Review Article

Alternatives to Laboratory Animals
2020, Vol. 48(3) 106–115

DOI: 10.1177/0261192920939876
journals.sagepub.com/home/atl


Some Reasons Why Preclinical Studies of Psychiatric Disorders Fail to Translate: What Can Be Rescued from the Misunderstanding and Misuse of Animal 'Models'?

S. Clare Stanford

Abstract

The repeated failure of animal models to yield findings that translate into humans is a serious threat to the credibility of preclinical biomedical research. The use of animals in research that lacks translational validity is unacceptable in any ethical environment, and so this problem needs urgent attention. To reproduce any human illness in animals is a serious challenge, but this is especially the case for psychiatric disorders. Yet, many authors do not hesitate to describe their findings as a 'model' of such a disorder. More cautious scientists describe the behavioural phenotype as 'disorder-like', without specifying the way(s) in which the abnormal behaviour could be regarded as being analogous to any of the diagnostic features of the disorder in question. By way of discussing these problems, this article focuses on common, but flawed, assumptions that pervade preclinical research of depression and antidepressants. Particular attention is given to the difference between putative 'models' of this illness and predictive screens for candidate drug treatments, which is evidently widely misunderstood. However, the problems highlighted in this article are generic and afflict research of all psychiatric disorders. This dire situation will be resolved only when funders and journal editors take action to ensure that researchers interpret their findings in a less ambitious, but more realistic, evidence-based way that would parallel changes in research of the cause(s), diagnosis and treatment of psychiatric problems in humans.

Keywords

animal model, depression and antidepressants, endophenotype, Forced Swim Test, predictive drug screen, psychiatry, psychopharmacology, translation, validity

Introduction

Immense effort is being invested in devising remedies for the poor reproducibility of preclinical biomedical research. It goes without saying that the design of experiments, and the way they are carried out, are crucial for ensuring that the conclusions are valid. But, in terms of helping to ensure successful translation of preclinical research into humans, the campaign to improve reproducibility is merely tinkering around the edges unless the scientific rationale for an experiment is sound. Taking steps to improve the reproducibility of experimental results will not improve translation if the interpretation of the research findings is misleading. A failure to translate is not only a major setback for the field but, when animals have been used at any stage of the process, it is ethically unacceptable if the failure was avoidable.

There have been several recent commentaries on factors that contribute to failed translation. These have covered many important points but have tended to focus on topics such as: species differences; methodology; experimental environment; neglected complications arising from the use of genetically altered animals; subjective and systematic bias; skills deficits; and even a need to redefine the whole strategy for preclinical research.[1–3] However, a

Department of Neuroscience, Physiology and Pharmacology, University College London, London, UK

Corresponding author:
S. Clare Stanford, Department of Neuroscience, Physiology and Pharmacology, University College London, Gower Street, London WC1E 6BT, UK.
Email: c.stanford@ucl.ac.uk

fundamental problem, which has received surprisingly little attention, is that some procedures are being used as 'models' of complex human disorders when that was not their intended purpose and when the underlying assumptions have never been validated.[4]

This article does not offer a detailed critique of the wide range of procedures that are used in research into the causes of and treatments for psychiatric disorders, not least because that would be a task with almost unlimited scope. It is also not intended as a discourse of the important ethical and philosophical debates arising from species differences in neuroanatomy and neuronal networks, which lead some authors to conclude that the development of an animal model for any human brain disorder can never be a realistic objective.[5,6] Instead, the baseline for this article is that not only do many scientists believe that such models exist already, but they are using them with impunity.

This scenario would be acceptable if the assumptions that underpin the model(s) had been validated, and if it is acknowledged that their validity depends on the experimental context and the research objectives. The aim of this article is to highlight: (a) the extent to which these vital steps are typically ignored; and (b) how the rationale for the experiments, as well as the conclusions arising from the work, risk being fundamentally flawed as a consequence. These problems do not rest on the fine details of each 'model' or psychiatric disorder; they are generic — and so this article draws on some striking examples from preclinical research on depression and antidepressants, in order to illustrate some ways in which validity can be compromised. Particular attention is devoted to an appraisal of the Forced Swim Test (FST) because that procedure has attracted a good deal of criticism and is the focus of a campaign to ban its use altogether.

All these 'models' have been used for many years and so it is realistic to assume that many scientists will continue to use them, as before, despite their limitations. However, if the problems discussed below are not resolved, there is a risk that confidence in the scientific merits of all procedures used in preclinical psychopharmacology will wither irrevocably, even though many have made invaluable contributions to the field.

To help prevent that from happening, the final section suggests remedial actions that could help to strengthen the validity of inferences from a programme of research that incorporates the use of a 'model' of a psychiatric disorder or 'disorder-like' behaviour.

Reasons for studying behavioural phenotypes

There are two main reasons for studying the behaviour of laboratory animals in preclinical research of psychiatric disorders. One is to induce behavioural changes in the animals that are analogous to the human disorder of interest: that is, to produce a 'model' of at least one aspect of the illness. The neuronal mechanisms underlying these responses can then be interrogated in more detail, which could point the way to new and better treatments. A problem arising from this type of research is that the behavioural phenotype produced by the procedure is often assumed to be analogous to the human disorder, even though it does not begin to qualify for such a 'diagnosis'.

