

PEER REVIEW HISTORY

BMJ Open Science publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (<http://openscience.bmj.com/pages/wp-content/uploads/sites/62/2018/04/BMJ-Open-Science-Reviewer-Score-Sheet.pdf>) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Methodological standards, quality of reporting, and regulatory compliance in animal research on amyotrophic lateral sclerosis: a systematic review
AUTHORS	Joana G Fernandes, Nuno H Franco, Andrew J Grierson, Jan Hultgren, Andrew JW Furley, I Anna S Olsson (Corresponding Author)

VERSION 1 - REVIEW

REVIEWER 1	<i>Malcolm Macleod</i> <i>University of Edinburgh</i>
REVIEW RETURNED	<i>16/09/2018</i>

GENERAL COMMENTS	<p>Thanks for the opportunity to review this interesting work. I have some comments. it may be that many of these things had been done but didn't make it to the manuscript. 1. The first search was done in 2013. PRISMA Item 5 asks to indicate whether a protocol exists, and your response is "N/A". On the basis that a protocol does not exist, you have great researcher degrees of freedom in what you did, and this is a major weakness which should be addressed. For instance, were your outcome measures articulated before data collection began, or not? 2. It's not clear whether citation screening, assessment of preclinical v exploratory, and annotation were performed by one person or two; and if annotation for quality was performed blinded to year of publication if this were possible. 3. Choice of web of science to search rather than pubmed is unusual for preclinical SRs - doesn't make it wrong, but might be worth mentioning why you took that approach? 4. There is a problem with using a checklist (present/ absent) and then weighting items and then summing them to give a measure of quality or reporting which you then consider at a ratio scale (4 is twice as good as 2) and plug the weighted sum as the dependent variable into a regression equation. In my view there is not sufficient evidence to assign weights to these items, and a better approach is the number of checklist items scored. 5. You might like to discuss the possible weakness of running several statistical tests each with a p threshold of 0.05. This can be justified if the questions are independent, but in this</p>
-------------------------	---

	<p>case I am not sure that that is the case. 6. Your MSR and RSR indexes are potentially interesting but unvalidated. I am concerned that - in expressing these with capitalisation and a name, rather than a list of items - you consider these now to be validated tools. You have not shown their validity, and so it would be preferable simply to list these items without claiming that you have developed an "index". What would be useful (and you have done this, I think) is to articulate a checklist which allows determination of compliance with the ALS reporting guidelines. 7. You claim this is the first demonstration of the impact of field specific reporting guidelines, but Minnerup (PMID: 26658439) and Ramirez (PMID: 28373349) have both looked at the change in quality in reporting of in vivo stroke studies, and we looked (Bahor PMID: 28373349) at changing quality in MCAO and lacunar stroke, and at the change in a specific drug in stroke (McCann, PMID: 27526101) ... so I don't think that claim is valid. 8. It's not clear whether you would include ex vivo SOD mice (eg tissue slice, myoculture) - I suspect not, but absent a clear articulation of inclusion and exclusion criteria it is difficult to be sure 9. References 33 and 79 appear to be duplicates. 10. There are a few typos - eg non capitalisation at the start of a sentence 11. The supplementary material still has inline comments, including apparently unresolved questions of the direction of effect, and a "is this correct?". Is it?</p>
--	---

REVIEWER 1	Shai D Silberberg
	<i>NIH</i>
REVIEW RETURNED	09/08/2018

GENERAL COMMENTS	<p>The study by Fernandes and colleagues evaluated in vivo studies of SOD1 mice models of ALS between 2005 and 2015 for the reporting of methodological and regulatory parameters. The authors examined whether the level of reporting improved following the publication of reporting guidelines. To this end they generated an index of items that contribute to "methodological standards" (labeled MSR) as well as an index of items that contribute to "regulatory compliance" (labeled RCR) and assigned numerical values to each of the items. The three main findings claimed by the authors, as summarized in the Discussion, are: 1. There is an overall improvement in regulatory compliance and the reporting of methods. 2. Reporting was better in studies testing the effects of drugs (preclinical studies) in</p>
-------------------------	--

comparison to studies investigating the mechanisms of disease (proof-of-concept studies). 3. The level of reporting depended on the country of origin of the first author. Overall, the manuscript could be greatly simplified given that two of the results strongly suggest that the ALS guidelines have had little impact. First, only approximately 10% of studies used at least 24 animals per group, as recommended by the guidelines (Figure 4a) and the average group size did not change over time (Figure 4b). Second, many of the studies did not use both females and males in each group, as recommended by the ALS guidelines. Of importance, the manuscript is peppered with imprecisions, lacks important details, and overstates results. Here I list these concerns.

