CLINICAL INVESTIGATION

Two decades testing interventions in transgenic mouse models of Alzheimer's disease: designing and interpreting studies for clinical trial success

Kieren Egan*,[1]
& Malcolm Macleod[1]
[1]Center for of Clinical Brain Sciences, The University of Edinburgh, Edinburgh, UK
*Author for correspondence: kieren.egan@gmail.com

The increasing prevalence of Alzheimer's disease (AD) poses considerable socioeconomic challenges. Decades of experimental research are yet to lead to the development of effective disease-modifying interventions. The limitations of *in vivo* research in AD are currently poorly understood and a deeper understanding of these might assist future research and trial design. Here we use examples from translational research in AD and across the neurosciences to illustrate how we might increase experimental rigour and thereby raise the validity of studies. We show that there are considerable weaknesses in the *in vivo* modeling of AD, and therefore clinical trials based on claims of efficacy in animals should proceed only after it has been shown that those claims are well founded.

Keywords: clinical trial design • experimental Alzheimer's disease • experimental validity • transgenic mouse models • translational failure

The global burden of Alzheimer's disease (AD) is expected to substantially increase in the years ahead. Dementia is currently is estimated to affect 44 million individuals, reaching 135 million by 2050 [1]. This increase will result in an unprecedented and indiscriminate socioeconomic challenge; patients need increasing assistance as the disease progresses. The cost of caring for each patient is thought to be €20,000 – exceeding the cost for both cancer and cardiovascular disease [2].

Our understanding of the condition has advanced considerably in the last 40 years; including the characterization of amyloid-β [3,4], tau neurofibrillary tangles [5] and the approval of a limited number of treatments, including acetylcholinesterase inhibitors (e.g., donepezil, galantamine, rivastigmine and tacrine) and the NMDA antagonist memantine [6,7]. These treatments provide moderate symptomatic benefits and are often used in the early-to-middle stages of the condition. In spite of their widespread use, these interventions are not useful for all dementia patients and do not address the progressive pathological and behavioral deterioration that occurs in clinical AD [8]. Therefore, despite our many advances, we have an urgent unmet medical need for interventions that are capable of slowing, halting and ultimately reversing the progressive neurodegenerative processes that occur.

This roadblock to developing novel intervention has been somewhat surprising considering the scientific communities' extensive research efforts and development of *in vivo* models. For example, there are numerous animal models of the condition that are currently in use, including *Caenorhabditis elegans, Drospohilla melongaster* and rodent models injected with amyloid to produce Alzheimer-like pathologies [9]. The animal model that has developed the greatest scientific interest is the transgenic mouse model and within this review we term amyloid- and tau-based animal models 'transgenic mouse models'. First produced by Games and colleagues in 1995 [10] transgenic mouse models of AD have been engineered to manifest different aspects of the condition including amyloid-β and tau expression, neurodegeneration and

FUTURE SCIENCE part of fsg

neurobehavioral deficits [11,12]. Such animal models have provided experimental settings to test interventions for the likelihood of clinical efficacy. The popularity of these models for this purpose is demonstrated by the observation that over 300 interventions have been tested in the Tg2576 mouse model [13]. However, this has not been reflected in clinical trial success.

This translational road block observed in AD has posed fundamental biological questions about construct validity (i.e., are we modeling what we think we are modeling?) and the conduct, reporting and the internal and external validity of studies in the field. It has been suggested, for instance, that the use of these models has encouraged the development of treatments that prevent the development of the transgenic phenotype rather than treating established disease [13]. However, the question remains whether other methodological reasons may also be contributing to translational failure.

This translational failure might occur if biological truths were not reflected in experimental results, either at the preclinical or the clinical trial stage (Tables 1 & 2). Alternatively, it could be that there is a mismatch between biological truths in animal experiments and biological truths in humans – that the animal models do not model human disease with sufficient fidelity to be useful in drug development (e.g., construct validity issues, see section 'Which transgenic model').

A deeper understanding of transgenic model studies may provide evidence to help address these questions while simultaneously aiding the design of future preclinical and clinical trials in AD. Here we examine the possible limitations of experimental AD through using a systematically collated data set of interventions tested in transgenic mouse models of Alzheimer's disease from the Collaborative Approach to Meta-analysis and Review of Animal Data from Experimental Studies (CAMARADES) database and examining similar issues across animal modeling in neuroscience.

Recent clinical trials in AD

Clinical trials in AD have faced considerable challenges. It is estimated that 93% of all clinically tested CNS interventions fail to make it to the marketplace (7% worse than the market average) and for those that do, it takes an average of 12.6 years [14]. Those interventions that do make it to the later stages in the clinic (e.g., Phase III) often become high profile. Subsequent failures can provoke fundamental questions asked regarding our understanding of the disease process as a whole and our ability to intervene [15].

Take for example, studies targeting the inhibition of γ-sectretase. Semagecaestat (LY450139) is a compound that acts as a specific γ-secretase inhibitor and demonstrated a lowering of amyloid-β in the brain and cerebrospinal fluid in animals and cerebrospinal fluid amyloid-β in patients [16]. Despite such encouraging data, two Phase III clinical trials in over 2000 mild-to-moderate AD patients suggested that the drug was associated with a decline in cognition and function alongside a substantial side-effect profile [15]. Elsewhere, other γ-secretase studies (e.g., Tarenflurbil [Myriad Genetics & Laboratories, UT, USA] and avagecestat) also failed to demonstrate clinical efficacy hence there are now considerable questions about the suitability of γ-secretase as a therapeutic target for clinical AD.

Complications have also arisen in the field of passive and active immunization [15]. For example, Bapineuzumab (Pfizer, NY, USA) is a monoclonal antibody that blinds to both soluble and insoluble fibrillar amyloid-β_{1-5}. Possible clinical efficacy was suggested in transgenic studies with a reduction of amyloid burden [19] and thus the intervention was taken forward to clinical trials. Although the intervention failed to meet primary end points at Phase II clinical studies, a modest improvement was identified in a subgroup population analysis [15]. Those individuals most likely to benefit were those with smaller brain volumes and those without the *ApoE4* allele. However, when the intervention

Table 1. Preclinical trial failures.