The other reason for studying such behaviour is to test whether or not a potential new treatment merits further development for a given clinical indication. This involves evaluating the effects of the candidate treatment on the behaviour of an animal during exposure to a standard test procedure, for example, the FST or the Tail Suspension Test. The important point about these procedures is that none requires the baseline behaviour (i.e. the behaviour of untreated animals) to be a model of any aspect of the target illness. All that is needed is for all treatments of the same class, with confirmed therapeutic efficacy in humans, to produce a clear and consistent change in any aspect of animal behaviour. That response then acquires *predictive validity* and can be used to screen new candidates for their potential as effective treatments in humans. However, a common misunderstanding, which is discussed in more detail below, is that the behaviour under evaluation is often assumed to be a model of a psychiatric disorder, or disorder-like behaviour, and that the prevention of this behaviour by a test treatment indicates that it has 'cured' the illness. This is a particularly common problem when these procedures are used to look for changes in the behaviour of animals that have been genetically altered in a way that is thought to be relevant to the psychiatric disorder of interest.

The challenges for animal 'models' of a human psychiatric disorder

As the publication archive confirms, many researchers believe that it is plausible to claim that an experimental intervention has induced an animal 'model' of a human illness when there is clear pathology or a biomarker, which is common to both species (e.g. when the animals have a specific type of cancer or a microbial infection). But, even in these cases, translation into humans and the development of effective treatments have not been as straightforward as expected. It is far more difficult to be confident about the validity of an animal model when the human disorder comprises several different contributory factors (e.g. hypertension or metabolic syndrome). The development of animal models of psychiatric disorders is particularly challenging because no defining pathology or biomarkers for any of these disorders have been discovered so far.

Several different systems are used to steer the diagnoses of psychiatric illnesses.[7,8] Although the details of the criteria that qualify for a formal diagnosis differ slightly from one to another, the qualitative aspects of symptoms and

Table 1. The core (qualitative) aspects of the symptoms and signs that contribute to a diagnosis of depression in humans and should also be expressed in animal 'models' of depression.

Diagnostic feature	DSM-5	ICD-10
Psychological		
Sadness/depressed mood	✓	✓
Diminished interest or pleasure in activities	✓	✓
Cognitive impairment/poor concentration/indecisiveness	✓	✓
Suicidal thoughts	✓	✓
Feelings of worthlessness or (inappropriate) guilt	✓	
Guilt or self-blame		✓
Low self-confidence		✓
Somatic/vegetative		
Weight loss or gain	✓	
Disturbed sleep (insomnia/hypersomnia)	✓	✓
Fatigue/loss of energy	✓	✓
Reduced or increased appetite	✓	✓
Behavioural		
Agitation with reduction in physical movement	✓	✓

DSM-5: Diagnostic and Statistical Manual of Mental Disorders[7]; ICD-10: World Health Organisation International Classification of Diseases.[8]

signs that are regarded as common in depressed humans are listed in Table 1.

Some of these features are self-evidently analogous in humans and other animals, and their evaluation is fairly straightforward. This would be the case for monitoring changes in body weight, disruption of sleep architecture or alterations to feeding patterns, for instance. Similarly, cognitive impairments that are arguably equivalent to those expressed by humans can also be evaluated in animals, by using tests to monitor their spatial memory, discrimination of novel aspects of their environment, focused attention and so on. By contrast, the equivalence of other abnormalities expressed in animals and certain aspects of depression is less clear-cut.

For instance, the loss of a rodent's innate preference for sweet fluids, after experiencing a prolonged series of mild stressors, is widely regarded as a sign of 'anhedonia', which is a prominent feature of depression. However, it should be borne in mind that the detailed DSM-5 criterion for anhedonia in depression (one of the criteria outlined in the Diagnostic and Statistical Manual of Mental Disorders (DSM-5)) is "a markedly diminished interest or pleasure in many, or all, activities, nearly every day" — that is, the anhedonia is not confined to a gustatory preference. Similarly, a rodent's submissive behaviour after social defeat is often described as 'depression-like' behaviour, but it is not clear what aspect of depression is being emulated. Low self-confidence is a remote possibility, even though, in depressed humans, that symptom is not normally attributed to threats of physical attack from a dominant conspecific.

In neither of these cases, is there any evidence to indicate an animal's state of mind and so the description of the behavioural abnormality as being 'depression-like' is based on risky anthropomorphic assumptions. That limitation taps into the major obstacle for modelling any psychiatric disorder, which is that many of the key diagnostic criteria are subjective and cannot be evaluated in animals at all (e.g. suicidal ideation in depression, flashbacks and survivor guilt in post-traumatic stress disorder (PTSD), hallucinations in schizophrenia).

To add to these problems, a diagnosis of a psychiatric disorder rests on patients expressing combinations of symptoms and signs, which are long-lasting (often lifelong). Some are given more weight than others, and their number and type can differ substantially from one patient to another and yet still qualify for a diagnosis of depression. Moreover, the severity of the illness can wax and wane, with periods of remission that have unpredictable duration. Another confounder is that some symptoms and signs are evident in more than one psychiatric disorder. Given that many disorders are often comorbid and that some of the features of comorbid illnesses overlap, it can be difficult, if not impossible, to attribute any particular behavioural abnormality to any single disorder alone.

It is reasonable to expect the criteria for a valid model of depression in animals to be equivalent to those required for a diagnosis of depression in humans. However, DSM-5, for instance, requires patients to express at least five of nine symptoms for at least 2 weeks. Even if the bar for a diagnosis in animals is lowered as much as possible — by ignoring important details of each feature that psychiatrists would look for in humans (e.g. its duration, as in Table 1) — it is clear that no experimental intervention (e.g. drug treatment, genetic alteration, neuronal lesion) has yet produced a collection of abnormalities in animals that would qualify as a 'model' of depression in terms of its symptom profile, let alone its severity and temporal instability.