Imprecisions a. Perhaps the most concerning aspect of the manuscript is the large number of numerical errors and inconsistencies. These errors reduced confidence in the analysis and in the numbers presented in tables. For examples: (1) on line 165 it states “382 full-text articles remained for analysis”. This is clearly wrong; the number should be 569. (2) The legend to Figure 2 states 1306 abstracts. Where did this number come from? (3) The sentence starting on line 331 contains several errors. 57/194 should be 57/294 and 35/194 should be 35/294. These changes lead to different percentages. Furthermore, even after the corrections the percentages do not add up to 100%. (4) Line 388 indicates that Supplemental Table 3 includes a total of 569 papers, yet the supplemental Table 3 indicates 490 papers. (5) In line 390, 62/382 should be 62/569. I find this error particularly troubling since the number 382 was also mistakenly used in example 1 above. Where does this number come from? b. There are a few statements that are not well substantiated and references that do not support the statements. For examples, (1) Line 83 states that “Systematic reviews of animal use in both neuroscience [5 6] and other fields of research e.g. [7 8] indicate that self-reported regulatory compliance – including of ethical approval of protocols – has steadily increased over the last decade”. The listed references don’t support this claim. (2) On line 94 it is not clear why references 11 and 12 are used. (3) The sentence starting on line 301 states that most of the items listed in Table 3 are covered by the ALS guidelines. However, only 5 of the 12 items in the Table received an Index weight of 1.5 assigned to items in the ALS guidelines. (4) On line 539, references 29 and 76 do not support the claim that “there is ample evidence for ALS that published studies which do not report measures to minimise bias (i.e. blinding, randomisation and allocation concealment) tend to present an exaggerated estimate of

the therapeutic effect of experimental drugs.” Lack of details

a. The sentence on line 161 indicates that searches spanned a four-year period. Does this mean that analysis spanned a similar period? If so, did the same individuals conduct the analysis over the entire time or were different years analyzed by different individuals? b. Line 228: there is insufficient information on the regression analysis. What justification is there to use linear regression? What do the fits look like? c. Line 237: it is not clear what “adjusted after checking for internal consistency” means. d. Line 243: No explanation is provided why the specific numerical value of 1.5 was used for items which are also part of the ALS guidelines. How was this number chosen? What would the results look like with other numerical values? e. In the legend to Table 3, as well as in the paragraph starting on line 278, no explanation or justification is given for the various statistical tests. f. In the legend to Table 3 it states that “The index score for each variable is provided (MSR index score ranging from 0 to 12.5, and RCR score ranging from 0 to 3)”, but the Table does not appear to include this information. g. Line 278: what are “all significant model effects”? h. Line 287: Why could only 490 observations be used? i. Figure 3: Are the error bars standard deviations? j. Line 372: “Studies with high severity seemed to have higher odds of high RCR values ($p=0.027$)” needs to be explained. k. Figure 5a presents the “predictive marginal means” as a function of publication year. There is no explanation what “predictive marginal means” are, how they were calculated, and why the data is presented in this way rather than in the same format of Figure 3a. Without this information/explanation it is difficult to assess the significance of the results presented in this figure.

Overstatements a. One of the important conclusions in the manuscript is that “General methodological standards improved significantly over a ten-year period”. The evidence for this statement is summarized in Figure 3a that shows the mean level (and presumably the standard deviation) of reporting of methodological items over the six years examined in this study. Given the large variability, and the similarity between the values in 2005 and 2015, I see no justification for the claim that “reporting standards improved over time”. b. Line 24: The ALS research community did not pioneer the adoption of methodology guidelines to improve preclinical research reproducibility. The stroke community published the STAIR criteria in 1999. c. Line 317: “Sweden appeared to have comparatively low MSR values” is not substantiated, given the large error bars (and small sample). d. Line 463 states “....suggesting widespread compliance to

	<p>published guidance in this respect”, yet the very next sentence contradicts this ‘suggestion’: “However, this endpoint was already broadly used before the publication of the guidelines suggesting that these reflect common practice at the time of publication”. e. Line 488 states that “Methodological standards reporting improved over the time period under study” yet on line 493 the authors state that “Throughout the period under study, the MSR scores remain below 50% of the maximum score, showing that the overall level of reporting of methodological detail did not change and remain substantially below the recommendations in the guidelines”.</p>
--	---

VERSION 1 – AUTHOR RESPONSE

We thank the two reviewers for the attentive and detailed review and constructive criticism. We have done our best to revise the manuscript accordingly, as outlined in detail in the following.

Please note that line numbers in our response refer to the revised manuscript, unless something else is indicated.

Both reviewers

In response to the questions about the use of weighting.