Scientific truth	Efficacy reported in preclinical trials	Efficacy not reported in preclinical trials
Truly positive studies	Positive preclinical trial results are a faithful representation of biological truth	Preclinical studies are falsely negative. Plausible reasons for this may include construct validity issues, inappropriate outcome measure selection or random chance
Truly negative studies	Preclinical trials are falsely positive. Possible reasons may include: insufficient sample size, random chance, outcome measure selection, publication bias and study quality bias	Negative preclinical trial results are a faithful representation of the biological truth

Table 2. Clinical trial failures.

Scientific truth	Efficacy reported in clinical trials	Efficacy not reported in clinical trials
Truly positive studies	Positive clinical trial results are a faithful representation of biological truth	Clinical trials are falsely negative. Possible reasons for this may include issues regarding selection of trial population, outcome measure based issues, compliance or other biases such as study quality
Truly negative studies	Clinical trials are falsely positive. Possible reasons may include: insufficient sample size, random chance, outcome measure selection, publication bias and study quality bias	Negative clinical trial results are a faithful representation of biological truth

reached four Phase III trials this subpopulation did not demonstrate improvements with or without the *ApoE3* allele. Similarly, for active immunization, the AB-1792 study was supported by data from transgenic studies showing improvements for both structural and behavioral outcomes. However, as well noted, the AN-1792 Phase II study was interrupted after the development of meningoenciphelitus in 6% of patients [15].

Therefore, while these represent a small selection of recent clinical trial failures, the reasons for clinical trial failure are often complex and multifactorial. While it could be that these clinical studies are falsely negative, the presence of subgroup analyses and the progression of interventions on moderate suggestions of efficacy through different clinical phases suggest this to be a less likely scenario. Therefore, if clinical findings are truly biologically negative, the explanation must be that either: there is a disconnect between current transgenic animals and humans that we will not be able to circumvent; or the way in which we design, perform and report preclinical studies is not optimal for clinical translation.

Might the transgenic mouse model experiments be internally flawed?

Experiments testing interventions in transgenic mouse models are designed to ascertain whether a given intervention is likely to improve clinical outcomes in AD patients. A fundamental cornerstone of reaching publication (by peer review) involves determining whether or not these differences reach statistical significance. However, could it be that as a scientific community we have inadvertently stacked in our favour the odds of finding such statistical significance through inappropriate experimental design? Here we discuss a number of internal validity issues regarding the testing of interventions in transgenic mouse models of AD, including sample size calculation, the reporting of blinded assessment of outcome and the reporting of random allocation to group.

Sample size calculation

Previous studies have suggested that animal studies of neurological disorders are usually underpowered [21,22]. Essentially, finding statistical significance depends on a number of features: the null hypothesis tested, the sample size, the size of effect and its variance and the critical p-value chosen to represent statistical significance (α level). Generally speaking, the larger the group size and the size of effect compared to the variance, the greater the likelihood of finding statistically significant differences between groups. Conversely, particularly where the prior probability of a drug truly improving outcome is low, underpowered studies are at greater risk that a statistically significant finding is falsely positive. The question is: how many preclinical studies in AD perform sample size calculations and how many animals per group are routinely used?

From a systematically collated data set in the CAMARADES database (over 400 published articles testing interventions in transgenic mouse models of AD) we did not identify any publications where a sample size calculation had been conducted. This finding is consistent with experience across the animal modeling of neurological disorders [23]. Perhaps as a consequence, the sample sizes were relatively small (seven in control group and nine in the treatment group, respectively). While it could be argued that cost, attrition and mouse availability may play a role in how feasible it is to design studies that are sufficiently powered, the absence of these calculations across neurological disorders suggests that general experimental design also plays a role. While the Animal Research: Reporting of *In Vivo* Experiments (ARRIVE) guidelines suggest sample size calculations should be conducted for *in vivo* experiments [24] these guidelines also state that explaining dropouts and how the number of animals used was decided on should be recorded: something that is not always clear in the published literature.

For the issue of sample size in AD, many parallels may be drawn from preclinical studies in amyotrophic

lateral sclerosis (ALS). For example, these animal models are often created using transgenic methods (e.g., *SOD1* mutation), which roughly represents 5% of the population who suffer from the progressive neurodegenerative condition. Translational failure has also been an issue here, perhaps most prominently where an intervention was tested in over 400 patients that caused a worsening of symptoms [25]. Subsequent systematic review work suggested that the single most influential factor that contributed to translational failure in the ALS field was underpowered *in vivo* studies [26].

Identifying the sample size required

To calculate the estimated effect size for a given experiment is relatively straightforward: one of the simplest approaches is to calculate Cohens d. Cohens d is a measure of effect that requires an estimate of the expected effect size and the variance (see formula below). It is calculated by taking the mean in the treatment group (M1) minus the mean in the control group (M2) divided by the pooled variance [65]:

$$d = \frac{M_1 - M_2}{\sigma_{pooled}}$$

with standard error:

$$SE = \sqrt{\frac{N}{n_1 n_2} + \frac{d_1^2}{2(N-2)}}$$

The difficulty for transgenic experiments is the idiosyncrasy of research: experiments will differ in terms of intervention tested, outcome assessed and transgenic mouse model used (see sections 'Which transgenic model' and 'Which outcome measure'). Therefore, this means that a 'one fits all' approach for sample size in experimental AD is difficult to identify. Say, however, we seek 80% power (the lower end of conventionally expected experimental power) to detect a standardized difference between experimental groups (Cohens d; a typical effect size in the AD literature) then we would need 20 animals per group (compared with the medians of seven [control] and nine [treatment] above). Figure 1 illustrates the relationship between Cohens d and sample size in practice.

One potential solution to obtaining (or affording) larger sample sizes might be to perform animal experiments across different centers, as is routinely performed in clinical trials. An example of how this might be achieved has recently been initiated where an intervention will be tested across many different research centers in animal models of stroke [63]. The authors plan to perform multicenter 'Phase III'-type preclinical trials of promising, novel ischemic stroke therapies to inform the possible transition from animal modeling to clinical trial. Performing studies in this manner would allow sample sizes (and therefore power) to be increased, reducing the likelihood of spurious findings being achieved.

