 28. (eds. Z. Rankovic, M. Bingham, E.J. Nestler & R. Hargreaves) 534–556 (The Royal Society of Chemistry, 2012).
19. Begley, C.G. & Ellis, L.M. Drug development: Raise standards for preclinical cancer research. *Nature* **483,** 531–533 (2012).
20. Hunter, J. Challenges for pharmaceutical industry: new partnerships for sustainable human health. *Philos. Trans. A. Math. Phys. Eng. Sci.* **369,** 1817–1825 (2011).
21. Geerts, H. Of mice and men: bridging the translational disconnect in CNS drug discovery. *CNS Drugs* **23,** 915–926 (2009).
22. Ioannidis, J.P. Why most published research findings are false. *PLoS Med.* **2,** e124 (2005).
23. Würbel, H. Behavioral phenotyping enhanced - beyond (environmental) standardization. *Genes Brain Behav.* **1,** 3–8 (2002).
24. Würbel, H. Behaviour and the standardization fallacy. *Nat. Genet.* **26,** 263 (2000).
25. Würbel, H. Ideal homes? Housing effects on rodent brain and behaviour. *Trends Neurosci.* **24,** 207–211 (2001).
26. Würbel, H. & Garner, J.P. Refinement of rodent research though environmental enrichment and systematic randomization. *NC3Rs* **9,** 1–9 (2007).
27. Richter, S.H., Garner, J.P., Auer, C., Kunert, J. & Würbel, H. Systematic variation improves reproducibility of animal experiments. *Nat. Methods* **7,** 167–168 (2010).
28. Richter, S.H., Garner, J.P. & Würbel, H. Environmental standardization: cure or cause of poor reproducibility in animal experiments? *Nat. Methods* **6,** 257–261 (2009).
29. Andreatini, R. & Bacellar, L.F.S. Animal models: Trait or state measure? The test-retest reliability of the elevated plus-maze and behavioral despair. *Prog. Neuropsychopharmacol. Biol. Psychiatry* **24,** 549–560 (2000).
30. Andrews, N. & File, S.E. Handling history of rats modified behavioural effects of drugs in the elevated plus-maze test of anxiety. *Eur. J. Pharmacol.* **235,** 109–112 (1993).
31. Chesler, E.J., Wilson, S.G., Lariviere, W.R., Rodriguez-Zas, S.L. & Mogil, J.S. Identification and ranking of genetic and laboratory environment factors influencing a behavioral trait, thermal nociception, via computational analysis of a large data archive. *Neurosci. Biobehav. Rev.* **26,** 907–923 (2002).
32. Crusio, W.E. Flanking gene and genetic background problems in genetically manipulated mice. *Biol. Psychiatry* **56,** 381–385 (2004).
33. Crusio, W.E., Goldowitz, D., Holmes, A. & Wolfer, D. Standards for the publication of mouse mutant studies. *Genes Brain Behav.* **8,** 1–4 (2009).
34. Wolfer, D.P., Crusio, W.E. & Lipp, H.P. Knockout mice: simple solutions to the problems of genetic background and flanking genes. *Trends Neurosci.* **25,** 336–340 (2002).
35. Nieuwenhuis, S., Forstmann, B.U. & Wagenmakers, E.-J. Erroneous analyses of interactions in neuroscience: a problem of significance. *Nat. Neurosci.* **14,** 1105–1107 (2011).
36. Brown, R.E. & Wong, A.A. The influence of visual ability on learning and memory performance in 13 strains of mice. *Learn. Mem.* **14,** 134–144 (2007).
37. Gerlai, R. Gene-targeting studies of mammalian behavior - is it the mutation or the background genotype. *Trends Neurosci.* **19,** 177–181 (1996).
38. Gerlai, R. & Clayton, N.S. Analysing hippocampal function in transgenic mice: An ethological perspective. *Trends Neurosci.* **22,** 47–51 (1999).
39. Garner, J.P. Stereotypies and other abnormal repetitive behaviors: potential impact on validity, reliability, and replicability of scientific outcomes. *ILAR J.* **46,** 106–117 (2005).
40. Insel, T.R. From animal models to model animals. *Biol. Psychiatry* **62,** 1337–1339 (2007).
41. Gaskill, B.N. & Garner, J.P. Stress out: providing laboratory animals with behavioral control to reduce the physiological impacts of stress. *Lab Anim. (NY)* **46,** 142–145 (2017).
42. Jirkof, P. Side effects of pain and analgesia in animal experimentation. *Lab Anim. (NY)* **46,** 123–128 (2017).
43. Prescott, M.J. & Lidster, K. Improving quality of science through better animal welfare: the NC3Rs strategy. *Lab Anim. (NY)* **46,** 152–156 (2017).