To wriggle out of that problem, many authors resort to describing the behavioural abnormality they have studied as 'depression-like' instead, although it is never made clear what 'depression-like' actually means (cf. Table 1; see also Garner[1]). Nevertheless, this is a promising approach because current research strategies are moving towards the study of endophenotypes, whereby a specific abnormality can be linked with an underlying biological process, such as a neuronal network or genetic mutation.[9] Even so, if it is to be claimed that an endophenotype is analogous in humans and other animals, its description still needs to be far more specific than merely 'depression-like.'

Some procedures that are used to produce animal 'models' of depression

This section describes some important examples of procedures that are used as preclinical models of depression in

rodents and outlines their strengths, limitations and common pitfalls.

Reserpine

This was the first animal 'model' of depression. Reserpine is a naturally occurring alkaloid that was used to treat hypertension and mania until around the early 1960s. At that time, it was asserted that this drug caused suicidal depression in a few patients (cf. 12–15%), but that now seems unlikely.[10] Nevertheless, that belief provided a rationale for studying the effects of reserpine in animals on the grounds that it would induce depression in other species too. This assumption was further encouraged by evidence that a new class of antidepressant drugs, which had just been discovered (the 'tricyclics'), prevented reserpine-induced hypothermia.[11] The finding that the tricyclics had negligible effect on the behavioural response to reserpine treatment (akinesia) was basically ignored.

As well as inducing hypothermia and hypotension, reserpine also depletes monoamine-releasing neurons of their pool of neurotransmitter (norepinephrine, serotonin or dopamine). This action seemed to strengthen the endorsement of the model because, at that time, it was thought that depression was caused by a deficit in monoamine transmission. That is no longer the case and, given that none of these abnormalities (hypothermia, hypotension or depletion of brain monoamines) are evident in depressed humans, reserpine treatment is now generally deprecated as a model of depression. Nevertheless, it is still used by some laboratories to study "reserpine-induced depression" [*sic*] and to screen novel compounds for their potential as antidepressants, even though prevention of the hypothermia only ever seemed to happen with drugs that block neuronal reuptake of norepinephrine.[12] Examples can be found on PubMed by using the keywords: reserpine/depression/model.

Learned helplessness

This procedure involves exposing animals to uncontrollable, unpredictable stress (bouts of electric footshocks) in an environment from which they cannot escape. The animals freeze on subsequent exposure to that environment, even when given the opportunity to escape (see Seligman[13]): that is, they develop an 'escape deficit.' When first discovered, this behaviour was somewhat arbitrarily called 'learned helplessness.' That label strongly implied that the stress induces a particular state of mind in the animals, which is analogous to the feeling of helplessness experienced by depressed humans. That implication was supported to the extent that the animals also show impaired grooming, loss of appetite and cognitive deficits, which are arguably analogous to features of depression in humans (cf. Table 1). The finding that all these changes are prevented by treatment with antidepressants[14] further strengthened that possibility.

Nevertheless, the question of whether or not learned helplessness is analogous to depression in humans has been debated continually over the last 50 years. It has even been suggested that it is a model of PTSD, rather than depression.[15] The prevailing view is that the escape deficit arises from an innate, default passivity, rather than 'depression', which is triggered by a lack of control over environmental stressors. A further suggestion is that finding ways of preventing this response (behavioural passivity) could hold the key to antidepression.[16]

The debate over learned helplessness illustrates how it can be really difficult to be confident about the human state that is being emulated by an abnormal behaviour in a given model. That setback, together with concerns about the severity of this procedure, means that it is ethically controversial, which probably explains why it is rarely used in jurisdictions with a strong emphasis on the Three Rs and welfare of laboratory animals.

Olfactory bulbectomy

This is the best validated model of depression and involves bilateral surgical removal of the olfactory bulbs in rodents. Notwithstanding species differences in the effects of bulbectomy on an animal's behaviour and physiology,[17] this procedure causes a wide range of abnormalities in the animals, many of which echo what happens in humans suffering from depression. These involve not only changes in neuronal networks that are consistent with our understanding of the neurobiology of depression,[18,19] but also disruption of the immune system (cytokine production) and endocrine system (raised plasma cortisol), in ways that are common in depressed patients. A strong piece of evidence that distinguishes olfactory bulbectomy from other models is that many of these changes are prevented by prolonged, but not short-term, treatment with an antidepressant drug[20] (cf. the FST, below).

This diverse evidence offers a good example of how validation of a putative animal model involves confirmation that a wide range of abnormalities seen in depressed humans are reproduced in animals too; these include neuroendocrine, immune, therapeutic profile and timescale, not merely a change in a single aspect of their behaviour. By contrast, most rodent 'models' of depression rely merely on comparisons of the change in their motor behaviour after an experimental challenge (e.g. when suspended upside down or immersed in a bucket of water) and do not consider the other physiological changes that are often associated with the human disorder.

Despite the strengths of this model, it does have some limitations, which are widely acknowledged. One is that olfactory bulbectomised animals are hyperactive, which is the opposite of what typically happens in depressed

patients. Another is that depression in humans clearly does not involve physical ablation or section of the olfactory bulbs. However, there are reports that the sense of smell is impaired in some depressed patients,[21] which raises the fascinating possibility that a functional deficit in this brain region could explain this subgroup of patients. Finally, the surgical intervention needed for olfactory bulbectomy is regarded as a severe procedure under the European legislation and so its use needs stringent ethical justification.