Blinded assessment of outcome

Blinded or 'masked' outcomes are a method of improving experimental rigour by masking the identity of the control and treatment groups from those who measure the outcome, handle the data or analyze data. Across experimental neuroscience the reporting of blinded assessment of outcome is relatively uncommon and using meta-analysis of data from animal models of multiple sclerosis Vesterinen and colleagues were able to demonstrate that the reporting of blinding was associated with smaller estimates of neurobehavioral effect size [22]. Similar findings have been identified in preclinical studies of stroke and Parkinson's disease [23]. While there are no empirical data at present to suggest that the blinded assessment of outcome impacts on observed effects in AD, guidelines produced by the Alzheimer's Drug Discovery Foundation call for its use and reporting in published literature [21]. We estimate that fewer than one in five publications modeling AD report this study quality item (16%). It is likely that most experimental outcome measures commonly used in preclinical AD research could be performed blindly. Notwithstanding this point, it is likely that some outcome measures may be more susceptible to influence from this bias than others (e.g., subjective neurobehavioral scoring opposed to automated histological analyses).

Random allocation to group

The reporting of random allocation to treatment group is another concern for internal validity. Thorough random allocation to group commonly involves random number generation that can be performed in basic computer spreadsheet programs. Randomly allocating animals to treatment and control groups reduces the selection bias, which can skew results to suggest an effect of an intervention when in fact the differences are due to pre-existing differences between cohorts of animals (e.g., weight and motor ability). As with blinding, randomization should be possible across most experiments in preclinical AD, although there would be occasional exceptions if animals are required to be matched in some way (e.g., to a variable such as weight or blood pressure). Only one in five publications reported this feature of internal validity in preclinical AD literature, and for those that did the details of

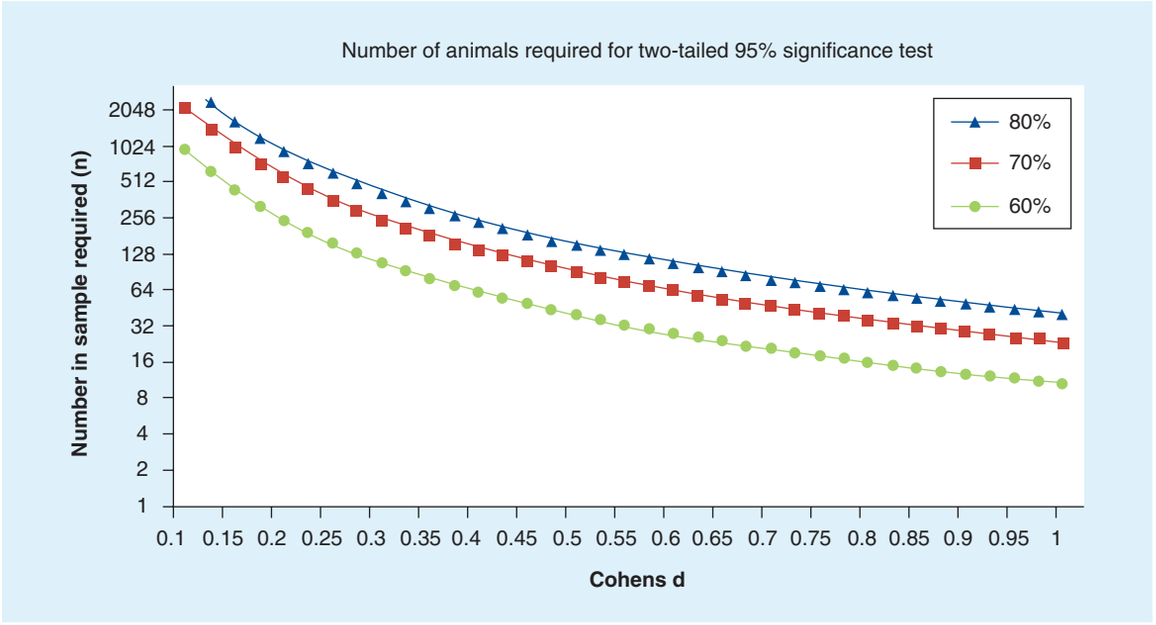

Figure 1. Sample size required according to Cohens d. Each line represents one arm of experimental design for 60, 70 and 80% power according to the estimate of Cohens d and number of animals required. The dotted line shows the number of animals currently used in experimental Alzheimer's disease.

how this was performed were seldom reported. Again, similar issues have been noted across the modeling of neurological disorders and in a number of such models the empirical evidence suggests that features of experimental rigor can influence observed effect sizes. For example, NXY-059 is free radical-trapping agent that was proposed from animal literature to significantly improve outcomes in focal ischemia models. However, when NXY-059 was tested at the clinical trial stage in over 3000 patients these improvements were not demonstrated [27]. Subsequent meta-analysis of the animal data for studies testing NXY-059 suggested that the reporting of random allocation to group (alongside other methodology and study quality items) was associated with significantly smaller estimates of improvements in infarct volume [28].

We are still without direct empirical evidence that demonstrates that failure to conduct sample size calculations, random allocation to group or a blinded assessment of outcome are an important cause of translational failure in AD. Nonetheless, the consistency of findings across experimental neuroscience suggests that this is highly likely, and more rigorous internal validity would be an asset for preclinical trial design in the field.

Are we testing interventions in conditions representative of the clinical setting?

Testing interventions in transgenic mouse models of AD is performed to help develop clinically effective interventions. While there are fundamental differences between the phenotype of transgenic AD mice and the clinical presentation of AD (e.g., overexpression of amyloid precursor protein [*APP*], accelerated accumulation of amyloid and the genetic cause [only present in 5% of clinical cases]), there are a number of external validity issues that are within the scope of experimental design. Here we discuss a number of these in more detail, including: transgenic model selection, outcome measure selection and age at intervention administration.

Which transgenic model?

There are substantial differences between different transgenic mouse models. In terms of pathology, some models will produce only specific features, such as amyloid plaques or tau neurofibrillary tangles in isolation, whereas others such as the triple transgenic (3×TgAD) are capable of producing a combination of these alongside an inflammation and neurobehavioral phenotype [29]. There are also many 'atypical' models, such as those based on inflammation, oxidative stress and sertonergic loss, and each has a specific array of capturing specific attributes of clinical AD [30,31]. The advantages, disadvantages and attributes of different animal model types is not the purpose of this review and have been reviewed extensively elsewhere [32,33]. The Tg2576 mouse model is the most commonly used transgenic model for testing interventions and our data set suggests that one in three studies report its use. The model itself constitutes of an overexpression of mutant form of *APP* with a mutation first identified in a Swedish

kindred (K670/671L) [34]. These mice develop both amyloid plaques and cognitive deficits from around 5 months of age. The four most commonly tested transgenic mouse models for interventions in AD are shown in Figure 2.

