Amyotrophic Lateral Sclerosis. 2008; 9: 4–15

informa
healthcare

REVIEW ARTICLE

Design, power, and interpretation of studies in the standard murine model of ALS

SEAN SCOTT[1], JANICE E. KRANZ[1], JEFF COLE[1], JOHN M. LINCECUM[1], KENNETH THOMPSON[1], NANCY KELLY[1], ALAN BOSTROM[2], JILL THEODOSS[1], BASHAR M. AL-NAKHALA[1], FERNANDO G. VIEIRA[1], JEYANTHI RAMASUBBU[1] & JAMES A. HEYWOOD[1]

[1]*ALS Therapy Development Institute, Cambridge, Massachusetts, and* [2]*Department of Epidemiology and Biostatistics, University of California, San Francisco, USA*

Abstract

Identification of SOD1 as the mutated protein in a significant subset of familial amyotrophic lateral sclerosis (FALS) cases has led to the generation of transgenic rodent models of autosomal dominant SOD1 FALS. Mice carrying 23 copies of the human $SOD1^{G93A}$ transgene are considered the standard model for FALS and ALS therapeutic studies. To date, there have been at least 50 publications describing therapeutic agents that extend the lifespan of this mouse. However, no therapeutic agent besides riluzole has shown corresponding clinical efficacy.

We used computer modeling and statistical analysis of 5429 $SOD1^{G93A}$ mice from our efficacy studies to quantify the impact of several critical confounding biological variables that must be appreciated and should be controlled for when designing and interpreting efficacy studies. Having identified the most critical of these biological variables, we subsequently instituted parameters for optimal study design in the $SOD1^{G93A}$ mouse model. We retested several compounds reported in major animal studies (minocycline, creatine, celecoxib, sodium phenylbutyrate, ceftriaxone, WHI-P131, thalidomide, and riluzole) using this optimal study design and found no survival benefit in the $SOD1^{G93A}$ mouse for any compounds (including riluzole) administered by their previously reported routes and doses.

The presence of these uncontrolled confounding variables in the screening system, and the failure of these several drugs to demonstrate efficacy in adequately designed and powered repeat studies, leads us to conclude that the majority of published effects are most likely measurements of noise in the distribution of survival means as opposed to actual drug effect. We recommend a minimum study design for this mouse model to best address and manage this inherent noise and to facilitate more significant and reproducible results among all laboratories employing the $SOD1^{G93A}$ mouse.

Key words: *G93A mice, riluzole, pre-clinical, SOD1, FALS*

Introduction

Amyotrophic lateral sclerosis (ALS) is a paralytic neurodegenerative disorder caused by the loss of motor neurons. About 3% of cases are caused by single point mutations in the gene encoding SOD1. The identification of SOD1 as the mutated protein in a significant subset of FALS (1) has led to the generation of transgenic rodent models of autosomal dominant SOD1 FALS. The 23-copy human SOD1-G93A transgenic mouse (referred to hereafter as $SOD1^{G93A}$) (2), a mixed hybrid strain, is the most widely used murine model of FALS and is accepted as a standard model for therapeutic studies since Gurney described its use (3).

At least 50 publications describe therapeutic agents, from small molecules to viral vectors, that extend the lifespan of this $SOD1^{G93A}$ mouse (reviewed in (4,5); complete list of references in Table S1). Due to the rapid, unremitting disease course of ALS, these agents often quickly advance into clinical trials. However, to date, no therapy besides riluzole (with its modest effect of possibly two months life extension (6)) has demonstrated a significant impact on the course of the human disease (4,5).

In the past five years we have screened more than 70 drugs in 18000 mice across 221 studies, using rigorous and appropriate statistical methodologies. We used the same strain of $SOD1^{G93A}$ mice used in

Correspondence: S. Scott, ALS Therapy Development Institute, 215 First Street, Cambridge, Massachusetts 02142, USA. Fax: 415 447 1996. E-mail: sscott@als.net

(Received 27 September 2007; accepted 11 December 2007)

ISSN 1748-2968 print/ISSN 1471-180X online
DOI: 10.1080/17482960701856300

all studies summarized in Table S1 (see Methods). We expected to replicate reports of efficacy and to establish both positive controls and metrics to gauge future therapeutic potential. While we were able to measure a significant difference in survival between males and females with great sensitivity, we observed no statistically significant positive (or negative) effects for any of the 70 compounds tested, including several previously reported as efficacious. Here we report, based on our analysis of these results, a possible explanation for these discrepancies.

We used computer modeling and statistical methods applied to our historical database of observations from 5429 SOD1^{G93A} mice in efficacy studies to identify and quantify the impact of the critical confounding biological variables that must be appreciated and controlled in designing studies and ultimately in interpreting them. Specifically, we demonstrate how gender, clustering within litters, and censoring criteria dramatically increase noise in the distribution of survival means. Using this analysis to optimize study designs, we have repeated several of the SOD1^{G93A} studies that have led to clinical trials for ALS. Our results with minocycline, creatine, ritonavir, celecoxib, sodium phenylbutyrate, ceftriaxone, WHI-P131, thalidomide, and riluzole indicate that these compounds have no survival benefit in the SOD1^{G93A} mouse at their reported routes and doses. These findings are generally applicable as they are derived from the same breeding colony at The Jackson Laboratory that has supplied virtually all animals in published efficacy studies in the SOD1^{G93A} mouse. Our recommendations for study design to best address and manage the noise inherent in this model should facilitate improved and more reproducible results among all laboratories employing the model.

Materials and methods

Mice

Efficacy studies used transgenic SOD1^{G93A} mice (strain name B6SJL-Tg(SOD1-G93A)1Gur, # 002726) from The Jackson Laboratory ('JAX', Bar Harbor, Maine), bred by JAX or by Genzyme Transgenic Corporation (GTC, Framingham MA)). Both sources maintain this mixed hybrid SOD1^{G93A} colony by breeding a hemizygous B6SJL-Tg (SOD1-G93A) male to B6SJLF1 dams. To check for presence of the transgene in the progeny, tail biopsies are collected by the breeder from 14-day-old pups, then PCR-genotyped (according to JAX April 03 SOD protocol). Transgenic mice are shipped to us at age 35–45 days, allowing at least a week to acclimatize to our facility (a 12-h light/ dark cycle) before assigning them to a study. All statistical modeling analyses were performed on data from JAX mice only.

Genotypic analysis

Mice, hemizygous for the transgene and living past 180 days, were assayed for a reduction in transgene copy number using quantitative Southern blotting (7), which confirmed that mice living past 180 days had 10–14 copies of the SOD1^{G93A} transgene. A reduction in transgene copy number can result from meiotic rearrangement of the transgene array in one or more hemizygous stud males. This was confirmed by historical pedigree analysis of a subset of the mice living past 180 days that also had reduced transgene copy number. The pedigree analysis demonstrated that the long-lived progeny were the offspring of a single hemizygous male stud.