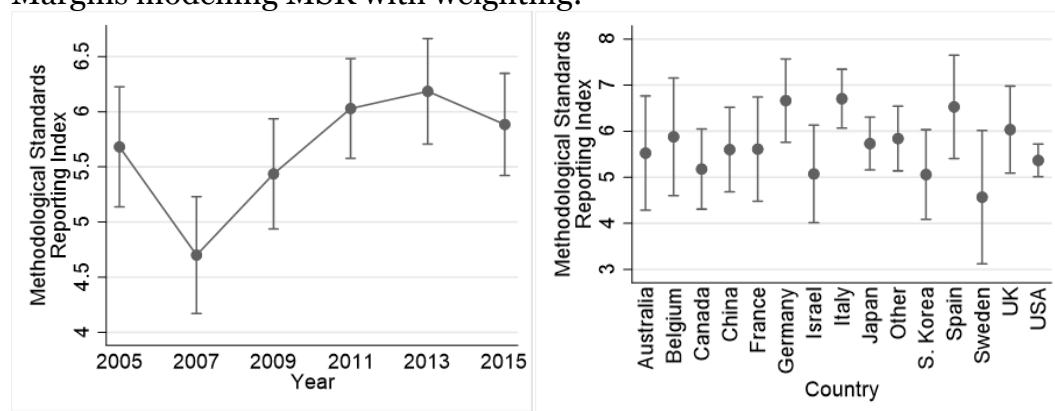
We chose to create two aggregate measures (the MSR and RCR measures) in order to be able to use statistics that are more robust. We included in these aggregate measures a number of items which we consider relevant for determining the methodological standard and regulatory compliance reporting, on basis of previous literature (see also our response to comment 1 above). Among these items, there are five that are also part of the ALS guidelines. Whereas we think that all the items we included are relevant, we also decided that a greater weight should be attributed to the items that are in the ALS guidelines.

To use an aggregate measure without weighting, as suggested by Reviewer 1, would mean attributing equal importance to the items identified by us (or other individual team of authors) and to the items recognized by the international research community in the field. Whereas we recognize that the choice of weight will always be somewhat arbitrary, we do not agree that avoiding weighting altogether would be a better approach.

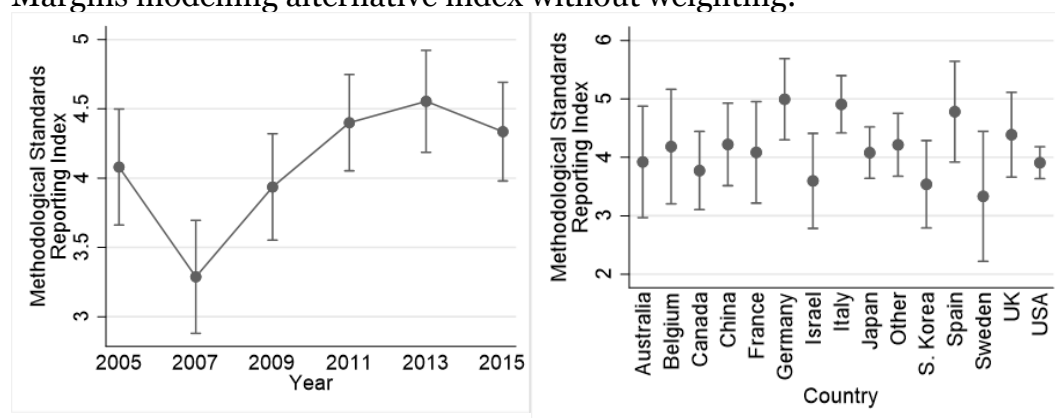
We have instead followed the recommendation by Reviewer 2, to test the consequences of weighting the MSR items, instead of using an unweighted sum, by modelling an alternative index formed without weighting. The plots below show that

the differences between years and countries remained virtually unchanged, although the scale on the y-axis changed because the alternative index values were generally lower. Hence the weighting did not create the reported results and our conclusions would have remained unchanged. We now mention the investigated alternative index calculation, on line 339.

Margins modelling MSR with weighting:



Margins modelling alternative index without weighting:



Reviewer 1 (Malcolm Macleod):

Thanks for the opportunity to review this interesting work. I have some comments. it may be that many of these things had been done but didn't make it to the manuscript.

1. The first search was done in 2013. PRISMA Item 5 asks to indicate whether a protocol exists, and your response is "N/A". On the basis that a protocol does not exist, you have great researcher degrees of freedom in what you did, and this is a major weakness which should be addressed. For instance, were your outcome measures articulated before data collection began, or not?

The review protocol was defined to cover relevant variables on regulatory compliance, refinement and research quality. It was based on a) the protocol we have previously used for systematic review of animal ethics and refinement measures in research (PMID: 23110093, 23215663, 18625582), b) the ALS guidelines and c) additional recommendations for preclinical research using mouse models of ALS (PMID 24678540). The review protocol was defined prior to data collection, as is now mentioned on line 194. We apologize for overlooking that this point on the PRISMA checklist did not refer only to pre-registered protocols.

2. It's not clear whether citation screening, assessment of preclinical v exploratory, and annotation were performed by one person or two; and if annotation for quality was performed blinded to year of publication if this were possible.

Citation screening, assessment of study type and annotation were performed by Joana Fernandes for all sampled years and all papers. Nuno Franco provided support for disambiguation, throughout the analysis. This information has been added on line 194.

3. Choice of web of science to search rather than pubmed is unusual for preclinical SRs - doesn't make it wrong, but might be worth mentioning why you took that approach?

We recognize that this may appear an unusual choice from the perspective of researchers performing typical systematic reviews of preclinical research. We approached the topic from our background in refinement and animal welfare research, an area for which Pubmed is not an ideal database, which is the reason why we used Web of Science as has traditionally been our publication database of choice. The database used is clearly stated in the paper.