One challenge to researchers is that there are now many AD transgenic mouse models that have been used to test interventions (>50 in current literature); and each one differs in complexity, progression and AD-like features. It appears that many research groups use the same model to address many different questions, but as the American Psychologist Abraham Maslow observed, "To the man who only has a hammer, everything he encounters begins to look like a nail", and the justification for choosing which model is most appropriate to address a given hypothesis is not always clear. While there have been suggestions in recent years that models capable of reproducing oligomer species (e.g., Aβ*56) [35] or models that can capture many different aspects of the disease [36] might be superior, there remains little consensus on which model is the most relevant. Meta-analysis techniques may be useful to help identify which transgenic models could give us the most precise or conservative estimates of efficacy, but it is probable that such analyses will face limitations owing to balancing specificity and power.

Therefore, in truth, it is likely to remain difficult to identify which transgenic mouse model is most clinically relevant for testing interventions until we have identified a gold standard as the model that predicts success in clinical trial. We estimate that in published literature only one in five interventions is tested in more than one transgenic mouse model, and it may be that the demonstration of efficacy in different models would provide a greater prospect of translational success. Furthermore, it would be reassuring to demonstrate efficacy in different models that address different aspects of the condition (i.e., plaque/amyloid pathology, tau tangles, neurodegeneration and neurobehavioral deficits).

Which outcome measure?

There have been many calls for more focus on the disease process of AD opposed to specific outcomes such as plaque pathology or tau neurofibrillary tangles [37]. This view has been solidified by a lack of clinical success

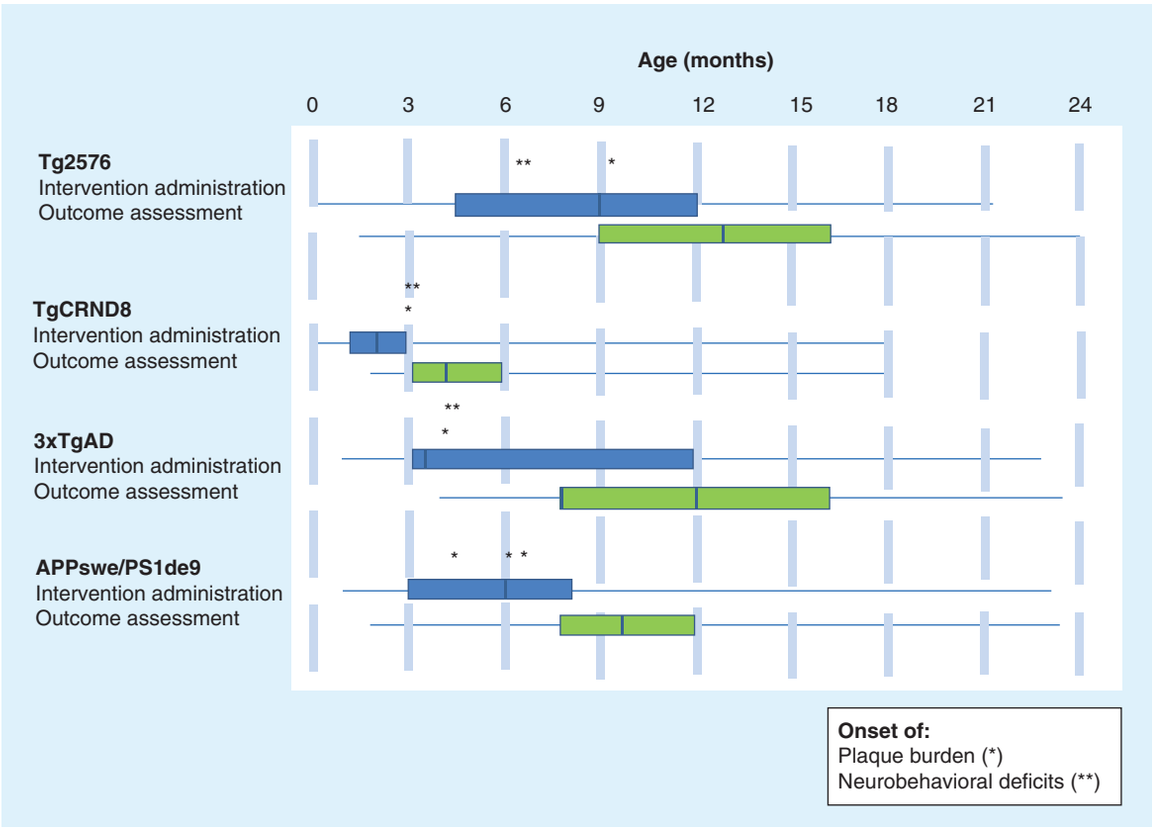

Figure 2. The four most commonly used transgenic mouse models of Alzheimer's disease that have been used for testing interventions (Tg2576, TgCRND8, 3×TgAD and APPswe/PS1de9). Data represents medians, interquartile ranges and minimum maximum. For each transgenic we describe when interventions were administered and outcomes assessed (quartile 1, quartile 2 and quartile 3). Also shown are suggested timescales for the onset of neurobehavioral deficits and plaque pathology [5,8,4,17,18,20].

of interventions that target specific end points such as amyloid-β outcomes (e.g., active immunizations). However, within transgenic mouse models studies, multiple outcomes are often observed with efficacy demonstrated across numerous different types of methodologies.

It is often unclear how the outcomes reported in transgenic mouse studies have been chosen. This is in contrast to other neurological conditions, such as stroke or ALS, where outcome measures are more self-evident, such as infarct volume (stroke), survival (ALS), or motor or sensory neurobehaviors (stroke and ALS). The concern with preclinical studies of AD is that such flexibility in study design makes experiments susceptible to Type 1 error [38], with the possibility that non-significant findings go unreported [39]. This point is demonstrated in the CAMARADES data set where numerous pathological outcomes were often assessed in transgenic studies including plaque burden, amyloid-β40, amyloid-β42, tau, cellular infiltrates (e.g., astrocytosis and microgliosis) alongside neurodegeneration.