SOD1^{G93A} efficacy studies

Mice are separated into treatment and vehicle cohorts at age day 45. To ensure minimal variability between cohorts, each cohort is defined by the following constraints: balanced for gender, $n=12$ males and $n=12$ females; age-matched; littermate-matched. Littermates are defined as offspring of the same non-transgenic dam and transgenic sire, born on the same day. Specifically, each male (and female) in the treatment group has a littermate brother (and sister, respectively) in vehicle group; bodyweight balanced. The weights of each mouse are recorded at day 50; the average weight is determined for males and females separately. Individual mice are exchanged between treatment and vehicle cohorts to ensure initial body weights of each group are as similar as possible (usually within 0.3 g). All experiments were approved by the ALSTDI Institutional Animal Care and Use Committee. Most studies were performed blinded. Over the time period reported, studies were performed in three physically distinct animal facilities having a wide range of murine pathogen exposure (basic, clean, SPF). Analysis of survival times based on animal facility has shown no significant effect of the facility on the average survival time.

Each drug reported in Table I of the manuscript was dosed as follows: WHI-P131 (acquired by custom synthesis) was administered intraperitoneally (IP) in 10% DMSO/ 90% PBS at 12.5 mg/kg/day, five days per week, starting at age 60 days (d). Celebrex (donated by Pfizer) was formulated in chow at 1500 ppm by either Research Diets, Inc. (New Brunswick NJ) or Test Diet (Richmond IN) and administered *ad libitum* to result in a dose of approximately 300 mg/kg/day, 7 d/week, starting at age 50 d. Ceftriaxone (purchased from Sigma) was administered IP at 200 mg/kg/day, 7 d/week, starting at day 84 or 90 (in 2 separate studies). Creatine (Sigma), formulated in chow at 2% by Research Diets, Inc. resulting in a dose ~4000 mg/kg/day, started at age 50 d. Minocycline (Sigma), formulated in chow at 1 g/kg by Research Diets, Inc.

Table I. Validation of published studies. Retest (with optimized design and power) of compounds published as efficacious in SOD1^{G93A} studies.

	Published studies					Retest with ALSTDI study design[1]					
	n		Survival (days)			*n*			Survival (days)		
							Censored[2]				
Drug(s)	Control	Treated	Control	Treated	Lifespan extension	per group at start	Control	Treated	Control	Treated	Lifespan extension
WHI-P131 (27)	24	28	134	200	49.00%	40	34	30	133.9	136.4	1.86%
Celebrex 0.012% (28)	12	12	126	150.2	19.00%	90[3]	88	64	129.8	130.5	0.52%
Celebrex 1500 ppm in chow (21)	27	28	NA	NA	25.00%						
Creatine 2% (28)	12	12	126.1	151.4	19.80%	40	38	39	126	126.9	0.67%
Creatine 2% (29)[4]	6	7	143.7	169.3	17.80%						
Minocycline 50 mpk (33)[5]	7	7	130.3	150.9	15.80%	48*	47	44	135.7	134.9	−0.60%
Ceftriaxone, 200 mpk (36)[6]	20	20	122	135	10.70%	63	59	62	128	129.3	1.02%
Riluzole 0.1 mg/ml in water (~22mpk) (19)	8	8	134	148	10.4%	40	34	35	132.3	134.9	1.96%
Riluzole 24 mpk in chow (20)	11	10	127	140	10.2%						
Riluzole 44 mpk in chow (20)	11	11	127	139	9.4%						
PBA (Sodium phenylbutyrate) at 400 mpk (37)	20	20	125.7	153.2	21.9%	24	24	22	132.6	132.4	−1.75%
Thalidomide, 50 mpk (38)[7]	10	10	130	145	12%						
Thalidomide, 100 mpk (38)	10	12	130	151	16%	24	22	22	134.5	131.9	−1.93%
Thalidomide, 200 mpk						76	73	73	130.2	133	2.20%

[1] All results except WHI-P131 used the optimized study design (col. 10 from Figure 2A; all started at day 50); where $n > 24$, results shown are meta-analyses of independent studies performed with identical conditions; of the two independent WHI-P131 studies, one was optimal design, one was not litter-matched due to lack of litter data from The Jackson Laboratory. [2] 'Censored' refers to exclusion of mice that either died of non-ALS causes or had a low-copy number of the SOD1G93A transgene [3] Dose used in ALSTDI repeat was 1500 ppm Celebrex in chow. [4] Only highest published effect for creatine is shown; other reports had lesser effects, or did not measure survival (30–3232). [5] Other minocycline studies reporting lesser effects were carried out at different doses, routes, or age at start (31,34,35) [6] Greatest effect reported with start age of 42 days (10.7%) vs. 84 days (8.2%) (36); ALSTDI repeats were at day 50. [7] Both published thalidomide studies (38) started at age 30 days; ALSTDI repeats were at day 50.

resulting in a dose of ~400 mg/kg/day started at age 50 d. Riluzole (Sigma) was administered *ad libitum* in water at 0.9 μg/ml or 1.8 μg/ml to result in a dose of ~22 mg/kg/day or ~44 mg/kg/day, respectively, starting at age 60 d. Sodium phenylbutyrate ('PBA', donated as 'Buphenyl' by Ucyclyd Pharma, Scottsdale AZ) was administered IP in PBS at a dose of 400 mg/kg/day, starting at age 50 d. Thalidomide (provided by Celgene Corp., Summit NJ) was administered by oral gavage as a 40 mg/ml solution in 0.25% HPMC daily at a dose of 200 mg/kg/ day, 7 d/week, starting at age 50 d. If administered via food or water, consumption was tracked to allow exact determination of actual dose. For all studies, each mouse was weighed and neurological score (on a scale of 0 to 4, with 0 being normal and 4 being completely paralyzed) of both hind legs assessed daily. Although not discussed in this report, neurological score is necessary in conjunction with body weight in identifying non-ALS deaths: true ALS deaths follow a progression of decreased body weight and increased neurological score. If mice die before attaining a score of '2', it is highly unlikely that the death is of ALS. Date and cause of death are recorded for each mouse. To determine 'survival' reliably and humanely, an artificial endpoint is used, defined by the inability of a mouse to right itself in 30 s after being placed on its side. The moribund mice are scored as 'Died of ALS', and are euthanized by carbon dioxide followed by cervical dislocation.

Table S5 contains the study design and individual survival data for all studies performed by ALSTDI described in Table I.

Statistical analysis

Statistical analyses of efficacy studies were performed with the Cox proportional hazard (CoxPH) model using gender as a covariate and litter as a frailty term. Ties are handled via the Breslow method. Analysis of gender differences was again performed by CoxPH. Litter clustering was tested by performing a Cox regression analysis where litter was introduced as a frailty term. Results from statistical modeling were analyzed by one-way ANOVA. GraphPad Prism (v. 4.03, GraphPad Software, SanDiego CA) was used to draw survival curves.

'SimLIMS'

Simulations were generated from an integration of datasets stored in our LIMS (Laboratory

Information Management System) built on MSQL 2000, TSQL stored procedures and R scripts run on Intel processors in Windows 2000 servers.

Results

Identifying confounding variables affecting SOD1^{G93A} survival

We examined survival data of 5429 SOD1^{G93A} mice used as either control or treated animals in therapeutic studies at ALSTDI over a four-year period (Figure 1). The strain (B6SJL hybrids – see Methods) was chosen because it is used in the majority of ALS mouse efficacy studies. This distribution was collected in three physically distinct animal facilities having a wide range of environments (from basic to clean to SPF (specific pathogen-free), suggesting that animal care does not significantly influence the average survival time of these mice. Three distinct clusters were observed independent of any therapeutic effect (Figure 1). The largest group (87%) has an average lifespan of 134 ± 10 days (range 104–179 days) and represents mice that died of ALS-related causes. It is noteworthy that this survival is virtually identical to Gurney's original report of the survival times of 155 mice collected from seven different laboratories (3). Another group (10%) has an average lifespan of 100 ± 27 days (range 50–177 days) and contains mice excluded from study analyses because they died of non-ALS-related causes such as injection, self-mutilation, or acute infection. (A complete list of tracked non-ALS death causes can be found in Table S3.) A third group (2.6%) has an average survival of 195 ± 18 days (range 180–250 days). Mice in this cluster have reduced copies of the SOD1 transgene, as determined by qPCR or pedigree analysis (data not shown, see Methods). The direct correlation between copy number and survival is well established in this model (3,8–10).