Chronic mild (unpredictable) stress

As mentioned above, rodents normally have a strong preference for sweet drinks, but this is diminished after experience of a series of mild stressors. The stressors do not need to be noxious or cause overt physical discomfort but can merely involve disrupting an animal's environment or daily routine. When used as intended, when it was first developed in the UK, the procedures apply stressors such as: short-term food or water restriction; continuous lighting; cage tilt; a change to group-housing; wet bedding; low temperature (10°C); intermittent white noise (85 dB); a novel object in the cage; or strange odour (e.g. air freshener). The animals experience one of these stressors each day, for several weeks, after which they develop a deficit in the Sucrose Preference Test.[22]

Because the loss of sucrose preference after chronic mild stress (CMS) is prevented by antidepressants, it is widely assumed that this change, alone, justifies the use of CMS to produce a model of depression or depression-like behaviour. A more realistic and specific description would be that this deficit is equivalent to a loss of the rodents' motivation to experience pleasure, which is arguably relevant to 'anhedonia' in depressed humans (but see above). Whether or not this is the case, it should be borne in mind that anhedonia is a common element of other human psychiatric disorders too (e.g. schizophrenia, bipolar disorder, obesity and PTSD).

A particular concern about the CMS/chronic mild (unpredictable) stress (CUMS) procedure is the variability in the types of stressors used in different laboratories. In some cases, the protocol involves social isolation in combination with stressors such as: 24-hour food deprivation and/or 24-hour water deprivation; forced swimming in ice-cold water (4–6°C); placement in an oven (45°C); several hours of physical restraint; or several hours of continuous loud white noise. During the procedure, the animals can experience at least one or two of these stressors, every day, for several weeks. Recent examples of these experiments can be found in PubMed by using the keywords: stress/chronic/mild/sucrose.

It is hard to understand how any of these types of stressors could be regarded as 'mild', still less so when the cumulative harm to the animals is taken into account. As well as the ethical considerations arising from these studies, it is important to bear in mind that neuronal responses to stress are strongly dependent on the type, duration, frequency and severity of the stress (reviewed by Stanford[23]). It cannot be assumed that findings that emerge from one protocol are typical, or that studies following different protocols can be compared in any meaningful way. It is also worth considering how these types of stress could be relevant to those that can trigger depression in humans, which often involve loss of control (e.g. bereavement or redundancy), as opposed to PTSD or other long-term consequences of experiencing physically traumatic, life-threatening stress.

Social defeat stress

There are several different types of tests in this category, but they all evaluate changes in animal behaviour after a social challenge between dominant and subordinate animals. Following several bouts of social defeat, the subordinate (defeated) animal typically develops a loss of sucrose preference, a deficit in grooming, reduced body weight and disruption of sleep architecture (Table 1).[24,25] On that basis, this is widely described as a model of 'depression-like' behaviour. However, that inference is somewhat undermined by the gamut of research that has used this procedure to study 'anxiety.'[26,27]

As with CUMS procedures, a major concern about these studies is that there are appreciable differences in the protocols used in different laboratories, particularly in respect of the time for which the animals are left to interact. In some laboratories, as is the case in the UK, the animals are separated immediately after the first contact but retain sight and smell of each other.[28] Elsewhere, the physical interaction between the animals is scored over periods that can last up to 30 minutes each day. The different versions of these tests have implications not only for the interpretation of the findings but also for Three Rs compliance, animal welfare and what qualifies as an ethically acceptable procedure in different jurisdictions.

Predictive drug screens for antidepressants are not models of depression

A second cluster of procedures comprises those that are used as high-throughput, predictive screens to test novel compounds at an early stage of their development. To achieve the status of a predictive screen, all drugs of a given class must affect any aspect of an animal's behaviour in a specific way. For instance, most drugs that bind to 5-HT_{2A} receptors induce the characteristic behavioural 'serotonin syndrome' in rodents — which includes head-twitches, 'wet-dog shakes' and Straub tail — induce hallucinations in humans. Although none of these behaviours are seen in humans (especially Straub tail!), there is a well-justified expectation that if a test drug induces this cluster of abnormal behaviours in rodents, there is a high risk that it will induce hallucinations in humans (see Fantegrossi et al.[29]).

Predictive screens for antidepressants have less pharmacological precision because these drugs interact with and affect many different types of molecular targets in the brain. These range from interactions with neurotransmitter receptors and transporters, to effects on genes that influence glial cell connectivity. Nevertheless, as discussed below, the behavioural response to certain experimental procedures is modified in a consistent way by all antidepressants. In such cases, the effect of a test drug on an animal's behavioural response can be an invaluable guide as to whether or not it is likely to act as an antidepressant in humans.

Some of the procedures discussed above, as 'models' of depression, are also used for the purpose of drug screening (e.g. CUMS and social defeat). Others that are similarly used include the Tail Suspension Test (for mice) and various measures of impaired semantic memory (e.g. novel object or novel location recognition). However, one test that is widely used for this purpose is the FST. This is discussed in detail below, because it is an excellent example of how the validity of an experimental procedure can depend on the experimental objectives — that is, the FST can be valid for one purpose (as a predictive drug screen) but not another (as a model of depression).