4. There is a problem with using a checklist (present/ absent) and then weighting items and then summing them to give a measure of quality or reporting which you then consider at a ratio scale (4 is twice as good as 2) and plug the weighted sum as the dependent variable into a regression equation. In my view there is not sufficient evidence to assign weights to these items, and a better approach is the number of checklist items scored.

See the response under “Both reviewers” above.

5. You might like to discuss the possible weakness of running several statistical tests each with a p threshold of 0.05. This can be justified if the questions are independent, but in this case I am not sure that that is the case.

Indeed, a large number of significance tests introduce a risk of falsely positive (significant) tests, even if these tests are independent, which may motivate e.g. an elevated significance threshold (lowered alpha). In this case, we estimated two models which were not particularly large. Our way to report the model results is consistent with standard procedures in observational research and multivariable modelling. We also agree that pairwise comparisons across several time-points require correction for multiple tests, which may have warranted caution had we drawn conclusions about specific years. However, we did not formulate hypotheses regarding the effect of specific publication years or countries. Instead, we wanted to point to the fact that the scores increased successively over the study period and differed between countries. Nevertheless, due to the nature of the study, we admit that the conclusions on lines 43 (now 44) and 554 (now 641) should be modified to state indications rather than proofs of effects. These have now been changed to read on line 44: “standards improved significantly over a ten-year period” to “standards improved gradually over an 8- to 10-year period” and on line 641 onwards: to “In contrast to previous research, this study indicated a gradual improvement in the methodological standards and regulatory compliance reporting scores over time. However, it is difficult to say to what extent this is the result of field-specific guidelines, as there is an overall increasing trend in these scores.”.

6. *Your MSR and RSR indexes are potentially interesting but unvalidated. I am concerned that - in expressing these with capitalisation and a name, rather than a list of items - you consider these now to be validated tools. You have not shown their validity, and so it would be preferable simply to list these items without claiming that you have developed an "index". What would be useful (and you have done this, I think) is to articulate a checklist which allows determination of compliance with the ALS reporting guidelines.*

We appreciate that the term index may be interpreted as a validated measure and have replaced it with the term "score".

7. *You claim this is the first demonstration of the impact of field specific reporting guidelines, but Minnerup (PMID: 26658439) and Ramirez (PMID: 28373349) have both looked at the change in quality in reporting of in vivo stroke studies, and we looked (Bahor PMID: 28373349) at changing quality in MCAO and lacunar stroke, and at the change in a specific drug in stroke (McCann, PMID: 27526101) ... so I don't think that claim is valid.*

We thank the reviewer for having pointed this out. Our study is the first to cover both methodological standard and regulatory compliance reporting, and we have changed the wording accordingly.

8. *It's not clear whether you would include ex vivo SOD mice (eg tissue slice, myoculture) - I suspect not, but absent a clear articulation of inclusion and exclusion criteria it is difficult to be sure*

The main inclusion criterion was reported use of mice carrying the SOD1 mutation. For studies where animals were killed before showing any clinical signs, a severity classification of "1" was attributed. In most cases, behavioural observations or other procedures were carried out on the animals before being euthanized, typically followed by ex-vivo work. Also the few studies when all work was carried out *ex vivo* without any prior intervention done to the animals, were included, as they met the main inclusion criterion.

9. *References 33 and 79 appear to be duplicates.*

The duplicate reference has been removed.

10. *There are a few typos - eg non capitalisation at the start of a sentence*

Typos have been checked and corrected.

11. *The supplementary material still has inline comments, including apparently unresolved questions of the direction of effect, and a "is this correct?". Is it?*

We apologise for not submitting the tables sans comments. The questions referred to issues which had been sorted out prior to submission, and the values are, indeed, correct. The question of direction of effect had been raised for the item "severity" but as this item was not included in the final version of the RCR, the question became irrelevant.

Reviewer 2 (Shai D Silberberg):

The study by Fernandes and colleagues evaluated in vivo studies of SOD1 mice models of ALS between 2005 and 2015 for the reporting of methodological and regulatory parameters. The authors examined whether the level of reporting improved following the publication of reporting guidelines. To this end they generated an index of items that contribute to “methodological standards” (labeled MSR) as well as an index of items that contribute to “regulatory compliance” (labeled RCR) and assigned numerical values to each of the items. The three main findings claimed by the authors, as summarized in the Discussion, are:

- 1. There is an overall improvement in regulatory compliance and the reporting of methods.*
 - 2. Reporting was better in studies testing the effects of drugs (preclinical studies) in comparison to studies investigating the mechanisms of disease (proof-of-concept studies).*
 - 3. The level of reporting depended on the country of origin of the first author.*
- Overall, the manuscript could be greatly simplified given that two of the results strongly suggest that the ALS guidelines have had little impact. First, only approximately 10% of studies used at least 24 animals per group, as recommended by the guidelines (Figure 4a) and the average group size did not change over time (Figure 4b). Second, many of the studies did not use both females and males in each group, as recommended by the ALS guidelines.*

It is not clear to us in which way the reviewer suggest that we simplify the manuscript. We agree that the low adherence to minimum sample sizes and use of male and female animals say something about the limited impact of the guidelines. However, we do not think that it would be appropriate to base the manuscript entirely on this observation, given that these are only two items on a much longer list of items of relevance for methodological standards.