Additionally, our data confirm a previously reported gender effect (3,8,11–15,39) that confers

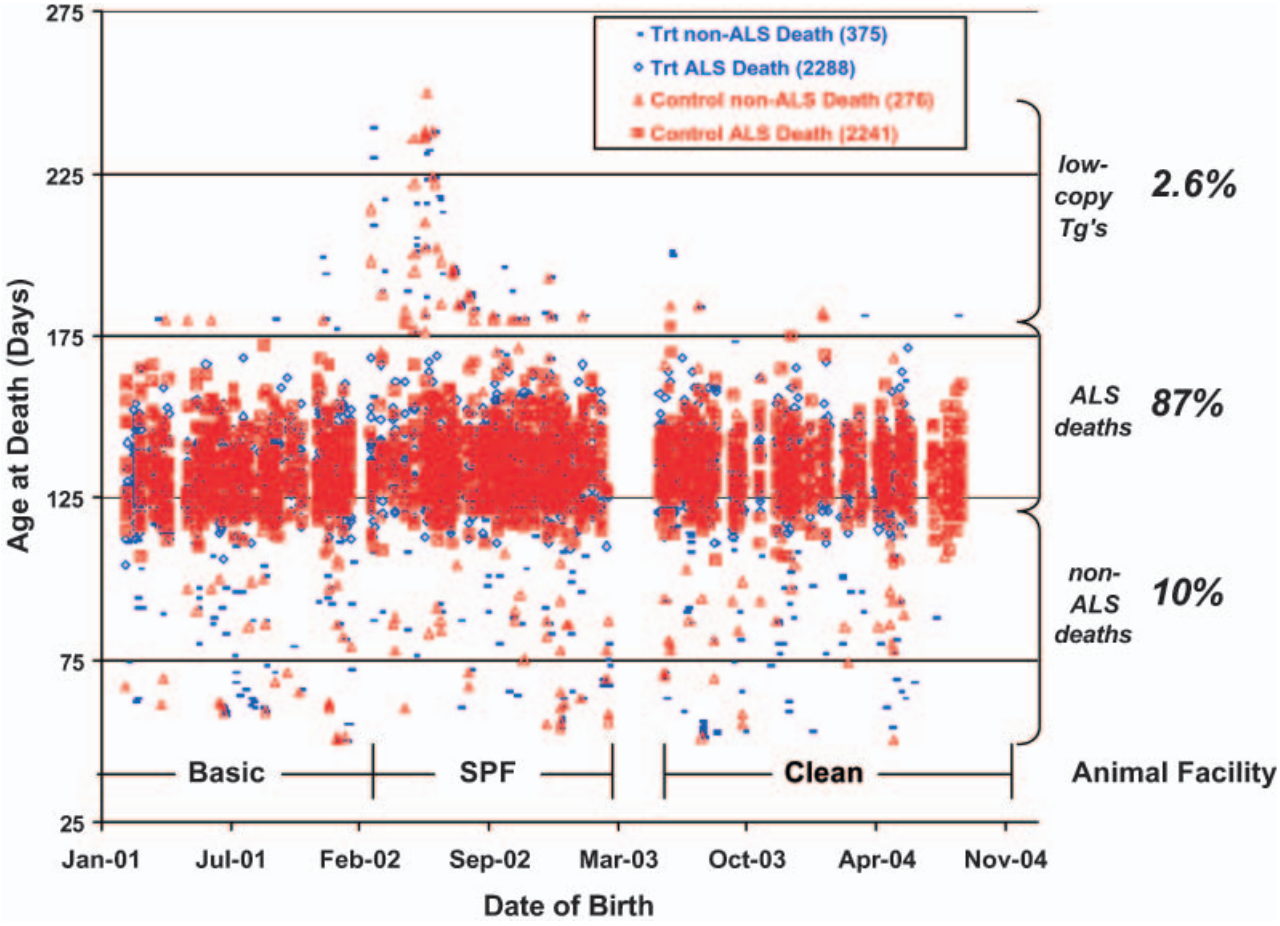

Figure 1. SOD1^{G93A} mouse colony survival over time. The horizontal axis is the date of birth for each animal while the vertical axis is the age at death. All mice were from The Jackson Laboratory and were used in either control (red) or treatment ('Trt', blue) groups in efficacy studies over the four-year period shown. Over this period the mice were housed in three physically distinct animal facilities labeled 'Basic', 'Clean' and 'SPF' (Specific Pathogen-Free). Animals were observed seven days a week and the cause of death was noted for every animal (ALS or Non-ALS). The data points between 200 and 250 days (from February 2002 to August 2002) were from long-lived mice that were eventually characterized as containing low-copy numbers of the human SOD1^{G93A} transgene. Since August 2002, our standard procedure is to check any mice alive at 180 days for transgene presence and copy number, resulting in few data points >200 days of age. See Methods for detailed explanation, Table S2 for full survival dataset. The brackets on the right show the major groupings by death reason.

a lifespan increase of four days (~3%) to females versus males (female survival=136 days; male survival=132 days ($p<0.001$); hazard ratio (HR)=1.41; 95% CI 1.27–1.56 (see Figure S1)). Overall, this survival analysis confirms that transgene copy number and gender are variables that affect survival and identifies exclusion criteria as a potentially large source of variability. The two types of exclusion criteria we establish are transgene copy number and death from non-ALS causes.

Because we observed that long-lived animals frequently turned out to be littermates (for example, animals surviving over 200 days in early 2002 in Figure 1), we examined the relationship between litter and survival in this set of 5429 mice. We employed the Cox proportional hazards statistical method which can take account of dependence between animals by introducing what are known as 'frailty' terms to detect clustering in survival times. An example of a frailty term is our observation that litter and sibling genotypes are independently predictive of the endpoint, and therefore siblings are more likely to have similar ages of onset and death than non-siblings. When litter was tested in this manner it was found to be highly statistically significant for males ($p<0.006$) and females ($p<0.001$) (see Figure S1 and Table S2). While this correlation also held across genders ($p<0.001$), females still tend to live longer. Therefore, consistent with previous reports (39) we propose that litter is an additional factor affecting survival along with gender, transgene copy number, and exclusion criteria.