44. Martin, P. & Bateson, P. *Measuring Behaviour: An Introductory Guide* (Cambridge University Press, Cambridge, England UK, 1986).

45. Arguello, P.A. & Gogos, J.A. Modeling madness in mice: one piece at a time. *Neuron* **52,** 179–196 (2006).
46. Campbell, D.T. & Fiske, D.W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.* **56,** 81–105 (1959).
47. Willner, P. Validation criteria for animal models of human mental disorders: Learned helplessness as a paradigm case. *Prog. Neuropsychopharmacol. Biol. Psychiatry* **10,** 677–690 (1986).
48. Flecknell, P.A. Do mice have a pain face? *Nat. Methods* **7,** 437–438 (2010).
49. Vierck, C.J., Hansson, P.T. & Yezierski, R.P. Clinical and pre-clinical pain assessment: are we measuring the same thing? *Pain* **135,** 7–10 (2008).
50. Sufka, K.J. Translational challenges and analgesic screening assays. *Pain* **152,** 1942–1943 (2011).
51. Fu, L., Pelicano, H., Liu, J., Huang, P. & Lee, C.C. The circadian gene period2 plays an important role in tumor suppression and DNA damage response *in vivo*. *Cell* **111,** 41–50 (2002).
52. Nakamura, Y. *et al.* Phospholipase Cdelta1 is required for skin stem cell lineage commitment. *EMBO J.* **22,** 2981–2991 (2003).
53. Rosbash, M. & Takahashi, J.S. Circadian rhythms: The cancer connection. *Nature* **420,** 373–374 (2002).
54. Garner, J.P. *et al.* Reverse-translational biomarker validation of abnormal repetitive behaviors in mice: an illustration of the 4Ps modeling approach. *Behav. Brain Res.* **219,** 189–196 (2011).
55. Sclafani, V. *et al.* Early predictors of impaired social functioning in male rhesus macaques (*Macaca mulatta*). *PLoS ONE* **11,** e0165401 (2016).
56. Pham, T.M. *et al.* Housing environment influences the need for pain relief during post-operative recovery in mice. *Physiol. Behav.* **99,** 663–668 (2010).
57. Van Loo, P.L. *et al.* Impact of 'living apart together' on postoperative recovery of mice compared with social and individual housing. *Lab. Anim.* **41,** 441–455 (2007).
58. Hurst, J.L. & West, R.S. Taming anxiety in laboratory mice. *Nat. Methods* **7,** 825–826 (2010).
59. Beura, L.K. *et al.* Normalizing the environment recapitulates adult human immune traits in laboratory mice. *Nature* **532,** 512–516 (2016).
60. Leys, L.J., McGaraughty, S. & Radek, R.J. Rats housed on corncob bedding show less slow-wave sleep. *J. Am. Assoc. Lab. Anim. Sci.* **51,** 764–768 (2012).
61. Mayeux, P., Dupepe, L., Dunn, K., Balsamo, J. & Domer, J. Massive fungal contamination in animal care facilities traced to bedding supply. *Appl. Environ. Microbiol.* **61,** 2297–2301 (1995).
62. Villalon Landeros, R. *et al.* Corncob bedding alters the effects of estrogens on aggressive behavior and reduces estrogen receptor-alpha expression in the brain. *Endocrinology* **153,** 949–953 (2012).
63. Howerton, C.L., Garner, J.P. & Mench, J.A. Effects of a running wheel-igloo enrichment on aggression, hierarchy linearity, and stereotypy in group-housed male CD-1 (ICR) mice. *Appl. Anim. Behav. Sci.* **115,** 90–103 (2008).
64. Gaskill, B.N. *et al.* Energy reallocation to breeding performance through improved nest building in laboratory mice. *PLoS ONE* **8,** e74153 (2013).
65. Gaskill, B.N. *et al.* Heat or insulation: behavioral titration of mouse preference for warmth or access to a nest. *PLoS ONE* **7,** e32799 (2012).
66. Van Loo, P.L.P., Kruitwagen, C.L.J.J., Van Zutphen, L.F.M., Koolhaas, J.M. & Baumans, V. Modulation of aggression in male mice: influence of cage cleaning regime and scent marks. *Anim. Welf.* **9,** 281–295 (2000).
67. Sorge, R.E. *et al.* Olfactory exposure to males, including men, causes stress and related analgesia in rodents. *Nat. Methods* **11,** 629–632 (2014).
68. Labelle, P. *et al.* Mousepox detected in a research facility: case report and failure of mouse antibody production testing to identify Ectromelia virus in contaminated mouse serum. *Comp. Med.* **59,** 180–186 (2009).
69. Kilkenny, C., Browne, W.J., Cuthill, I.C., Emerson, M. & Altman, D.G. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLoS Biol.* **8,** e1000412 (2010).