Of importance, the manuscript is peppered with imprecisions, lacks important details, and overstates results. Here I list these concerns.

Imprecisions

a. Perhaps the most concerning aspect of the manuscript is the large number of numerical errors and inconsistencies. These errors reduced confidence in the analysis and in the numbers presented in tables.

We understand this concern, The discrepancies in numbers derive from the fact that the manuscript was initially written up with results from 2005-2013, with results from 2015 being added when the manuscript was revised for submission to BMJ Open Science. Despite our best efforts in cross checking the text, some of the values regrettably remained unchanged. We are thankful to the reviewer for pointing out the following discrepancies.

(1) on line 165 it states “382 full-text articles remained for analysis”. This is clearly wrong; the number should be 569.

This has been corrected.

(2) The legend to Figure 2 states 1306 abstracts. Where did this number come from?

We apologise for this mistake, and wish to thank the reviewer for spotting it. In revising the manuscript to reflect the inclusion of 2015 data, we apparently failed to spot this sentence. The figure provided is hence correct, as it refers to the actual sample reported in this paper (N=569), and the legend has been updated accordingly.

(3) The sentence starting on line 331 contains several errors. 57/194 should be 57/294 and 35/194 should be 35/294. These changes lead to different percentages. Furthermore, even after the corrections the percentages do not add up to 100%.

We thank the reviewer for identifying this inconsistency, for which we apologise. This derives from a typo and a minor inconsistency in the dataset, both of which have been corrected.

(4) Line 388 indicates that Supplemental Table 3 includes a total of 569 papers, yet the supplemental Table 3 indicates 490 papers.

It is unclear to us which part of the text the reviewer refers to, as there is no mention of Supplemental Table on line 388 (which reads “*Over the entire period, most papers (67.0%; 381/569) reported that studies had been appraised*”). As to why descriptive statistics for the different variables are expressed as a proportion of 569 papers, whereas Supplemental Table 3 indicates 490 papers, this is because only papers where there were unambiguous data for all items on the respective MSR and RCR register could be included in the logistic analysis.

(5) In line 390, 62/382 should be 62/569. I find this error particularly troubling since the number 382 was also mistakenly used in example 1 above. Where does this number come from?

The reviewer is correct in spotting the same number repeated twice in the paper. This was indeed the same issue mentioned before. Our first sample, before extending the study to include more years, comprised 382. This has now been corrected.

b. There are a few statements that are not well substantiated and references that do not support the statements. For examples, (1) Line 83 states that “Systematic reviews of animal use in both neuroscience [5 6] and other fields of research e.g. [7 8] indicate that self-reported regulatory compliance – including of ethical approval of protocols – has steadily increased over the last decade”. The listed references don’t support this claim.

This has been corrected to include only our own previous work on Huntington’s and mycobacteria models (PMID 23215663 and 23110093) which indeed support the claim.

(2) On line 94 it is not clear why references 11 and 12 are used.

Lines 93-94 refer to refinement measures to facilitate access to food and water for locomotor impaired animals. References 11 and 12 are from studies reporting the use of such measures.

(3) The sentence starting on line 301 states that most of the items listed in Table 3 are covered by the ALS guidelines. However, only 5 of the 12 items in the Table received an Index weight of 1.5 assigned to items in the ALS guidelines.

The wording has been changed to remove the reference to the ALS guidelines here.

(4) On line 539, references 29 and 76 do not support the claim that “there is ample evidence for ALS that published studies which do not report measures to minimise bias (i.e. blinding, randomisation and allocation concealment) tend to present an exaggerated estimate of the therapeutic effect of experimental drugs.”

This has been reworded to remove the specific reference to ALS research and maintain only the references to other areas of research.

Lack of details

a. The sentence on line 161 indicates that searches spanned a four-year period. Does this mean that analysis spanned a similar period? If so, did the same individuals conduct the analysis over the entire time or were different years analyzed by different individuals?

This has been clarified in the manuscript, see point 2 in the response to reviewer 1.

b. Line 228: there is insufficient information on the regression analysis. What justification is there to use linear regression? What do the fits look like?

On line 288, we only state that regression modelling was applied; the type of models used is explained on lines 275-276. In the following paragraph, lines 278-282, we describe how the models were validated by examining residuals and using several tests. The model of MSR was checked with the Breusch-Pagan / Cook-Weisberg test for heteroscedasticity ($p=0.35$, which was excellent), the Ramsey RESET test for omitted variables ($p=0.87$, which was excellent), the Pregibon link test ($p\text{-hat}=0.23$ and $p\text{-hatsq}=0.64$, which was acceptable), and by examining residuals. The distribution of standardized residuals was close to Normal.