The Forced Swim Test (FST)

This procedure involves immersing rodents in a tank of water from which they cannot escape. After a few minutes, the animals stop swimming and adopt a posture ('immobility'), which enables them to float with their noses held above the surface of the water. All established antidepressant drugs increase the latency to adopt this immobile posture[30] and this test has been used as a predictive screen for antidepressants for over 40 years. It is striking that even S(+)-ketamine, which is a stereoisomer of the anaesthetic, ketamine, reduces immobility in the FST[31] and was given FDA approval, in 2019, as a lead compound for a completely new class of fast-acting antidepressants.

One criticism of the FST points to evidence that many factors can affect immobility when the animals are tested in a drug-free state. It is inferred that the FST must be intrinsically unreliable because the variables that affect an animal's baseline behaviour in this test will vary in different laboratories.[32,33] However, the fact that all established antidepressants reduced immobility in this test, despite this baseline variability, actually strengthens, rather than undermines, its value as a high-throughput predictive screen for therapeutic efficacy in humans: that is, the antidepressant 'signal' still shows up above the 'baseline noise.'

Obviously, we cannot know how many compounds that would have made effective antidepressants returned a false-negative result in the FST, and so were not developed for the clinic. On the flip-side, the possibility that there are true false positives (i.e. drugs that reduce immobility in the test but do not translate to effective antidepressants in the clinic) cannot be ruled out either, but strident criticisms of the FST have focused on specific examples. These need to be addressed because there are many reasons why a compound that looked promising in the FST might not reach the clinic. A verdict of a false-positive result in the FST might not be as clear-cut as is often claimed. Some straightforward issues that could be associated with an outcome that is apparently a false positive are outlined below:

- In the test for efficacy in Phase III clinical trials, the estimation of the optimal dose in humans, which is informed by anisotropic estimates from animal studies, might have been incorrect: that is, a different dose might have turned out to be efficacious.[34]
- The drug might not have met safety requirements, or it had undesirable side-effects: that is, its development stumbled after Phase I or Phase II, and efficacy was never tested in the clinic.
- The drug simply turned out to be highly effective at treating a different disorder with a more promising market niche. For example, sibutramine was marketed as an anti-obesity agent instead of an antidepressant, and amoxapine was renamed atomoxetine and is now used to treat attention deficit hyperactivity disorder (ADHD).

A specific group of compounds that has been highlighted as producing false-positive results in the FST include antihistamines. This is a particularly interesting example because many antidepressant drugs are ligands for histamine receptors. In fact, an entire class of antidepressants, known as 'tricyclics', are molecular derivatives of the prototypical antihistamine, promazine, and are antagonists of H_1 receptors. It is possible that antihistamines would make useful antidepressants, were it not for the problem that H_1 receptor antagonists are highly sedative if they reach the brain, as anyone who suffers from hay fever will testify. Indeed, drugs that selectively target other types of histamine receptors are currently being investigated for their potential as antidepressants.

Amphetamine and caffeine are often claimed to be convincing examples of drugs that produce false positives in the FST. Apart from the problem that amphetamine is highly addictive and so has restricted licensed approval (for the treatment of ADHD), that criticism disregards the fact that it was the first drug treatment for depression and was used for over 30 years, until the 1960s, when the first antidepressants were discovered. Amphetamine was regarded as being especially beneficial for patients expressing profound fatigue.[35] Recent studies, based on meta-analysis, continue to support the view that amphetamine can have beneficial effects in the treatment of depression.[36] Similarly, there is mounting evidence that caffeine[37] could make a useful adjunct to established antidepressant drug

regimens.[38] In short, recent evidence shows that neither of these compounds should be regarded as being false positives.

A final reason why some compounds that reduce immobility in the FST have not translated for use as antidepressants in the clinic is that the observed reduction in animal immobility in the FST could actually be attributed to a non-specific increase in their motor activity, rather than a reduction in swim-stress induced immobility (which is completely different), but no experiments were carried out to check that possibility (see below Open Field Test (OFT)). Examples of substances that could fall into this category include: green tea;[39] ghrelin;[40] probiotic supplements;[41] and NK_1 receptor antagonists. Evidence that a lack of functional NK_1 receptors causes locomotor hyperactivity in mice did not come to light until several years after the development of NK_1 receptor antagonists as antidepressants had been abandoned.[42]

However, the main problem with the FST is the widespread belief that, if antidepressants delay the onset of immobility of rodents, then immobility must be a 'model' of depression in humans. There is no evidence to support that assumption, which is an example of what Garner has called 'a logical trap.'[1] This misunderstanding seems to derive from early publications describing the FST, in which the authors stated that: "We suggested that this characteristic and readily identifiable behavioural immobility reflects a state of despair in the rat" and they went on to describe the immobility as "a behavioral state resembling depression".[43] A later paper by the same group speculated that "the immobile behaviour may reflect a state of lowered mood in the animal".[44] Even by 2001, this position had not changed: "It was hypothesized that immobility reflected the animals having learned that escape was impossible and their having given up hope. Immobility was therefore given the name 'behavioral despair'".[45] It is clear that there was no evidence then that immobility in the FST has anything to do with depression, and that still remains the case.

The possibility that the immobility is analogous to some aspect of depression is most unlikely, not least because depression is a chronic, relapsing disorder, whereas any change in an animal's mood while immersed in water would presumably dissipate soon after they are returned to their home cage. Also, the immobility is prevented by short-term treatment with an antidepressant drug (within 24 hours of the test), whereas the therapeutic lag for antidepressants in humans is typically 6–8 weeks or more. These points, alone, are sufficient to rule out the immobility in the FST as a model of depression or even 'depression-like' behaviour.