Impact of confounding variables on SOD1^{G93A} efficacy studies

To determine how these four variables would affect study results if left uncontrolled, we used our cumulative dataset to 'simulate' results of studies performed with various study designs. Since treatment effects in SOD1^{G93A} mice are routinely reported as an increase in the treatment group's mean survival (in days) as a percent of control, we used this objective measure to quantify how the four variables could affect survival analyses. If the variables were causing noise in the distribution of survival times, such noise should become evident when survival times are sampled and averaged independent of any pharmacological treatment. We created a computer model (called SimLIMS) that samples the actual survival data recorded in our database (LIMS). Actual individual survival times were taken from 2241 animals assigned exclusively to control groups in efficacy studies (see Figure 1). This database of survival times reflects the actual range of lifespans in this SOD1^{G93A} mouse colony. SimLIMS therefore samples (without replacement) the age at death (in days) of animals, based on any chosen set of study design restrictions and distributes them into groups arbitrarily designated 'treatment' and 'control'. The percent difference in mean survival times is then calculated between treatment and control. Because effects in the 0–5% range are typically not reported in the literature, we disregarded effects between 5% and −5%. Effects outside this range were designated 'apparent effects' and tabulated. SimLIMS repeats this operation iteratively 1000 times per 'study' using the criteria of the study design restrictions. Values reported here are the averages of five such separate studies (5×1000 iterations $\pm$ SEM, designated 'Frequency of Apparent Effects' in Figure 2, panel A). Apparent effects are determined without regard to statistical significance because they are an independent descriptive measure of the noise in the distribution of survival times around the mean.

For example, to determine the effect of powering a study at 10 mice in a cohort (cohort size being the only defined variable), SimLIMS would be programmed to randomly select 10 mice from the database and assign them to the 'treatment group', and select 10 random mice and assign them to the 'control group'. Average survival times would be calculated for each group and the percent difference between the treatment and control groups calculated. This process is repeated iteratively to make possible a statistical analysis of the effect of powering a study with 10 mice per group. It is important to note that none of the mice in the database has received treatment, therefore any statistical differences between the 'treated' and 'control' cohorts must be derived from variables such as gender, transgene copy, censoring criteria, etc. By introducing each of these variables we are able to calculate both the frequency and magnitude of apparent effects.

The frequency of these apparent effects (graphed in Figure 2, panel A) is essentially the percentage of SOD1^{G93A} studies that would have generated a result that could be interpreted as an apparent effect of greater than 5% between the 'treatment' and 'control'; however, as there is no treatment actually given to the animals used in these simulated studies, the variation can be attributed solely to the study design variables altered in the simulations. Clearly, the results indicate that including non-ALS deaths (i.e. not applying exclusion criteria) is the largest potential source of noise in the SOD1^{G93A} mouse, followed by inclusion of low-copy transgenics (Figure 2, panel A, columns 1–3). When no exclusion criteria are enforced, an apparent effect would be seen in 58% of studies (Figure 2, panel A, column 1). Removing low-copy transgenic animals drops the apparent effect rate to 54% (Figure 2, panel A, column 2), while enforcing both exclusion criteria and removing low-copy transgenic animals reduces the apparent effect rate to 36% (Figure 2, panel A, column 3). Litter clustering is the next

largest source of noise; addition of litter-matching to the study design further decreases the apparent effect rate to 30% (Figure 2, panel A, column 10). Surprisingly, gender distribution is not a significant contributor to noise in the distribution of survival means despite females living on average four days longer than males (Figure 2, panel A, compare columns 3–6). However, the litter effect appears to be stronger within females (Figure 2, panel A, compare column 8 to column 7), making a female-only study the least prone to noise. For this demonstration, a smaller sample size (n=4) was

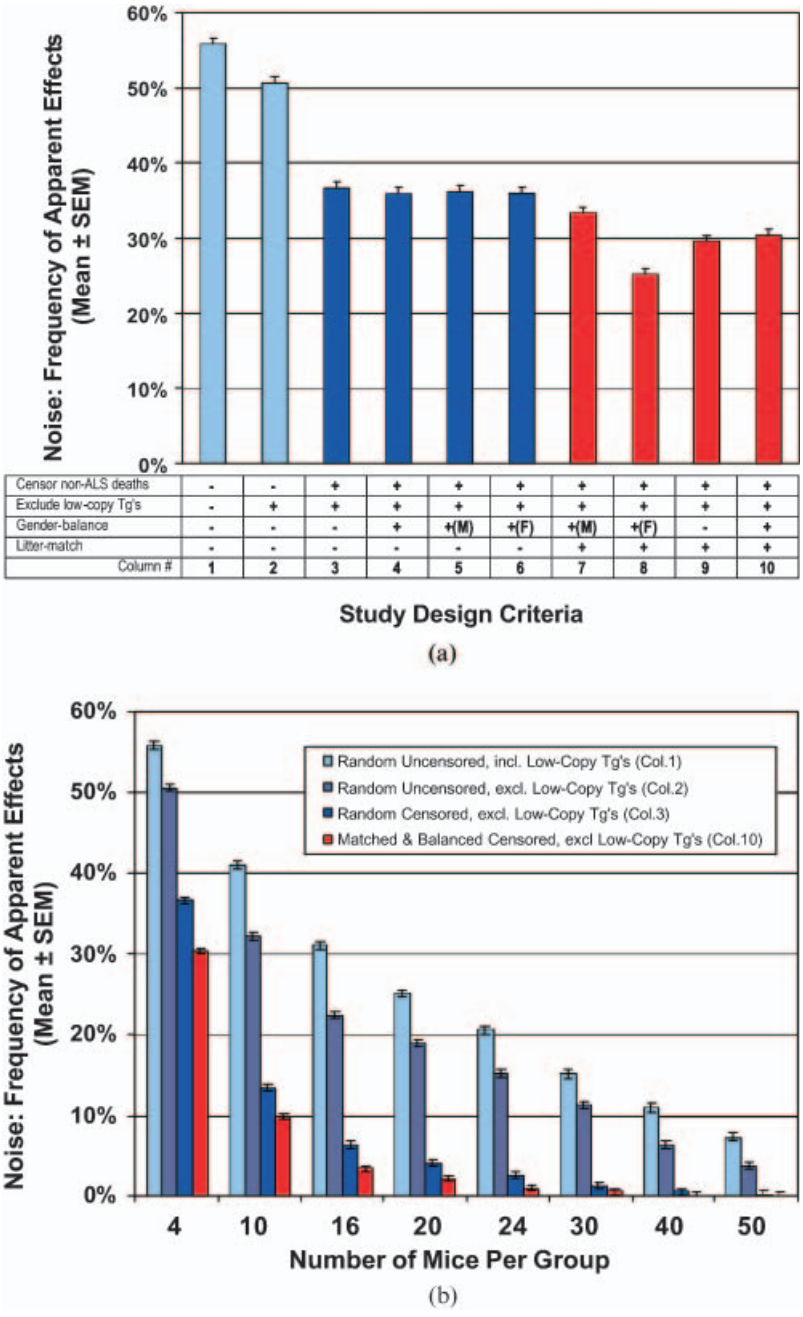

Figure 2. Quantitation of inherent noise in the SOD1^{G93A} mouse model. Panel A) Cumulative decrease in noise ('apparent effects') by successively more stringent criteria. Computer simulations ('SimLIMS') were performed using actual survival data from 2579 untreated SOD1^{G93A} mice. Mice (n=4) were 'assigned' to either control or treatment groups, survival ages averaged, and the % difference in means calculated. Any difference $>5\%$ (or $<-5\%$) was considered to be an 'apparent' effect. Five sets of 1000 iterations each (to allow SEM determination) were performed to give each data point, reflecting the percentage of studies ('conducted' by each of 10 designs) in which an apparent effect would be observed. The 10 study designs (chosen to reflect likely or actual published study designs) encompass the four variables of gender, littermates, and censoring of either low-copy transgenics and/or non-ALS deaths. For gender-balancing, '+(M) or (F)' refers to studies that are balanced by using only a single gender, either male of female. This figure shows only n=4 animals per group to best illustrate the effects of the confounding variables. Panel B) Apparent effects vs. cohort size (n per group) for various study designs and analysis criteria. With larger sample sizes the noise is reduced, although the contributions of each variable are similar as detailed in Panel A. Column numbers refer to the conditions described in Panel A. One-way ANOVA indicates high significance of difference between each column in a group ($p<0.0001$).

used to illustrate the effects of the confounding variables; larger sample sizes reveal lower noise with similar relative contributions of variables to that noise (Figure 2, panel B and data not shown).