70. Vaidya, D., Morley, G.E., Samie, F.H. & Jalife, J. Reentry and fibrillation in the mouse heart: a challenge to the critical mass hypothesis. *Circ. Res.* **85,** 174–181 (1999).
71. Wakimoto, H. *et al.* Induction of atrial tachycardia and fibrillation in the mouse heart. *Cardiovasc. Res.* **50,** 463–473 (2001).
72. Moberg, G.P. in *The Biology of Animal Stress: Basic Principles and Implications for Animal Welfare* (eds. G.P. Moberg & J.A. Mench) 1–22 (CABI, Wallingford, UK, 2000).
73. Dominguez, A.A., Lim, W.A. & Qi, L.S. Beyond editing: repurposing CRISPR-Cas9 for precision genome regulation and interrogation. *Nat. Rev. Mol. Cell Biol.* **17,** 5–15 (2016).
74. Sanna, C.R., Li, W.-H. & Zhang, L. Overlapping genes in the human and mouse genomes. *BMC Genomics* **9,** 169 (2008).
75. Graham, M.L. *et al.* Successful implementation of cooperative handling eliminates the need for restraint in a complex non-human primate disease model. *J. Med. Primatol.* **41,** 89–106 (2012).
76. Lapin, B.A., Gvozdik, T.E. & Klots, I.N. Blood glucose levels in rhesus monkeys (*Macaca mulatta*) and cynomolgus macaques (*Macaca fascicularis*) under moderate stress and after recovery. *Bull. Exp. Biol. Med.* **154,** 497–500 (2013).
77. Graham, M.L. & Schuurman, H.-J. Validity of animal models of type 1 diabetes, and strategies to enhance their utility in translational research. *Eur. J. Pharmacol.* **759,** 221–230 (2015).
78. Harding, E.J., Paul, E.S. & Mendl, M. Animal behavior: cognitive bias and affective state. *Nature* **427,** 312 (2004).
79. Paul, E.S., Harding, E.J. & Mendl, M. Measuring emotional processes in animals: the utility of a cognitive approach. *Neurosci. Biobehav. Rev.* **29,** 469 (2005).
80. Garner, J.P., Meehan, C.L. & Mench, J.A. Stereotypies in caged parrots, schizophrenia and autism: evidence for a common mechanism. *Behav. Brain Res.* **145,** 125–134 (2003).
81. Garner, J.P. & Mason, G.J. Evidence for a relationship between cage stereotypies and behavioural disinhibition in laboratory rodents. *Behav. Brain Res.* **136,** 83–92 (2002).
82. Gould, T.D. & Gottesman, I.I. Psychiatric endophenotypes and the development of valid animal models. *Genes Brain Behav.* **5,** 113–119 (2006).
83. Garner, J.P., Thogerson, C.M., Würbel, H., Murray, J.D. & Mench, J.A. Animal neuropsychology: validation of the intra-dimensional extra-dimensional set shifting task in mice. *Behav. Brain Res.* **173,** 53–61 (2006).
84. Abelson, J.F. *et al.* Sequence variants in SLITRK1 are associated with Tourette's syndrome. *Science* **310,** 317–320 (2005).
85. Zuchner, S. *et al.* SLITRK1 mutations in Trichotillomania. *Mol. Psychiatry* **11,** 888–889 (2006).
86. Shmelkov, S.V. *et al.* Slitrk5 deficiency impairs corticostriatal circuitry and leads to obsessive-compulsive-like behaviors in mice. *Nat. Med.* **16,** 598–602 (2010).
87. George, N.M. *et al.* Antioxidant therapies for ulcerative dermatitis: a potential model for skin picking disorder. *PLoS ONE* **10,** e0132092 (2015).
88. Wahlsten, D. Standardizing tests of mouse behavior: reasons, recommendations, and reality. *Physiol. Behav.* **73,** 695–704 (2001).
89. Binder, E., Droste, S.K., Ohl, F. & Reul, J.M.H.M. Regular voluntary exercise reduces anxiety-related behaviour and impulsiveness in mice. *Behav. Brain Res.* **155,** 197–206 (2004).
90. Miller, K.A., Garner, J.P. & Mench, J.A. Is fearfulness a trait that can be measured with behavioural tests? A validation of four fear tests for Japanese quail. *Anim. Behav.* **71,** 1323–1334 (2006).
91. Holmes, P.V. Rodent models of depression: reexamining validity without anthropomorphic inference. *Crit. Rev. Neurobiol.* **15,** 143–174 (2003).
92. Langford, D.J. *et al.* Coding of facial expressions of pain in the laboratory mouse. *Nat. Methods* **7,** 447–449 (2010).
93. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29,** 1165–1168 (2001).
94. Kroenke, C.H., Kubzansky, L.D., Schernhammer, E.S., Holmes, M.D. & Kawachi, I. Social networks, social support, and survival after breast cancer diagnosis. *J. Clin. Oncol.* **24,** 1105–1111 (2006).