Neurobehavioral deficits can also be measured in transgenic animals and there are now well over 20 different neurobehavioral paradigms that have been used to demonstrate efficacy, each differing in methodological variation from one laboratory to the next (e.g., Morris water maze, radial arm water maze and fear conditioning).

Interestingly, transgenic studies differ from clinical studies in the respect that pathological outcome measures are much more commonly reported than neurobehavioral outcomes. For example, the CAMARADES database suggests that pathological outcomes are roughly three-times more likely to be reported than neurobehavioral outcomes. Where pathological outcomes are reported there is a bias for reporting plaque- and amyloid-related outcomes (>50% of publications reported these). Only around one in ten publications report tau or neurodegeneration outcomes: both fundamental features of clinical AD.

Even within specific outcome measures there are different ways to interpret the apparent efficacy of an intervention. For example, for studies testing interventions in transgenic models we found reports focusing on eight different phosphorylation patterns for tau (Table 3). While a degree of experimental flexibility is probably an asset for the external validity of studies, it is not clear which of these would be most reflective of the clinical setting, or indeed how experimenters should choose which antibody to stain with. It is reassuring to note that authors frequently study multiple phosphorylation sites within a single study.

Such experimental flexibility is at times mirrored in the clinic. Commonly used dementia and function scales in use for AD include: the mini-mental state examination [40], The Alzheimer's Disease Assessment Scale cognitive behavior section [41] the clinical dementia rating [42], Adenbrooke's cognitive examination [43] and the Montreal cognitive assessment [44]. Therefore, flexibility of behavioral paradigms must be accommodated for in preclinical experiments; however, it would be advantageous to demonstrate efficacy across multiple methodologies or paradigms the before considering interventions for clinical trial.

For pathological outcomes, one empirical way to identify outcomes that might be of greater clinical interest would be ones with: strong associations with changes in neurobehavior; and a demonstration of low variance and therefore greater reliability of results. While there have been some studies in transgenic models investigating the association between pathological outcomes and neurobehavior [20], there are no definitive data for which pathological outcome measure would be of the greatest interest within each transgenic model. What seems certain is that the ultimate goal of assessing any intervention is to demonstrate behavioral improvements in the clinic. Therefore, if we are to continue to use pathological outcome measures as surrogate measures of neurobehavioral improvement it may be more relevant to peruse interventions if they can demonstrate reasonably strong associations with neurobehavioral improvements.

What age have we administered interventions & assessed outcomes?

There is some debate as to the appropriate age at which animal models best reflect human disease. There is of course a tension between the desire to identify treatment effective in late-stage disease (when the diagnosis is not in question) and the potential benefits of treating either in very early disease or indeed before any disease features are manifest. It seems likely that developing treatments for established disease will be substantially more challenging; and that treatments that are effective early in the course of genetic models of AD are likely to be most effective when given to humans with the same genetic mutations at a very early stage of illness development.

It is evident from our own work and the work of others [51] that studies are being conducted extremely early in the mouse lifespan (e.g., <3 months of age). Using preliminary data we have confirmed these findings across the four most commonly tested transgenic mouse models and Figure 2 compares the age at which interventions are administered and outcomes assessed corresponding to neurobehavioral deficits and plaque burden onset. For example, triple transgenic models most commonly have outcomes assessed at the 6-month stage (shortly after the onset of neurobehavioral deficits

Table 3. Clinical trial failures.

Antibody	Epitope	Example	Ref.
AT8	202/205	Halagappa *et al.* (2007)	[45]
AT180	231-235	Lanz *et al.* (2007)	[46]
PHF1	396/404	Asuni *et al.* (2007)	[47]
AT270	181	Cacammo *et al.* (2007)	[48]
AT100	212/214	Cacammo *et al.* (2007)	[48]
CP13	202	Matsuoka *et al.* (2008)	[49]
12 E8	262/356/394	Cacammo *et al.* (2007)	[48]
AP422	422	Le Corre *et al.* (2006)	[50]

From a systematically identified data set of interventions tested in transgenic mouse models of Alzheimer's disease, we identified eight different phosphorylation states of tau, mapping to different regions of the tau protein.

and plaque pathology). The use of meta-analysis to quantify the impact of age at treatment initiation or outcome assessment may be able to provide empirical guidance as to whether age is a likely contributor to clinical failures.

These will remain crucial external validity questions, unless and until we can identify in early adult life individuals who are at high risk of developing AD. Nonetheless, there are still numerous reasons to be optimistic that external validity in transgenic studies is probable in the near future. Firstly, early identification of AD is looking increasingly feasible: demonstrated by the approval of Florbetapir (Eli Lilly and Company, IN, USA) [52] by the US FDA for the early detection of clinical AD. Furthermore, there appears to be an increasing appreciation that the administration of interventions should be performed late in the transgenic lifespan, as illustrated in the recent studies of the insulin-like drug Liraglutide (Eli Lilly and Company) where APP/PS1 mice of 14 months of age were used [53]. Liraglutide will now be tested at the clinical trial stage.

Can we see all the preclinical data?

Across animal models of neurological disorders, there have been suggestions of a presence of publication bias. Most comprehensively, this has been suggested in animal models of stroke where Sena and colleagues in 2010 [28] identified a suggestion of publication bias in ten out of 16 meta-analysis data sets. Where the authors used the 'trim and fill' function (to account for the potential missing studies) this resulted in a relative reduction in efficacy 31%. Elsewhere a suggestion of publication bias has been observed in animal studies of Parkinson's disease [54].

As previously discussed, there are numerous outcomes of pathological and behavioral interest in transgenic studies. The CAMRARADES preliminary analysis of all interventions tested in transgenic mouse models suggests that approximately one in five pathological outcome measures are missing and one in seven neurobehavioral outcomes are missing, which if included would result in a relative reduction in efficacy of 78.8 and 48.4%, respectively.