While there are publications using $n=4$ animals per group (16–18), the majority (~60%) of published SOD1^{G93A} efficacy studies we reviewed employed 5–10 animals per group. To determine the effect of group size on SOD1^{G93A} studies, we next altered the group size in SimLIMS to reflect numbers commonly used. Predictably, every variable tested is reduced by increasing the number of animals per group (Figure 2, panel B). However, adding animals cannot overcome all noise. If low-copy transgenics are included and exclusion criteria are not enforced, apparent effects still occur at unacceptable levels (>5%) even with as many as 50 animals per group. In contrast, imposing both of these criteria can drive noise to a virtual zero when $n=24$ or more. Based on this analysis, we established a standard efficacy study design of $n=24$ animals per cohort, employing same-gender litter matching. (This low noise level could also be achieved by using $n=30$ per cohort without litter matching; however this design offers the practical advantage of requiring ~6 fewer animals per group.) We also balanced studies at 50% female, 50% male in case a drug shows a gender-specific therapeutic benefit or pharmacological effect.

Comparison of potential noise in SOD1^{G93A} mouse model published efficacy studies

Having quantified the noise caused by each of the confounding variables, we asked how that inherent noise compared to the results of both our own studies and publications of efficacy in the SOD1^{G93A} mouse. Thus, in Figure 3 we looked at the magnitude of the apparent (and reported) effects instead of their frequency. Most published studies do not discuss any type of exclusion criteria, gender, or litter issues, and usually rely on standard PCR genotyping that cannot discriminate low- from high-copy transgenics. Therefore, to best evaluate published studies, we reasoned that the most controlled study design would likely employ random assignments of mice with exclusion of both non-ALS deaths and low-copy number transgenics (correlating with column 3 in Figure 2, panel A). The line in Figure 3 labeled 'Random Censored Noise Level' corresponds to the highest level of noise found by SimLIMS for this study design, at varied group size n. Similarly, we assumed an upper level of noise would result from a random study design lacking exclusion criteria of both non-ALS deaths and low-copy number transgenics (as described in column 1 of Figure 2, panel A). The line in Figure 3 labeled 'Random Uncensored Noise Level' corresponds to the highest level of noise expected from this study design, as calculated by SimLIMS, for each group size n.

The majority (69 of 81) of published effects in the SOD1^{G93A} mouse model fall within the noise inherent to the model, as demonstrated by the box and whiskers plots in Figure 3. Therefore, such differences reported between the control and treatment groups could be explained by factors other than the treatment being investigated. The results of 12 studies are above all calculated noise lines and 24 other studies report results that fall in the grey area between the 'Random Censored' and 'Random Uncensored' lines.

Retest of select compounds with optimized SOD1^{G93A} study design

We selected WHI-P131, Celebrex and several of the therapeutics in the grey area to test with our optimized study design described above (see Figure 3, inset). We were particularly interested in testing compounds that had efficacy reported by multiple groups, as well as compounds that had gone to clinical trial (such as Celebrex, creatine, and minocycline) or were headed to clinical trial (such as ceftriaxone, thalidomide, or sodium phenylbutyrate). Despite the reported survival extension in SOD1^{G93A} mice (~5–10% (15,19,20)) falling within the noise range for the relevant study design, we also tested riluzole since it is the only drug approved by the FDA for ALS.

As seen in Table I, none of the retested compounds showed a significant effect on survival of the SOD1^{G93A} mouse when run under study design conditions that minimize background noise. The retested studies (including some studies run at even higher power (n) in an attempt to identify a positive effect), often with double or more the number of mice used by all laboratories combined, were calculated to yield 0–3% apparent effects. Furthermore, power analysis indicated that these repeat studies had greater than 90% power to detect the published effect (see Methods and Table S4).

Discussion

The SOD1^{G93A} mouse model relies on tightly controlled breeding, quantitative genotyping, animal shipping, husbandry, pharmacology, handling, and evaluation, all of which are subject to their own opportunities for errors in standardization. This report quantitatively characterizes the effect on survival studies of several critical variables – both those previously recognized (transgene copy number and gender) and some identified by our analysis (exclusion criteria and litter). Analysis of published studies in this model, in the context of this quantified 'noise floor' revealed that 'efficacy' reported in the majority of published studies (69 of