95. Hermes, G.L. *et al.* Social isolation dysregulates endocrine and behavioral stress while increasing malignant burden of spontaneous mammary tumors. *Proc. Natl. Acad. Sci. USA* **106,** 22393–22398 (2009).
96. Kerr, L.R., Grimm, M.S., Silva, W.A., Weinberg, J. & Emerman, J.T. Effects of social housing condition on the response of the Shionogi mouse mammary carcinoma (SC115) to chemotherapy. *Cancer Res.* **57,** 1124–1128 (1997).
97. Baumans, V., Schlingmann, F., Vonck, M. & Van Lith, H.A. Individually ventilated cages: beneficial for mice and men? *Contemp. Top. Lab. Anim. Sci.* **41,** 13–19 (2002).

98. Kallnik, M. et al. Impact of IVC housing on emotionality and fear learning in male C3HeB/FeJ and C57BL/6J mice. *Mamm. Genome* **18,** 173–186 (2007).
99. Neigh, G.N., Bowers, S.L., Korman, B. & Nelson, R.J. Housing environment alters delayed-type hypersensitivity and corticosterone concentrations of individually housed male C57BL/6 mice. *Anim. Welf.* **14,** 249–257 (2005).
100. DiVincenti, L., Moorman-White, D., Bavlov, N., Garner, M. & Wyatt, J. Effects of housing density on nasal pathology of breeding mice housed in individually ventilated cages. *Lab Anim. (NY)* **41,** 68–76 (2012).
101. Nagamine, C. *et al.* Ammonia production and nasal histopathology in mice housed in 4 IVC systems for 14, 21, or 28 days. *J. Am. Assoc. Lab. Anim. Sci.* **53,** 566 (2014).
102. Markaverich, B. *et al.* A novel endocrine-disrupting agent in corn with mitogenic activity in human breast and prostatic cancer cells. *Environ. Health Perspect.* **110,** 169–177 (2002).
103. Jankowsky, J.L. *et al.* Environmental enrichment mitigates cognitive deficits in a mouse model of Alzheimer's disease. *J. Neurosci.* **25,** 5217–5224 (2005).
104. Adams, S.C., Garner, J.P., Felt, S.A., Geronimo, J.T. & Chu, D.K. A "Pedi" cures all: toenail trimming and the treatment of ulcerative dermatitis in mice. *PLoS ONE* **11,** e0144871 (2016).
105. Nordgreen, J. *et al.* in *Proceedings of the 42nd International Congress of the International Society for Applied Ethology 13* (Dublin, UK, 2008).
106. Shanks, D.R. *et al.* Romance, risk, and replication: Can consumer choices and risk-taking be primed by mating motives? *J. Exp. Psychol. Gen.* **144,** e142–e158 (2015).
107. Parker, K.J. *et al.* Plasma oxytocin concentrations and OXTR polymorphisms predict social impairments in children with and without autism spectrum disorder. *Proc. Natl. Acad. Sci. USA* **111,** 12258–12263 (2014).
108. Yuen, K.W. *et al.* Plasma oxytocin concentrations are lower in depressed vs. healthy control women and are independent of cortisol. *J. Psychiatr. Res.* **51,** 30–36 (2014).
109. Gottesman, I.I. & Gould, T.D. The endophenotype concept in psychiatry: Etymology and strategic intentions. *Am. J. Psychiatry* **160,** 636–645 (2003).
110. Fisher, R.A. *The Design of Experiments* (Oliver and Boyde, Edinburgh, London, 1935).
111. Garner, J.P., Dufour, B., Gregg, L.E., Weisker, S.M. & Mench, J.A. Social and husbandry factors affecting the prevalence and severity of barbering ('whisker trimming') by laboratory mice. *Appl. Anim. Behav. Sci.* **89,** 263–282 (2004).
112. Ader, D.N., Johnson, S.B., Huang, S.W. & Riley, W.J. Group-size, cage shelf level, and emotionality in nonobese diabetic mice: impact on onset and incidence of IDDM. *Psychosom. Med.* **53,** 313–321 (1991).
113. Walker, M. *et al.* Mixed-strain housing for female C57BL/6, DBA/2, and BALB/c mice: validating a split-plot design that promotes refinement and reduction. *BMC Med. Res. Methodol.* **16,** 11 (2016).
114. Chen, S.K. *et al.* Hematopoietic origin of pathological grooming in Hoxb8 mutant mice. *Cell* **141,** 775–785 (2010).
115. Kilkenny, C. *et al.* Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PLoS ONE* **4,** e7824 (2009).
116. Kuhn, T.S. & Hacking,, I. *The Structure of Scientific Revolutions* 4th edn. (The University of Chicago Press, Chicago; London, 2012).