For the logistic model of RCR, the options for validation is more limited. The Pregibon link test was used ($p\text{-hat}=0.000$ and $p\text{-hatsq}=0.42$, which was excellent) and residuals were examined. The model was re-fitted while excluding three observations with high residuals, which showed that the estimates of year and country did not change dramatically. The estimates of the complete model (without excluding observations) were also generally more conservative. We also tried a different approach, using ordinal regression (a generalized ordered logit model), which was less successful. We therefore regarded the used logistic model to be justified and valid.

Not to burden the readers with too much information, we did not include validation results. However, in response to this reviewer comment, we have now included short statements on lines 366 and 446.

c. Line 237: it is not clear what “adjusted after checking for internal consistency” means.

On line 237-238 (now 259) we only mention that internal consistency was checked. The procedure is described in much more detail on lines 252-259. Not to confuse readers, we have now deleted the first mentioning. We also improved the sentence on lines 253-254 (now 277).

d. Line 243: No explanation is provided why the specific numerical value of 1.5 was used for items which are also part of the ALS guidelines. How was this number chosen? What would the results look like with other numerical values?

See response under Both reviewers above.

e. In the legend to Table 3, as well as in the paragraph starting on line 278, no explanation or justification is given for the various statistical tests.

These are tests of model specification and fit, although they are far from always used in similar situations. Citations of relevant publications have been provided and we feel that this is sufficient explanation and justification.

f. In the legend to Table 3 it is states that “The index score for each variable is provided (MSR index score ranging from 0 to 12.5, and RCR score ranging from 0 to 3)”, but the Table does not appear to include this information.

The information as such is not in the table, which is why we provided it in the legend. However, this is the result of summing the minimum versus maximum values for each item. Scoring 0 on each of the individual items would give a total score of 0. Scoring the maximum value on each of the individual items would give a total score of 12.5 for MSR and 3 for RCR.

g. Line 278: what are “all significant model effects”?

In statistical modelling terminology, an ‘effect’ is an independent variable (or covariate) included in the model. Thus, ‘all significant model effects’ are all the independent variables in the regression models with $p \leq 0.05$ in a joint Chi-square test. To increase clarity, we changed “model effects” to “independent variables” (now line 319).

h. Line 287: Why could only 490 observations be used?

Only papers where there were unambiguous data for all items on the respective MSR and RCR scale could be included.

i. Figure 3: Are the error bars standard deviations?

In Figures 3 and 5 the error bars are 95% confidence intervals. This explanation was unfortunately overlooked by us and has now been inserted in the figure captions. In Figure 4, on the other hand, the bars represent the standard deviation, which is explained in the caption.

j. Line 372: “Studies with high severity seemed to have higher odds of high RCR values ($p=0.027$)” needs to be explained. ~

It is not clear to us what is to be explained here. This sentence uses standard wording for reporting a result from a logistic regression. In plain English it means that a study with high severity is more likely to have a high RCR value. However, the original wording is the way that we report logistic regression results throughout the paper.

k. Figure 5a presents the “predictive marginal means” as a function of publication year. There is no explanation what “predictive marginal means” are, how they were calculated, and why the data is presented in this way rather than in the same format of Figure 3a. Without this information/explanation it is difficult to assess the significance of the results presented in this figure.

Margins are statistics calculated from predictions of a fit model at fixed values of some covariates and averaging or otherwise integrating over the remaining covariates. Margins are therefore a way to show model results in a more easily understood way than a table of coefficient estimates. There are different types of margins. Least-squares means (also called estimated marginal means), which are perhaps the margins most known to some researchers, assume that all levels of categorical covariates are equally likely or, equivalently, that the design is balanced. Such margins are therefore most relevant in experimental research. ‘Predictive marginal means’ or ‘predictive margins’, on the other hand, are margins where the remaining covariates (those not fixed for the calculation, in this case year and country, respectively) are assumed to have the values observed in the data. This type of margins are more applicable in observational research.

Depending on the type of model, margins are expressed in different ways. Based on a linear regression model, the margins simply show the predicted values of the outcome in the different categories of the covariate (year or country). When using logistic regression, the margins instead usually show the probability that the outcome is positive (1, in contrast to negative or 0), because that is what the model estimates.

Figure 3 therefore shows the predicted MSR values, according to the linear regression model, for different values of year and country, assuming that all other variables in the model have their observed values. Similarly, Figure 5 shows the predicted probabilities of a ‘high’ RCR (>1 , according to the logistic regression model), for different values of year and country, assuming that all other variables in the model have their observed values.

To help the readers, we have now included an explanation of the margins on line 319.

Overstatements

a. One of the important conclusions in the manuscript is that “General methodological standards improved significantly over a ten-year period”. The evidence for this statement is summarized in Figure 3a that shows the mean level (and presumably the standard deviation) of reporting of methodological items over the six years examined in this study. Given the large variability, and the similarity between the values in 2005 and 2015, I see no justification for the claim that “reporting standards improved over time”.