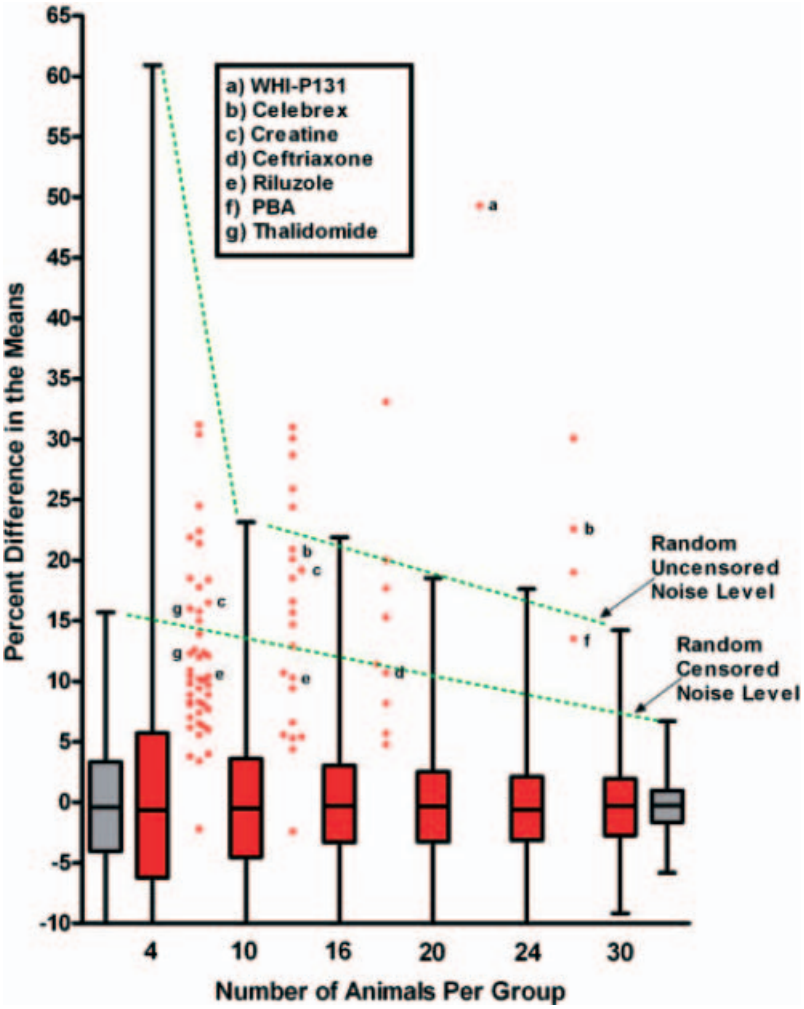

Figure 3. Comparison of results from published studies to calculated noise range. Results from 81 efficacy studies in high-copy SOD1^{G93A} mice from 50 publications (see Table S1 for complete list and selection criteria) are plotted (in red points) along the x-axis according to the number of mice used per group and along the y-axis according to the percent effect (% increase in lifespan) reported for that particular study. The whiskers on the box plots show the upper range of apparent effects generated by SimLIMS for two types of study design: the upper dashed line (labeled 'Random Uncensored Noise Level') marks the upper range of apparent effects across the various group sizes (plotted in red box and whiskers) if performed under the least stringent set of design criteria – random assignment of mice and lacking both non-ALS death censoring and exclusion of low-copy number transgenics (as described in column 1 of Figure 2, panel A). The lower dashed line (labeled 'Random Censored Noise Level') marks the lower range of apparent effects across the various group sizes if performed under a more stringent set of design criteria – random assignment of mice but censoring non-ALS deaths and low-copy transgenics (as described in column 3 of Figure 2, panel A). For simplicity, only the highest and lowest group size (n=4 and n=30) is shown (in grey). While the whiskers go equivalent distances in both the positive and negative direction, the negative whiskers have been truncated to focus on the area containing most published results. Points for drugs retested and detailed in Table I are labeled with letters a–g.

81) using this SOD1^{G93A} mouse model could actually be attributed to random distribution of survival times under the given study designs. Using our improved study design (which controls for the identified variables and offers higher power to detect), we retested eight compounds including riluzole and found no significant effect. The implications and limitations of these findings are discussed.

The results reported here also conclusively support the claim that there is indeed a distinct average lifespan (and standard deviation) for the SOD1^{G93A} mouse. Ten years ago, Gurney described the SOD1^{G93A} model as robust given that seven different laboratories reported similar survival times from a total of 155 mice (ranging from 132$\pm$12 to 138$\pm$16 days; mean$\pm$SD) (3). Aligning completely with these times, we observed an average lifespan of 134$\pm$10 days from 4529 mice dying of ALS (red line in Figure 4). The fact that this average lifespan is observed across all vehicle and route control, as well as treated, groups of mice (see Figure 1 and Table S2) further attests to the robustness of the model. However, several reports show treated animals within this range while the control animals have significantly shorter survival times ((21–23); results from (23) plotted in Figure 4)). This comparison raises questions about potential differences in SOD1^{G93A} mouse colony environment, handling, and/or genetics that could yield such a significant deviation. For example, one simple explanation might be that the groups were not run in parallel but rather were run at different times and/

or places, as when 'historical controls' are used repeatedly as a reference (possibly to reduce the number of mice needed). Another potential notion is that the mice could suffer from some non-ALS related illness and it is this illness rather than the ALS disease that is alleviated in such treatment groups. Consistent with this hypothesis, it is of note that several of the compounds initially reported as efficacious in SOD1^{G93A} mice but not retested here are broad-spectrum antibiotics and general anti-inflammatory agents.

Our failure to replicate the effect of riluzole in this mouse model at $n>50$ animals per cohort raises obvious questions about both the validity of our methods and studies and of published studies with riluzole, and the utility of riluzole as a positive control compound. Indeed, the effect of riluzole in humans is marginal (~2 months for an average age of ~50 years) and requires many thousands of patients to detect (6). If the true effect of riluzole was correspondingly modest in the SOD1^{G93A} mouse (say, ~1%), even studies of $n>50$ would be insufficiently powered to detect it. For example, our power calculations suggest that the gender effect of 3–4 days (~3%) requires >200 animals per cohort to be detected consistently. The original riluzole studies (19,20) reported survival effects ~8% and controlled for exclusion criteria, transgene number, and gender but did not control for litter and used very few animals (n ~10 per cohort). Our report here indicates that these conditions could potentially yield apparent effects of at least 5% to as high as 30% and only 30–60% power to detect a 5–10% effect. Additionally, the Student's t-test used in these published reports is not appropriate to survival studies in general and cannot address the litter clustering inherent in the SOD1^{G93A} mouse. Even with this invalid statistical method, the best reported p-values were 0.037 and 0.049 (no confidence intervals reported (19,20)), consistent with a nominal or no significant therapeutic effect; hence, the efficacy claimed in these reports could simply be a consequence of common misinterpretation or misuse of statistical analysis, as generally acknowledged recently (24,25).

Furthermore, one of these studies (19) describes all mice being treated with both ivermectin and piperazine hydrochloride to treat pinworm infection, with some mice dying after these treatments; yet ivermectin itself was recently reported as effective in the SOD1^{G93A} mouse, extending survival almost 10% (26), a similar level as reported for riluzole. This type of palliative treatment would not be allowed in our system and the infected animals would be excluded from the study. In contrast, when we tested riluzole, we did so at two doses, each with $n>50$. Our power analysis indicates that 40 or more animals per cohort would yield a >90% chance of detecting the effect (8% extension in survival) previously described (19,20). Thus, we believe the most likely explanation for the lack of riluzole efficacy in our report is that the effect published in the original riluzole studies must be attributed to type I error (false positives) or other unknown variables (such as the increased chance of a statistical positive when employing three-arm studies). In addition, the effect of riluzole may simply be too small to detect at the given power. Riluzole is an example of how bias toward type I error is propagated when negative results are not routinely reported in the literature. This remains a widespread phenomenon in pre-clinical studies and could be greatly ameliorated if investigators would report negative, or even toxic, effects in well-designed pre-clinical studies.

The above example of the use of the Student's t-test for the published riluzole (and other SOD1^{G93A}) studies highlights the point that while statistics are

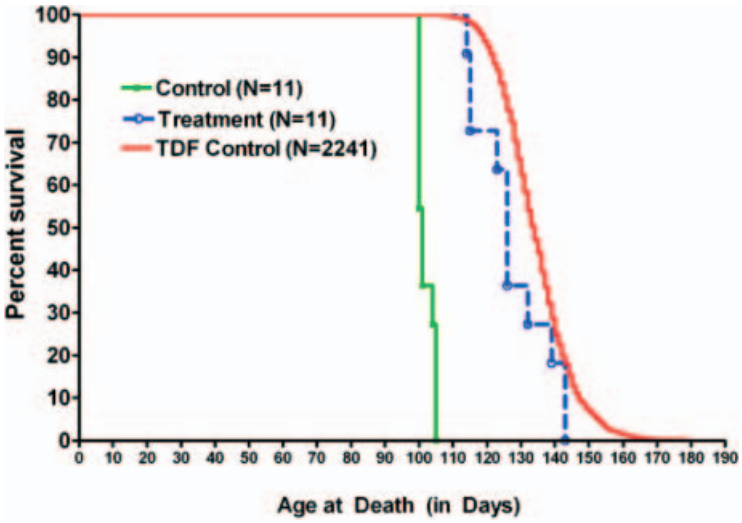

Figure 4. Survival analysis. Control and treated SOD1^{G93A} mice from one publication compared to all of our 2241 control animals (acquired over four years – data from Table S2) that died of ALS.

important, they ultimately do not offer protection from inappropriate control of the variables in a system. Each of the variables we have identified, if uncontrolled, violates certain assumptions essential to the valid application of the statistical methods employed, i.e. when the statistical method does not support the inherent structure of the data, any interpretation based on that statistical analysis is invalid. For instance, the correlation in survival times of littermates violates the assumption of independence while the existence of non-ALS deaths or low-copy transgenics violates the assumption that all animals in the study come from the same survival distribution. Both assumptions are critical to the validity of any resulting *p*-value calculations and the subsequent confidence they impart to a given study. Indeed, there may be sources of noise in addition to the four variables described here. For example, 20 slightly different death criteria (humane endpoints) have been reported by 27 different laboratories. Because our survival times are collected under a uniform standard, there is no way for us to readily gauge the impact of this variable. The lack of appropriate blinding of some studies is another variable not testable by our analysis, as we routinely perform all studies blinded. It is also important to emphasize that this report deals only with the most commonly used strain for ALS preclinical studies, which is a mixed hybrid (B6SJL) strain. Although an inbred strain might be expected to have lower noise (due to less or no litter effect, since all animals would be genetically identical), the variables inherent in any model (or strain) need to be identified and quantified before the model can be appropriately used and interpreted.