Publication bias is found across experimental medicine. While there have been advancements in greater understanding of the importance of publication of neutral and negative findings (e.g., the journal of neutral and negative results in biomedicine) and an increasing trend of universities developing and sharing open data repositories [55] there is still much room for improvement. It is probable that the solution lies in a multidisciplinary approach where authors, journal editors and funding bodies collectively ensure neutral and negative studies are reported and published in full.

A further issue is the selective reporting of outcomes; in a recent analysis of data from across the neurosciences [39] we found twice as many statistically significant results as would be expected. This would occur if authors analyzed data in numerous different ways and picked the statistically significant result to report, or measured multiple outcomes and reported the ones that happened to reach a statistical threshold.

Preclinical trials: making the most of the data we have

There have been extensive efforts to test interventions in clinical trials in recent years in AD, none of which has proven effective [56]. The use of evidence-based trial design may be one way to help researchers use existing data better in the design of trials in the neurosciences. Data from meta-analysis has been extensively used to develop guidelines for future animal studies [24,57–59], but there are emerging examples of how empirical data from preclinical studies can be collated to inform clinical trial design.

A particularly prominent example is in the field of stroke where there have been similar challenges of translational failure where of over 1000 interventions tested

for efficacy in animal models only one led to effective treatment in clinical practice [60]. When hypothermia was suggested to be a suitable intervention to test at the clinical trial stage investigators first set out to systematically review literature on hypothermia in animal models of focal ischemia [61,62]. Through exploring sources of excess heterogeneity the authors were able to identify: hypothermia unequivocally improved outcomes in animals, efficacy could be demonstrated across a number of different experimental designs (e.g., sex, species and anesthetic) and was retained in high-quality studies (e.g., studies with blinding and randomization). Furthermore (and crucially), the authors were able to demonstrate that the intervention appeared effective under conditions relevant to those that might be achieved in the clinical setting (e.g. time to treatment duration and depth of hypothermia), and even after accounting for the potential impact of publication bias. Such work performing meta-analysis of preclinical studies played an essential role of the €11 million Seventh Framework Programme of the European Union-funded clinical trial for hypothermia in acute ischemic stroke (recruitment began in 2013 [64]). Although we are still await the clinical findings of this study, the methodology provides a systematic evaluation of the available evidence and thus demonstrates how empirical evidence can be used to help guide future trial design.

Conclusion

In this paper, we set out to critically review two decades of testing interventions in transgenic mouse models of Alzheimer's disease. There are many plausible reasons for the translational failure from bench to bedside and here we have discussed some of the internal and external validity issues that may have played a role alongside the potential impact of publication bias. It is unquestionable that our knowledge of Alzheimer's disease has advanced considerably in recent years: transgenic mouse model development may be considered one the pinnacles of this.

Despite these successes we have used over 10,000 transgenic animals to test interventions without this leading to novel clinical interventions. Therefore, it is essential that the AD community utilizes the knowledge obtained so far to help future trials in AD (preclinical and clinical) and beyond. Current evidence from across the neurosciences suggest that if we are to improve the translational hit rate in AD we must be rigorous in assessing evidence before embarking on clinical trials.

This means we must first be able to first identify that there is a substantial enough data set to merit clinical trial testing, second, we must trust the results of the individual experiments (e.g., sufficient power, blinded and randomized), and third, that the intervention is tested in settings relevant to the clinical trial environment (e.g., age and outcome relevant). Furthermore, greater emphasis should be placed on designing and reporting experiments beyond amyloid, and more specifically for outcomes of interest to clinical trial design (e.g., tau and neurodegeneration). Where possible, evidence synthesis techniques (as demonstrated by the use of systematic review and meta-analysis for hypothermia in stroke animal models) are reasonably robust methods for AD to ascertain whether claims of efficacy are well founded or not.

Future perspective

There remains a prominent interest in testing interventions in transgenic mouse models that is still ongoing in experimental Alzheimer's disease. If we are to circumvent future clinical trial failures it is vital that we

Executive summary

Background

- Alzheimer's disease is an increasingly prevalent condition currently without effective long-term treatment. Candidate interventions in recent years have consistently failed to demonstrate statistical benefit at clinical trial stage.

Lack of clinical trial success: possible causation

- Clinical trial failures pose a number of questions regarding the conduct of experimental science. It could be that the way in which we have used transgenic mouse models of the condition has not had sufficient rigor to encourage translational hits.

Potential important issues regarding testing interventions in transgenic models

- Studies testing interventions in transgenic mouse models of Alzheimer's disease are often lacking in internal validity (with no prior sample size calculation or statement regarding blinding or randomization). In addition, there are concerns regarding the external validity of outcomes (e.g., the age of animals used and the outcome measurement used). Such issues are common to experimental models of a number of neurological conditions (e.g., stroke, multiple sclerosis, Parkinson's disease and motor neuron disease). Publication bias is also an issue in these studies, and taking these findings together collectively there are considerable weaknesses in current preclinical research that are of concern.

maximize the utility of existing data using both clinical and preclinical evidence. Evidence synthesis techniques (e.g., meta-analysis) can be used to empirically guide future research in Alzheimer's disease. This can be: to inform guidelines for future conduct in experimental science; and to rigorously assess whether there is substantial evidence in favor of a candidate intervention before embarking on clinical trials. We intend to make data sets collated available freely to researchers in due course.

References

1 Prince M. The Global Impact of Dementia 2013–2050 (2014). www.alz.co.uk/research/G8-policy-brief

2 Hampel H, Prvulovic D, Teipel S *et al.* The future of Alzheimer's disease. The next 10 years. *Prog. Neurobiol.* 95(4), 718–728 (2011).

3 Glenner GG, Wong CW. Alzheimer's disease and Down's syndrome. Sharing of a unique cerebrovascular amyloid fibril protein. *Biochem. Biophys. Res. Commun.* 122(3), 1131–1135 (1984).

4 Janus C, Pearson J, McLaurin J *et al.* A beta peptide immunization reduces behavioural impairment and plaques in a model of Alzheimer's disease. *Nature* 408(6815), 979–982 (2000).

5 Grundke-Iqbal I, Iqbal K, Tung YC, Quinlan M, Wisniewski HM, Binder LI. Abnormal phosphorylation of the microtubule-associated protein tau (tau) in Alzheimer cytoskeletal pathology. *Proc. Natl Acad. Sci. USA* 83(13), 4913–4917 (1986).