It is true that the values for 2005 and 2015 are similar. But it would not be correct to state that there is no improvement given that there is a significant effect of time, and

most of the curve has an upward trajectory indicating that the scores improved gradually. We have modified the statement on line 44 slightly (as explained above). As now also explained, the error bars in Figures 3 and 5 do not represent standard deviation but 95% confidence intervals.

b. Line 24: The ALS research community did not pioneer the adoption of methodology guidelines to improve preclinical research reproducibility. The stroke community published the STAIR criteria in 1999.

This has been corrected to “The ALS research community was one of the first to adopt”.

c. Line 317: “Sweden appeared to have comparatively low MSR values” is not substantiated, given the large error bars (and small sample).

This sentence has been removed.

d. Line 463 states “...suggesting widespread compliance to published guidance in this respect”, yet the very next sentence contradicts this ‘suggestion’: “However, this endpoint was already broadly used before the publication of the guidelines suggesting that these reflect common practice at the time of publication”.

“suggesting widespread compliance to published guidance” has been changed to “suggesting researchers to a great extent act in accordance with published guidance” (line 547)

e. Line 488 states that “Methodological standards reporting improved over the time period under study” yet on line 493 the authors state that “Throughout the period under study, the MSR scores remain below 50% of the maximum score, showing that the overall level of reporting of methodological detail did not change and remain substantially below the recommendations in the guidelines”.

We thank the reviewer for pointing out this contradiction. We have changed the wording on line 493 (now 579) to “showing that the overall level of reporting of methodological detail remain substantially below the recommendations in the guidelines”.

BMJ Open Science

BMJ Open Science is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers' comments and the authors' responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Open Science is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (<http://openscience.bmj.com>).

If you have any questions on BMJ Open's open peer review process please email info.bmj@bmj.com

1 **Methodological standards, quality of reporting, and regulatory compliance in**
2 **animal research on amyotrophic lateral sclerosis: a systematic review**

3 Joana G Fernandes^{1,2¶}, Nuno H Franco^{1,2¶}, Andrew J Grierson³, Jan Hultgren⁴, Andrew JW
4 Furley^{5&}, I Anna S Olsson^{1,2&*}

5 ¹ Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Portugal

6 ² IBMC-Instituto de Biologia Molecular e Celular, Universidade do Porto, Portugal

7 ³ Sheffield Institute for Translational Neuroscience, Department of Neuroscience, University of
8 Sheffield, Sheffield, United Kingdom

9 ⁴ Department of Animal Environment and Health, Swedish University of Agricultural Sciences,
10 Skara, Sweden

11 ⁵ Department of Biomedical Science, University of Sheffield, Western Bank, Sheffield, United
12 Kingdom

13 *Corresponding author

14 E-mail: olsson@ibmc.up.pt

15 ¶ These authors share the first authorship

16 & These authors share the last authorship

17

18 **Key words:** Amyotrophic Lateral Sclerosis, ALS, Guidelines, Methodology, Reporting, Quality,
19 Compliance, Animal Welfare, Reproducibility

20

21 **Word count:** 5248

22 **Abstract**

23 **Objectives**

24 The ALS research community pioneered the adoption of methodology guidelines to improve
25 preclinical research reproducibility. We here present results of a systematic review to
26 investigate how the standards in this field changed over the ten-year period during which the
27 guidelines were first published (2007) and updated (2010).

28

29 **Methods**

30 We reviewed 569 research papers reporting research with SOD1 mice, published between 2005
31 and 2015.

32

33

34 **Results**

35 Reporting standards improved over time. Of papers published after the first ALS guidelines were
36 made public, fewer than 9% referred specifically to these. Of key research parameters, only
37 three (genetic background, number of transgenes and group size) were reported in >50% of the
38 papers. Information on housing conditions, randomization and blinding were absent in over two
39 thirds of papers. Group size was among the best reported parameters, but the majority reported
40 using fewer than the recommended sample size and only two studies clearly justified group size.

41

42 **Conclusions**

43 General methodological standards improved significantly over a ten-year period but
44 remained generally comparable to related fields with no specific guidelines, except with
45 regard to severity. only 11% of ALS studies were classified in the highest severity level (animals
46 allowed to reach death or moribund stages), substantially below the proportion in studies of
47 comparable neurodegenerative diseases such as Huntington's. The existence of field-specific
48 guidelines, though a welcome indication of concern, seems insufficient to ensure adherence
49 to high methodological standards and other mechanisms may be required to improve
50 methodological and welfare standards.

51

52 **Strengths and limitations:**

53 - This systematic review is the first to assess quality in a research field with community guidelines
54 on methodological standards.

55 - Our large sample (N=569 papers) includes half the total population of published papers
56 between 2005-2015

57 - The approach for this systematic review is unique in covering methodological quality,
58 regulatory compliance and severity / animal welfare

59 - We built two comprehensive indexes (for methodological standard and for reporting quality)
60 which were analysed with regression analysis in order to investigate how the two indexes
61 (dependent variables) were related to publication year, type of study, country of origin and
62 journal (explanatory or predictor variables).