The lack of significant survival benefit in this $SOD1^{G93A}$ mouse model from the only FDA-approved drug for ALS, riluzole, makes it difficult to ascertain the predictive value of the model in the human population. If one ignored riluzole because its modest human effect makes the results difficult to correlate, then the lack of efficacy in the mouse with compounds such as celecoxib, creatine, and minocycline and their subsequent failure in human clinical trials could indeed be interpreted as the mouse being predictive of the human disease. However, one must also consider the possibility that the mouse suffers from a phenotype that is so aggressive and so overdriven by its 23 copies of the transgene that no pharmacological intervention outside of the direct inhibition of SOD1 will ever affect survival. If such is the case, then other molecular endpoints must be considered to glean potential therapeutic information from the $SOD1^{G93A}$ model. However, each alternative outcome measure (such as motor neuron counts, molecular biomarkers, body weight, neurological scores, rotarod performance, gait analysis, wheel distance, etc.) must be fully characterized and quantified in the model to appropriately analyze and interpret the results.

Conclusion

Two conclusions can be drawn from the data described in this report. First, the high noise floor of the model and the failure of the selected studies to replicate support the conclusion that the bulk of published studies using the $SOD1^{G93A}$ mouse model may unfortunately be measurements of biological variability due to inappropriate study design. Secondly, this inherent noise can be substantially eliminated by using the matched and balanced study design, applying appropriate and uniform exclusion criteria, and applying properly quantitative genotyping techniques. The most confounding variables for use of the $SOD1^{G93A}$ mouse model, both in our own use and likely in other published studies, are, in order: the inevitable occurrence of non-ALS deaths, the incidence of low-copy transgenics (due to recombination events in the array as previously described (3)), the genetic background or epigenetic influences causing littermate clustering, and a slight gender effect.

Specifically, our results and analyses generate the following recommendations for improved preclinical study design for therapeutic testing of survival in the $SOD1^{G93A}$ mouse: 1) Each cohort should have at least 24 litter-matched gender-balanced mice. 2) Each study should be blinded to both animal technicians and investigators. A single uniform endpoint criterion should be employed (the most common being if the mouse cannot right itself in 30 s after being placed on its side). The importance of running double-blind studies with uniform endpoints cannot be overstated. 3) Non-ALS deaths must be tracked and excluded from final analysis. Likewise, we recommend that the sibling be excluded as well and that this exclusion be reported. (As seen in Figure 2, panel B, excluding an unaffected sibling (when the matched littermate is excluded) is critical below $n=24$). 4) Initial quantitative analysis of transgene copy number prior to assigning mice to a study could alleviate confounding issues of interference from low-copy animals. If this is not feasible, long-lived animals due to a low number of transgene copies must be identified and excluded from $SOD1^{G93A}$ study data since they are actually different from the model intended to be tested. 5) For statistical analysis, because the $SOD1^{G93A}$ model has multiple variables, the Cox proportional hazards model is most appropriate since it can handle gender as a covariate, litter as a frailty term, as well as utilize censored data. Additional variables could also be incorporated in this analysis, with the important caveat that any variable that is used as a matching variable in the design should also be included as a covariate in the final analysis. Additionally, other log-rank tests cannot properly deal with litter clustering, and the commonly used *t*-test/ANOVA analyses are not suitable for survival statistics.

In conclusion, future interpretation and application of $SOD1^{G93A}$ survival studies requires that the criteria presented here be addressed. We consider the recommendations detailed in this report as a starting point for standardizing the field's use of murine models of ALS. Furthermore, although the current study design is sufficient to control for the variables described here, other criteria may be needed to address other important variables such as age at study start (presymptomatic vs. postsymptomatic). Additionally, including all raw data in published reports would serve to clarify alternative methodologies between studies or analyses and make possible the potential stratification of mice that progress differently thereby providing potentially valuable information. Finally, the use of the above pre-clinical study design will ease the confusion faced by patients, researchers and clinicians in their efforts to identify the most promising clinical trial therapeutics for this devastating disease.