6 Birks J. Cholinesterase inhibitors for Alzheimer's disease. *Cochrane Database Syst. Rev.* 25(1), CD005593 (2006).

7 McShane R. Memantine for dementia. *Cochrane Database Syst. Rev.* 19(2), CD003154 (2006).

8 Sterniczuk R, Antle MC, LaFerla FM, Dyck RH. Characterization of the 3×Tg-AD mouse model of Alzheimer's disease. Part 2. Behavioral and cognitive changes. *Brain Res.* 1348, 149–155 (2010).

9 Van Dam D, De Deyn PP. Animal models in the drug discovery pipeline for Alzheimer's disease. *Br. J. Pharmacol.* 164(4), 1285–1300 (2011).

10 Games D, Adams D, Alessandrini R *et al.* Alzheimer-type neuropathology in transgenic mice overexpressing V717F beta-amyloid precursor protein. *Nature* 373, 523–527 (1995).

11 Oddo S, Caccamo A, Shepherd JD *et al.* Triple-transgenic model of Alzheimer's disease with plaques and tangles. intracellular Abeta and synaptic dysfunction. *Neuron* 39(3), 409–421 (2003).

12 Hochgräfe K, Sydow A, Mandelkow EM. Regulatable transgenic mouse models of Alzheimer disease. onset, reversibility and spreading of tau pathology. *FEBS J.* 280(18), 4371–4381 (2013).

13 Zahs KR, Ashe KH. 'Too much good news' – are Alzheimer mouse models trying to tell us how to prevent, not cure, Alzheimer's disease? *Trends Neurosci.* 33(8), 381–389 (2010).

Financial & competing interests disclosure

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.

14 Pangalos MN, Schechter LE, Hurko O. Drug development for CNS disorders: strategies for balancing risk and reducing attrition. *Nat. Rev. Drug Discov.* 6(7), 521–532 (2007).

15 Mullane K, Williams M. Alzheimer's therapeutics: continued clinical failures question the validity of the amyloid hypothesis-but what lies beyond? *Biochem. Pharmacol.* 85(3), 289–305 (2013).

16 Henley DB, May PC, Dean RA, Siemers ER. Development of semagacestat (LY450139), a functional gamma-secretase inhibitor, for the treatment of Alzheimer's disease. *Expert Opin. Pharmacother.* 10(10), 1657–1664 (2009).

17 Savonenko A, Xu GM, Melnikova T *et al.* Episodic-like memory deficits in the APPswe/PS1dE9 mouse model of Alzheimer's disease. Relationships to beta amyloid deposition and neurotransmitter abnormalities. *Neurobiol. Dis.* 18(3), 602–617 (2005).

18 Xiong H, Callaghan D, Wodzinska J *et al.* Biochemical and behavioral characterization of the double transgenic mouse model (APPswe/PS1dE9) of Alzheimer's disease. *Neurosci. Bull.* 27(4), 221–232 (2011).

19 Bard F, Cannon C, Barbour R *et al.* Peripherally administered antibodies against amyloid beta-peptide enter the central nervous system and reduce pathology in a mouse model of Alzheimer disease. *Nat. Med.* 6(8), 916–919 (2000).

20 Westerman MA, Cooper-Blacketer DF, Mariash AF *et al.* The relationship between Abeta and memory in the Tg2576 mouse model of Alzheimer's disease. *J. Neurosci.* 22(5), 1858–1867 (2002).

21 Shineman D, Basi G, Bizon J *et al.* Accelerating drug discovery for Alzheimer's disease. Best practices for preclinical animal studies. *Alzheimers Res. Ther.* 3(5), 28 (2011).

22 Vesterinen HM, Sena ES, French-Constant CF, Williams AF, Chandran, Macleod MR. Improving the translational hit of experimental treatments in multiple sclerosis. *Mult. Scler.* 16(9), 1044–1055 (2010).

23 van der Worp HB, Howells DW, Sena ES *et al.* Can animal models of disease reliably inform human studies? *PLoS Med.* 7(3), e1000245 (2010).

24 Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLoS Biol.* 8(6), e1000412 (2010).

25 Gordon PH, Moore DH, Miller RG *et al.* Efficacy of minocycline in patients with amyotrophic lateral sclerosis. A Phase III randomised trial. *Lancet Neurol.* 6(12), 1045–1053 (2007).

26 Scott S, Kranz JE, Cole J *et al.* Design, power, and interpretation of studies in the standard murine model of ALS. *Amyotroph. Lateral Scler.* 9(1), 4–15 (2008).

27 Shuaib A, Lees KR, Lyden P *et al.* NXY-059 for the treatment of acute ischemic stroke. *N. Engl. J. Med.* 357(6), 562–571 (2007).

28 Macleod MR, van der Worp HB, Sena ES, Howells DW, Dirnagl U, Donnan GA. Evidence for the efficacy of NXY-059 in experimental focal cerebral ischaemia is confounded by study quality. *Stroke* 39(10), 2824–2829 (2008).

29 Schwab C, Klegeris A, McGeer P. Inflammation in transgenic mouse models of neurodegenerative disorders. *Biochim. Biophys. Acta Mol. Basis Dis.* 1802(10), 889–902 (2010).

30 Oikawa D, Akai R, Tokuda M, Iwawaki T. A transgenic mouse model for monitoring oxidative stress. *Sci. Rep.* 2, 229 (2012).

31 Rodríguez JJ, Noristani HN, Verkhratsky A. The serotonergic system in ageing and Alzheimer's disease. *Prog. Neurobiol.* 99(1), 15–41 (2012).

32 McGowan E, Eriksen JF, Hutton M. A decade of modeling Alzheimer's disease in transgenic mice. *Trends Genet.* 22(5), 281–289 (2006).

33 Gotz J, Ittner LM. Animal models of Alzheimer's disease and frontotemporal dementia. *Nat. Rev. Neurosci.* 9(7), 532–544 (2008).

34 Hsiao K, Chapman P, Nilsen S *et al.* Correlative memory deficits, Abeta elevation, and amyloid plaques in transgenic mice. *Science* 274, 99–102 (1996).

35 Lesne Koh MT, Kotilinek L *et al.* A specific amyloid-[beta] protein assembly in the brain impairs memory. *Nature* 440(7082), 352–357 (2006).