Others share this scepticism and have suggested that the immobility is actually a passive coping (survival) mechanism,[46] or that the immobility is driven by anxiety, not depression.[47] Whatever the case, in discussing the use of the FST as a model of depression, and the failure of certain findings to translate into humans, Anyan and Amir[47] hit the nail on the head:

> There are two possible explanations for the discrepancies between human and animal research: either the underlying mechanisms driving depression and anxiety are distinct in humans and rodents or we are misinterpreting animal behavior. We believe the underlying mechanisms are conserved and therefore it is more likely due to interpretation error.

It is interesting to speculate that immobility in the FST could be analogous to the psychomotor retardation seen in depression. If so, perhaps that is exacerbated by (swim) stress? There is evidence from one early study, which tested the psychomotor speed of depressed patients, and found that this was reduced when the task was made more stressful.[48] Until this is confirmed, the only safe conclusion is that the FST enables measurement of 'stress-induced immobility.' To describe it as anything else, or to align it with any aspect of depression, is mere conjecture.

In short, the merits of the FST as a useful, high-throughput predictive screen for antidepressant drugs are borne out by more than 40 years of evidence, but there is no reason to suppose that it produces an animal 'model' of depression. Indeed, it is not obvious what aspect of depression would be induced, or exacerbated, by immersion in a tank of water for a few minutes.

The Open Field Test (OFT) as a follow-up (secondary) drug screen

As discussed above, if a drug is claimed to have beneficial effects on mood, on the basis that it increases an animal's motor activity in procedures such as the FST, then it is essential to rule out the possibility that these drugs are merely increasing animal motor activity, non-specifically. To achieve that, most studies go on to check the effect of the test drug on an animal's ambulation in an arena, known as an 'Open Field'. The OFT is alluringly simple to carry out because it merely involves placing the animal in an arena for a few minutes and monitoring how much it moves around — apparently, neither the animals nor the experimenters need any expert training. However, there are many reasons why the evaluation of locomotor activity in an Open Field is far from straightforward (see Stanford[49]).

The first, and most important, point to bear in mind is that the OFT was developed to study animal emotionality, not motor activity. However, it soon became clear that the evaluation of such animals' emotionality parameters is confounded by differences in their ambulation inside the arena and *vice versa*.[50] Secondly, the behaviour of the animals in this test depends on the physical characteristics of the arena, including its size, shape, the material used for its construction, light intensity and so on (see Walsh and Cummins[51]). Thirdly, animals express many types of behaviour

in the Open Field, and some of these will be incompatible with ambulation. For instance, a change in grooming or rearing, following a drug treatment or genetic mutation, will inevitably affect an animal's locomotor activity indirectly, because they cannot express these behaviours and move around at the same time.[49,51–53]

It follows that locomotor activity in the Open Field is not a single, independent experimental variable. As with all tests of this sort, changes in animal behaviour in the Open Field are uninterpretable unless all other aspects of the behavioural profile have been taken into account. This is especially the case when the drug or genetic mutation being tested is intended to affect the mood of the animals. The best way to evaluate the effects of a drug on locomotor activity, specifically, is to monitor its effects on behaviour while the animals are in a familiar, naturalistic environment — ideally their home cage[54] — and there is now equipment that makes that feasible.

What can be rescued from the misunderstanding and misuse of animal 'models'?

One of the recurrent themes of this article is that, despite there being no validated rodent model for full-blown depression, some protocols might induce behavioural change(s) that could be analogous to specific element(s) of depression. In such cases, authors should avoid using flimsy descriptions of what they have studied ('disorder-like') and should specify and scientifically justify the aspect(s) of the disorder they have in mind (e.g. anhedonia, low self-esteem, psychomotor retardation). Of course, such specific inferences also need careful validation, together with an acknowledgment that they are not necessarily confined to a single psychiatric disorder.

That change of mindset would align closely with the growing interest in how human endophenotypes map onto specific neuronal pathways and/or genetic polymorphisms. This is the rationale for the international Prism Project (https://prism-project.eu). This programme of research aims to gather data on social withdrawal and cognitive deficits, which are shared across multiple psychiatric disorders (Alzheimer's disease, schizophrenia and major depressive disorder), and to identify how the underlying biological parameters differ in these illnesses. One objective of this project is to discern distinct endophenotypes and then, through back-translation, to develop valid, evidence-based animal models that would enable meaningful basic science research that is not permitted in humans.[9] This approach will be complemented, and informed, by parallel exploratory investigational new drug studies in humans (pre-Phase 1, clinical trial), which will, to some extent, reduce the number of animals needed for approval of a full clinical trial.[55] A better understanding of the underlying endophenotypes is essential for personalised (stratified) medicine, and to help explain why some patients respond to a given drug treatment, while others do not.

This approach, which regards psychiatric disorders not as single, defined entities but as assemblies of abnormal endophenotypes[9] is well advanced, but its basic principles have yet to permeate preclinical laboratories.

What can funders and journals do to help?

Research funders and peer-reviewed scientific journals are well placed to take steps that will strengthen the rigour and validity of animal models: that is, how they are used and how the findings are interpreted and described. Obviously, funders need to ensure the validity of the research that they sponsor, but journals can help too. Yet, journal editorial boards have been remarkably nervous about taking any action to improve matters. An exception is the *Journal of Psychopharmacology,* which took the lead in 2019 by revising their Guidelines for Authors to include the point:

> Studies aimed at producing an animal 'model' of a psychiatric disorder (or 'disorder-like' behaviour) in humans should also include a statement to justify the extent to which the experimental procedure produces a validated animal analogue of the human condition, bearing in mind the diagnostic criteria specified in the Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)

(see https://journals.sagepub.com/author-instructions/JOP#Preclinical%20Studies).