Acknowledgments

We thank Samantha Crocker, Andy Moreno and all of our many animal technicians for their thorough and dedicated technical assistance, Howard Cabral for his help with the original statistical analysis, Jerry DeZutter and John McCarty for critical reading of the manuscript, Harvey Lodish for guidance on this work, and the entire ALS TDI community for their support and interest. We are especially grateful to the numerous people, families, and foundations that funded this work although they are too numerous to list individually here. We dedicate this paper to Vanna Forrester and Stephen Heywood for inspiring this effort. This work was funded entirely by donations to ALS TDI from families affected by ALS and from individual philanthropists. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Rosen DR, Siddique T, Patterson D, Figlewicz DA, Sapp P, Hentati A, et al. Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. Nature. 1993;362:59–62.
2. Tu PH, Raju P, Robinson KA, Gurney ME, Trojanowski JQ, Lee VM. Transgenic mice carrying a human mutant superoxide dismutase transgene develop neuronal cytoskeletal pathology resembling human amyotrophic lateral sclerosis lesions. Proc Natl Acad Sci U S A. 1996;93:3155–60.
3. Gurney ME. The use of transgenic mouse models of amyotrophic lateral sclerosis in preclinical drug studies. J Neurol Sci. 1997;152(Suppl 1):S67–73.
4. Nirmalananthan N, Greensmith L. Amyotrophic lateral sclerosis: recent advances and future therapies. Curr Opin Neurol. 2005;18:712–9.
5. Rothstein JD. Preclinical studies: how much can we rely on? Amyotroph Lateral Scler Other Motor Neuron Disord. 2004;5(Suppl 1):22–5.
6. Miller RG, Mitchell JD, Lyon M, Moore DH. Riluzole for amyotrophic lateral sclerosis (ALS)/motor neuron disease (MND). John Wiley & Sons, Ltd. Available: http://www.cochrane.org/reviews/en/ab001447.html.
7. Gurney ME, Pu H, Chiu AY, Dal Canto MC, Polchow CY, Alexander DD, et al. Motor neuron degeneration in mice that express a human Cu/Zn superoxide dismutase mutation. Science. 1994;264:1772–5.
8. Alexander GM, Erwin KL, Byers N, Deitch JS, Augelli BJ, Blankenhorn EP, Heiman-Patterson TD. Effect of transgene copy number on survival in the G93A SOD1 transgenic mouse model of ALS. Brain Res Mol Brain Res. 2004;130:7–15.
9. Dal Canto MC, Gurney ME. A low expressor line of transgenic mice carrying a mutant human Cu/Zn superoxide dismutase (SOD1) gene develops pathological changes that most closely resemble those in human amyotrophic lateral sclerosis. Acta Neuropathol (Berl). 1997;93:537–50.
10. Chiu AY, Zhai P, Dal Canto MC, Peters TM, Kwon YW, Prattis SM, Gurney ME. Age-dependent penetrance of disease in a transgenic mouse model of familial amyotrophic lateral sclerosis. Mol Cell Neurosci. 1995;6:349–62.
11. Trieu VN, Uckun FM. Genistein is neuroprotective in murine models of familial amyotrophic lateral sclerosis and stroke. Biochem Biophys Res Commun. 1999;258:685–8.
12. Cudkowicz ME, Pastusza KA, Sapp PC, Mathews RK, Leahy J, Pasinelli P, et al. Survival in transgenic ALS mice does not vary with CNS glutathione peroxidase activity. Neurology. 2002;59:729–34.
13. Heiman-Patterson TD, Deitch JS, Blankenhorn EP, Erwin KL, Perreault MJ, Alexander BK, et al. Background and gender effects on survival in the TgN(SOD1-G93A)1Gur mouse model of ALS. J Neurol Sci. 2005;236:1–7.
14. Miana-Mena FJ, Munoz MJ, Yague G, Mendez M, Moreno M, Ciriza J, et al. Optimal methods to characterize the G93A mouse model of ALS. Amyotroph Lateral Scler Other Motor Neuron Disord. 2005;6:55–62.
15. Snow RJ, Turnbull J, da SS, Jiang F, Tarnopolsky MA. Creatine supplementation and riluzole treatment provide similar beneficial effects in copper, zinc superoxide dismutase (G93A) transgenic mice. Neuroscience. 2003;119:661–7.
16. Ende N, Weinstein F, Chen R, Ende M. Human umbilical cord blood effect on SOD mice (amyotrophic lateral sclerosis). Life Sci. 2000;67:53–9.
17. Chen R, Ende N. The potential for the use of mononuclear cells from human umbilical cord blood in the treatment of amyotrophic lateral sclerosis in SOD1 mice. J Med. 2000;31:21–30.
18. Bruce KM, Narayan K, Kong HC, Larmour I, Lopes EC, Turner BJ, et al. Chemotherapy delays progression of motor neuron disease in the SOD1 G93A transgenic mouse. Chemotherapy. 2004;50:138–42.
19. Gurney ME, Cutting FB, Zhai P, Doble A, Taylor CP, Andrus PK, Hall ED. Benefit of vitamin E, riluzole, and gabapentin in a transgenic model of familial amyotrophic lateral sclerosis. Ann Neurol. 1996;39:147–57.
20. Gurney ME, Fleck TJ, Himes CS, Hall ED. Riluzole preserves motor function in a transgenic model of familial amyotrophic lateral sclerosis. Neurology. 1998;50:62–6.
21. Drachman DB, Frank K, Dykes-Hoberg M, Teismann P, Almer G, Przedborski S, Rothstein JD. Cyclooxygenase 2 inhibition protects motor neurons and prolongs survival in a transgenic mouse model of ALS. Ann Neurol. 2002;52:771–8.
22. Acsadi G, Anguelov RA, Yang H, Toth G, Thomas R, Jani A, et al. Increased survival and function of SOD1 mice after glial cell-derived neurotrophic factor gene therapy. Hum Gene Ther. 2002;13:1047–59.
23. Crow JP, Calingasan NY, Chen J, Hill JL, Beal MF. Manganese porphyrin given at symptom onset markedly extends survival of ALS mice. Ann Neurol. 2005;58:258–65.

24. Ioannidis JP. Why most published research findings are false. PloS. 2005;2:124.
25. Sterne JA, Davey SG. Sifting the evidence – what's wrong with significance tests? Br Med J. 2001;322:226–31.
26. Andries M, van DP, Robberecht W, van den BL. Ivermectin inhibits AMPA receptor-mediated excitotoxicity in cultured motor neurons and extends the lifespan of a transgenic mouse model of amyotrophic lateral sclerosis. Neurobiol Dis. 2006.
27. Trieu VN, Liu R, Liu XP, Uckun FM. A specific inhibitor of janus kinase-3 increases survival in a transgenic mouse model of amyotrophic lateral sclerosis. Biochem Biophys Res Commun. 2000;267:22–5.
28. Klivenyi P, Kiaei M, Gardian G, Calingasan NY, Beal MF. Additive neuroprotective effects of creatine and cyclooxygenase 2 inhibitors in a transgenic mouse model of amyotrophic lateral sclerosis. J Neurochem. 2004;88:576–82.
29. Klivenyi P, Ferrante RJ, Matthews RT, Bogdanov MB, Klein AM, Andreassen OA, et al. Neuroprotective effects of creatine in a transgenic animal model of amyotrophic lateral sclerosis. Nat Med. 1999;5:347–50.
30. Andreassen OA, Jenkins BG, Dedeoglu A, Ferrante KL, Bogdanov MB, Kaddurah-Daouk R, Beal MF. Increases in cortical glutamate concentrations in transgenic amyotrophic lateral sclerosis mice are attenuated by creatine supplementation. J Neurochem. 2001;77:383–90.
31. Zhang W, Narayanan M, Friedlander RM. Additive neuroprotective effects of minocycline with creatine in a mouse model of ALS. Ann Neurol. 2003;53:267–70.
32. Derave W, van den BL, Lemmens G, Eijnde BO, Robberecht W, Hespel P. Skeletal muscle properties in a transgenic mouse model for amyotrophic lateral sclerosis: effects of creatine treatment. Neurobiol Dis. 2003;13:264–72.
33. Van den Bosch L, Tilkin P, Lemmens G, Robberecht W. Minocycline delays disease onset and mortality in a transgenic model of ALS. Neuroreport. 2002;13:1067–70.
34. Zhu S, Stavrovskaya IG, Drozda M, Kim BY, Ona V, Li M, et al. Minocycline inhibits cytochrome c release and delays progression of amyotrophic lateral sclerosis in mice. Nature. 2002;417:74–8.
35. Kriz J, Nguyen MD, Julien JP. Minocycline slows disease progression in a mouse model of amyotrophic lateral sclerosis. Neurobiol Dis. 2002;10:268–78.
36. Rothstein JD, Patel S, Regan MR, Haenggeli C, Huang YH, Bergles DE, et al. Beta-lactam antibiotics offer neuroprotection by increasing glutamate transporter expression. Nature. 2005;433:73–7.
37. Ryu H, Smith K, Camelo SI, Carreras I, Lee J, Iglesias AH, et al. Sodium phenylbutyrate prolongs survival and regulates expression of anti-apoptotic genes in transgenic amyotrophic lateral sclerosis mice. J Neurochem. 2005;93:1087–98.
38. Kiaei M, Petri S, Kipiani K, Gardian G, Choi DK, Chen J, et al. Thalidomide and lenalidomide extend survival in a transgenic mouse model of amyotrophic lateral sclerosis. J Neurosci. 2006;26:2467–73.
39. Ramesh TM, Buradagunta S, Thompson K, et al. Analysis of critical parameters for preclinical drug screening in the SOD1 G93A mouse model for amyotrophic lateral sclerosis. Abstract at the 13th International Symposium on ALS/MND; November, 2002; Melbourne.