36 Oddo S, Caccamo A, Shepherd JD *et al.* Triple-transgenic model of Alzheimer's disease with plaques and tangles. Intracellular Abeta and synaptic dysfunction. *Neuron* 39, 409–421 (2003).

37 Iqbal K, Liu F, Gong CX. Alzheimer disease therapeutics: focus on the disease and not just plaques and tangles. *Biochem. Pharmacol.* 88(4), 631–639 (2014).

38 Ioannidis JPA. Why most published research findings are false. *PLoS Med.* 2(8), e124 (2005).

39 Tsilidis KK, Panagiotou OA, Sena ES *et al.* Evaluation of excess significance bias in animal studies of neurological diseases. *PLoS Biol.* 11(7), e1001609 (2013).

40 Folstein MF, Folstein SE, McHugh PR. Mini-mental state: a practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* 12(3), 189–198 (1975).

41 Cano SJ, Posner HB, Moline ML *et al.* The ADAS-cog in Alzheimer's disease clinical trials. Psychometric evaluation of the sum and its parts. *J. Neurol. Neurosurg. Psych.* 81(12), 1363–1368 (2010).

42 Hughes CP, Berg LF, Danziger WL, Coben LA, Martin RL. A new clinical scale for the staging of dementia. *Br. J. Psych.* 140, 566–572 (1982).

43 Mathuranath PS, Nestor PJ FAU, Berrios GE FAU, Rakowicz WF, Hodges JR. A brief cognitive test battery to differentiate Alzheimer's disease and frontotemporal dementia. *Neurology* 55(11), 1613–1620 (2000).

44 Freitas S, Simões MR, Alves L, Santana I. Montreal cognitive assessment: validation study for mild cognitive impairment and Alzheimer disease. *Alzheimer Dis. Assoc. Disord.* 27(1), 37–43 (2013).

45 Halagappa VKM, Guo Z, Pearson M *et al.* Intermittent fasting and caloric restriction ameliorate age-related behavioral deficits in the triple-transgenic mouse model of Alzheimer's disease. *Neurobiol. Dis.* 26(1), 212–220 (2007).

46 Lanz TA, Salatto CT, Semproni AR *et al.* Peripheral elevation of IGF-1 fails to alter A-beta clearance in multiple *in vivo* models. *Biochem. Pharmacol.* 75(5), 1093–1103 (2008).

47 Asuni AA, Boutajangout A, Quartermain D, Sigurdsson EM. Immunotherapy targeting pathological tau conformers in a tangle mouse model reduces brain pathology with associated functional improvements. *J. Neurosci.* 27(34), 9115–9129 (2007).

48 Caccamo A, Oddo S, Tran LX, LaFerla FM. Lithium reduces tau phosphorylation but not Abeta or working memory deficits in a transgenic model with both plaques and tangles. *Am. J. Pathol.* 170(5), 1669–1675 (2007).

49 Matsuoka Y, Jouroukhin Y, Gray AJ *et al.* A neuronal microtubule-interacting agent, NAPVSIPQ, reduces tau pathology and enhances cognitive function in a mouse model of Alzheimer's disease. *J. Pharmacol. Exper. Ther.* 325(1), 146–153 (2008).

50 Le Corre S, Klafki HW, Plesnila N *et al.* An inhibitor of tau hyperphosphorylation prevents severe motor impairments in tau transgenic mice. *Proc. Natl Acad. Sci. USA* 103(25), 9673–9678 (2006).

51 Zahs KR, Ashe KH. '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).

52 Fleisher AS, Chen KF, Quiroz YT *et al.* Florbetapir PET analysis of amyloid-beta deposition in the presenilin 1 E280A autosomal dominant Alzheimer's disease kindred: a cross-sectional study. *Lancet Neurol.* 1057–1065 (2012).

53 McClean PL, Hölscher C. Liraglutide can reverse memory impairment, synaptic loss and reduce plaque load in aged APP/PS1 mice, a model of Alzheimer's disease. *Neuropharmacology* 76, 57–67 (2014).

54 Rooke EDM, Vesterinen HM, Sena ES, Egan K, Macleod MR. Dopamine agonists in animal models of Parkinson's disease: a systematic review and meta-analysis. *Parkinsonism Relat. Disord.* 17(5), 313–320 (2011).

55 Sandercock P. Negative results: why do they need to be published? *Int. J. Stroke* 7(1), 32–33 (2012).

56 Mangialasche F, Solomon A, Winblad B, Mecocci P, Kivipelto M. Alzheimer's disease: clinical trials and drug development. *Lancet Neurol.* 9(7), 702–716 (2010).

57 Ludolph AC, Bendotti CF, Blaugrund E FAU, Chio A *et al.* Guidelines for preclinical animal research in ALS/MND. *Amyotroph. Lateral Scler.* 11(1–2), 38–45 (2010).

58 Macleod MR, Fisher M, O'Collins V *et al.* Reprint: good laboratory practice: preventing introduction of bias at the bench. *Int. J. Stroke* 4(1), 3–5 (2009).

59 Landis SC, Amara SG, Asadullah K *et al.* A call for transparent reporting to optimize the predictive value of preclinical research. *Nature* 490(7419), 187–191 (2012).

60 O'Collins VE, Macleod MR, Donnan G, Horky LL, van der Worp B, Howells DW. 1,026 experimental treatments in acute stroke. *Ann. Neurol.* 59(3), 467–477 (2006).

61 van der Worp HB, Sena ES, Donnan GA, Howells DW, Macleod MR. Hypothermia in animal models of acute ischaemic stroke: a systematic review and meta-analysis. *Brain* 130(12), 3063–3074 (2007).

62 van der Worp HB, Macleod MR, Kollmar R. Therapeutic hypothermia for acute ischemic stroke: ready to start large randomized trials?. *J. Cereb. Blood Flow Metab.* 30(6), 1079–1093 (2010).

63 Multi-PART. Multicentre preclinical animal research team. www.multi-PART.org

64 Euro HYP. The European Stroke Research Network for Hypothermia – launch of the EuroHYP-1 trial. www.eurohyp.org

65 Egger M, Smith G, Altman D. *Systematic Reviews in Health Care. Meta-Analysis in Context (2nd Edition)*. BMJ, London, UK (2001).