It is too early to assess how conscientiously authors are responding to this rubric, but it is certainly a nudge in the right direction. The adoption of a similar policy by other journals, which publish papers in which the authors claim to have used an animal model of a complex human disorder, is long overdue.

Declaration of conflicting interests

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

Ethical approval

Ethics approval was not required for this review article.

Informed consent

Informed consent was not required for this review article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. Garner JP. 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* 2014; 55: 438–456.

2. Garner JP, Gaskill BN, Weber EM, et al. Introducing therioepistemology: the study of how knowledge is gained from animal research. *Lab Anim (NY)* 2017; 46: 103–113.
3. Pound P and Ritskes-Hoitinga M. Is it possible to overcome issues of external validity in preclinical animal research? Why most animal models are bound to fail. *J Transl Med* 2018; 16: 304.
4. Stanford SC. Confusing preclinical (predictive) drug screens with animal 'models' of psychiatric disorders, or 'disorder-like' behaviour, is undermining confidence in behavioural neuroscience. *J Psychopharmacol* 2017; 31: 641–643.
5. Lafollette H and Shanks N. The origin of speciesism. *Philosophy* 1996; 71: 41–61.
6. Greek R and Kramer LA. The scientific problems with using non-human animals to predict human response to drugs and disease. In: Herrmann K and Jayne K (eds) *Animal experimentation: working towards a paradigm change*. Leiden: Brill, 2019, pp. 391–416.
7. Anon. *The diagnostic and statistical manual of mental disorders*. 5th ed. (DSM-5). Philadelphia: American Psychiatric Association, 2013, 947 pp.
8. World Health Organization. *ICD-10 international statistical classification of diseases and related health problems, 10th Revision*. Geneva: World Health Organization, 2016.
9. Kas MJ, Penninx B, Sommer B, et al. A quantitative approach to neuropsychiatry: the why and the how. *Neurosci Biobehav Rev* 2019; 97: 3–9.
10. Baumeister AA, Hawkins MF and Uzelac SM. The myth of reserpine-induced depression: role in the historical development of the monoamine hypothesis. *J Hist Neurosci* 2003; 12: 207–220.
11. Garattini S, Giachetti A, Jori A, et al. Effect of imipramine, amitriptyline and their monomethyl derivatives on reserpine activity. *J Pharm Pharmacol* 1962; 14: 509–514.
12. Pawłowski L and Nowak G. Biochemical and pharmacological tests for the prediction of ability of monoamine uptake blockers to inhibit the uptake of noradrenaline *in-vivo*: the effects of desipramine, maprotiline, femoxetine and citalopram. *J Pharm Pharmacol* 1987; 39: 1003–1009.
13. Seligman ME. Learned helplessness. *Annu Rev Med* 1972; 23: 407–412.
14. Sherman AD, Sacquitne JL and Petty F. Specificity of the learned helplessness model of depression. *Pharmacol Biochem Behav* 1982; 16: 449–454.
15. van der Kolk B, Greenberg M, Boyd H, et al. Inescapable shock, neurotransmitters, and addiction to trauma: toward a psychobiology of posttraumatic stress. *Biol Psychiatr* 1985; 20: 314–325.
16. Maier SF and Seligman ME. Learned helplessness at fifty: insights from neuroscience. *Psychol Rev* 2016; 123: 349–367.
17. Hendriksen H, Korte SM, Olivier B, et al. The olfactory bulbectomy model in mice and rat: One story or two tails? *Eur J Pharmacol* 2015; 753: 105–113.
18. Song C and Leonard BE. The olfactory bulbectomised rat as a model of depression. *Neurosci Biobehav Rev* 2005; 29: 627–647.
19. Rajkumar R and Dawe GS. OBscure but not OBsolete: perturbations of the frontal cortex in common between rodent olfactory bulbectomy model and major depression. *J Chem Neuroanat* 2018; 91: 63–100.
20. Kelly JP, Wrynn AS and Leonard BE. The olfactory bulbectomized rat as a model of depression: an update. *Pharmacol Ther* 1997; 74: 299–316.
21. Chen B, Klarmann R, Israel M, et al. Difference of olfactory deficit in patients with acute episode of schizophrenia and major depressive episode. *Schizophr Res* 2019; 212: 99–106.
22. Willner P, Towell A, Sampson D, et al. Reduction of sucrose preference by chronic unpredictable mild stress, and its restoration by a tricyclic antidepressant. *Psychopharmacology (Berl)* 1987; 93: 358–364.
23. Stanford SC. Central noradrenergic neurones and stress. *Pharmacol Ther* 1995; 68: 297–342.
24. Patel D, Kas MJ, Chattarji S, et al. Rodent models of social stress and neuronal plasticity: relevance to depressive-like disorders. *Behav Brain Res* 2019; 369: 111900.
25. Pagliusi M Jr, Bonet IJM, Brandão AF, et al. Therapeutic and preventive effect of voluntary running wheel exercise on social defeat stress (SDS)-induced depressive-like behavior and chronic pain in mice. *Neuroscience* 2020; 428: 165–177.
26. Khalifeh M, Hobeika R, El Hayek L, et al. Nicotine induces resilience to chronic social defeat stress in a mouse model of water pipe tobacco exposure by activating BDNF signaling. *Behav Brain Res* 2020; 382: 112499.
27. Misiewicz Z, Iurato S, Kulesskaya N, et al. Multi-omics analysis identifies mitochondrial pathways associated with anxiety-related behavior. *PLoS Genet* 2019; 15: e1008358.
28. Fan Y, Chen P, Raza MU, et al. Altered expression of Phox2 transcription factors in the locus coeruleus in major depressive disorder mimicked by chronic stress and corticosterone treatment *in vivo* and *in vitro*. Neuroscience 2018; 393: 123–137.
29. Fantegrossi WE, Murnane KS and Reissig CJ. The behavioral pharmacology of hallucinogens. *Biochem Pharmacol* 2008; 75: 17–33.
30. Kara NZ, Stukalin Y and Einat H. Revisiting the validity of the mouse forced swim test: systematic review and meta-analysis of the effects of prototypic antidepressants. *Neurosci Biobehav Rev* 2018; 84: 1–11.
31. Fitzgerald PJ, Yen JY and Watson BO. Stress-sensitive antidepressant-like effects of ketamine in the mouse forced swim test. *PLoS One* 2019; 14: e0215554.
32. Bogdanova OV, Kanekar S, D'Anci KE, et al. Factors influencing behavior in the forced swim test. *Physiol Behav* 2013; 118: 227–239.
33. Kokras N, Antoniou K, Mikail HG, et al. Forced swim test: What about females? *Neuropharmacology* 2015; 99: 408–421.

34. Rupniak NMJ and Kramer MS. NK1 receptor antagonists for depression: Why a validated concept was abandoned. *J Affect Disord* 2017; 223: 121–125.
35. Rasmussen N. Making the first anti-depressant: amphetamine in American medicine, 1929–1950. *J Hist Med Allied Sci* 2006; 61: 288–323.
36. McIntyre RS, Lee Y, Zhou AJ, et al. The efficacy of psychostimulants in major depressive episodes: a systematic review and meta-analysis. *J Clin Psychopharmacol* 2017; 37: 412–418.
37. Kitada Y, Miyauchi T, Satoh A, et al. Effects of antidepressants in the rat forced swimming test. *Eur J Pharmacol* 1981; 72: 145–152.
38. López-Cruz L, Salamone JD and Correa M. Caffeine and selective adenosine receptor antagonists as new therapeutic tools for the motivational symptoms of depression. *Front Pharmacol* 2018; 9: 526.
39. Teng J, Zhou W, Zeng Z, et al. Quality components and antidepressant-like effects of GABA green tea. *Food Funct* 2017; 8: 3311–3318.
40. Fan J, Li BJ, Wang XF, et al. Ghrelin produces antidepressant-like effect in the estrogen deficient mice. *Oncotarget* 2017; 8: 58964–58973.
41. Marin IA, Goertz JE, Ren T, et al. Microbiota alteration is associated with the development of stress-induced despair behavior. *Sci Rep* 2017; 7: 43859.
42. Yan TC, McQuillin A, Thapar A, et al. NK1 (TACR1) receptor gene 'knockout' mouse phenotype predicts genetic association with ADHD. *J Psychopharmacol* 2010; 24: 27–38.
43. Porsolt RD, Anton G, Blavet N, et al. Behavioural despair in rats: a new model sensitive to antidepressant treatments. *Eur J Pharmacol* 1978; 47: 379–391.
44. Porsolt RD. Animal model of depression. *Biomedicine* 1979; 30: 139–140.
45. Porsolt RD, Brossard G, Hautbois C, et al. Rodent models of depression: forced swimming and tail suspension behavioral despair tests in rats and mice. *Curr Protoc Neurosci* 2001; Chapter 8: Unit 8.10A.
46. Molendijk ML and de Kloet ER. Coping with the forced swim stressor: current state-of-the-art. *Behav Brain Res* 2019; 364: 1–10.
47. Anyan J and Amir S. Too depressed to swim or too afraid to stop? A reinterpretation of the forced swim test as a measure of anxiety-like behavior. *Neuropsychopharmacology* 2018; 43: 931–933.
48. Blackburn IM. Mental and psychomotor speed in depression and mania. *Br J Psychiatry* 1975; 126: 329–335.
49. Stanford SC. The open field test: reinventing the wheel. *J Psychopharmacol* 2007; 21: 134–135.
50. Wilcock J and Broadhurst PL. Strain differences in emotionality: open-field and conditioned avoidance behavior in the rat. *J Comp Physiol Psychol* 1967; 63: 335–338.
51. Walsh RN and Cummins RA. The open-field test: a critical review. *Psychol Bull* 1976; 83: 482–504.
52. Roth KA and Katz RJ. Stress, behavioral arousal, and open field activity — a reexamination of emotionality in the rat. *Neurosci Biobehv Rev* 1979; 3: 247–263.
53. Ennaceur A. Tests of unconditioned anxiety — pitfalls and disappointments. *Physiol Behav* 2014; 135: 55–71.
54. Porter AJ, Pillidge K, Tsai YC, et al. A lack of functional NK1 receptors explains most, but not all, abnormal behaviours of NK1R−/− mice. *Genes Brain Behav* 2015; 14: 189–199.
55. US Department of Health and Human Services Food and Drug Administration Center for Drug Evaluation and Research (CDER). *Exploratory IND studies. Guidance for industry, investigators, and reviewers*. Rockville: CDER, 2006, 13 pp.