63 - While more models of ALS are now available, only studies using the SOD-1 mouse were
64 included

65 **1. Introduction**

66 Amyotrophic lateral sclerosis (ALS) is a rapidly progressing neurodegenerative disease typically
67 resulting in death within two to five years after the onset of symptoms. There is no known cure
68 and the most widely used treatment– riluzole – extends survival by just two months ¹. ALS
69 research using animal models focuses primarily on two main interconnected goals:
70 understanding the underlying mechanisms involved in motor neuron death in the brain and
71 spinal cord, and development and testing of potential drug therapies ². This research relies
72 substantially on genetically modified animals, particularly transgenic mice expressing mutant
73 forms of the human Superoxide Dismutase 1 (SOD1) gene, which manifest several important
74 characteristics of the human disease ^{3 4}.

75

76 While the use of animal models is relevant for advancing knowledge and considered essential
77 for testing putative treatments, it also presents ethical challenges and this may be a reason for
78 public concern. As a result, a common legal requirement in many countries is that animal
79 research projects undergo an evaluation process intended to ensure that protocols are designed
80 and carried out in compliance with the 3Rs principle: *replacement* of animal use by non-animal
81 methods, *reduction* of animal numbers needed to achieve the scientific objectives, and
82 *refinement* of procedures to reduce or prevent harm to animals and improve their wellbeing.
83 Systematic reviews of animal use in both neuroscience ^{5 6} and other fields of research e.g. ^{7 8}
84 indicate that self-reported regulatory compliance – including of ethical approval of protocols –
85 has steadily increased over the last decade, but that significant progress could still be made to
86 minimise and prevent avoidable suffering of laboratory animals. One key measure for
87 accomplishing this is the termination of experiments during less severe stages of disease
88 development where it is scientifically valid to do so. Endpoints based on early obtainable and
89 scientifically sound indicators of phenotype progression can not only improve the ethical
90 acceptability of animal studies, but also prevent the confounding influence of secondary factors;
91 in the case of animal models of neurodegenerative diseases, starvation and dehydration arising
92 from difficulties in eating and drinking due to progressive motor impairment can affect the
93 phenotype and the readout of survival studies ⁸⁻¹⁰. Simple refinements – such as adding mash
94 food and longer bottle spouts – can however help reduce the influence of such factors ¹¹⁻¹³.

95

96 Of related concern are reports that a number of published animal studies fail to uphold basic
97 standards regarding experimental design – e.g. random assignment of animals to treatment
98 groups, blinding of observers – or use too few animals often leading to irreproducible results of

99 limited translational value¹⁴⁻¹⁹. This also holds true for neuroscience²⁰⁻²³, with concerns over the
100 overall quality and reproducibility of published results being raised for several neuroscience sub-
101 fields, including multiple sclerosis²⁴, stroke²⁵, spinal cord injury²⁶ Alzheimer's²⁷, Parkinson's²⁸,
102 Huntington's¹³ and ALS²⁹ research. This has led major science funders, including the National
103 Institutes of Health³⁰ and Research Councils UK³¹ to demand that future grant proposals attest
104 to the likelihood of providing reliable results, by including details of experimental design and
105 adequate justification of sample sizes. Reproducibility is further hindered by insufficient
106 provision of information on methodology in published research³² – including failure to account
107 for key variables such as sex, genotype, age, and weight of animals, anaesthetics used or
108 methods of euthanasia. Omitting information also makes it impossible to evaluate the study
109 quality and there is evidence that papers that do not report randomization or blinding report
110 exaggerate biological effects³³⁻³⁵.

111

112 Broadly, the public conditionally approves of animal studies on the assumption that the harm
113 caused is offset by the benefits achieved and that scientists strive to minimise the former and
114 optimise the latter^{36,37}. Doing so requires scientists to critically revise their methods to maximise
115 translational relevance^{19,38}. Scientists are rightly concerned and, within the self-correcting
116 process of science, must rely on themselves to both identify the main obstacles hindering its
117 progress and find adequate solutions. To address the issue of methodological standards and
118 quality of reporting of basic and applied ALS studies, the ALS research community held two
119 meetings in 2006 and 2009, resulting in the publication of guidelines for animal studies in this
120 field^{2,39}. These guidelines aim to improve and standardise research methodology, and
121 encourage authors and journals to publish negative results in order to avoid publication bias.
122 The actual impact of such guidelines on how the ALS community carries out and reports research
123 has however not been assessed.

124

125 The present systematic review of animal studies of ALS uniquely aimed to assess, over an
126 extended period, the attention given to relevant methodological parameters (as a proxy for the
127 likely reliability of the study) and to examine how the principles of *refinement* and *reduction*
128 (measures to minimise animal harm) were considered. Both proof-of-concept and preclinical
129 studies were included in order to assess the influence of type of study.

130 2. Methods

131 2.1 Database search

132 An advanced search was conducted on the *ISI Web of Science*[®] database with the query *TS =*
133 *((mice OR mouse) SAME (ALS OR "amyotrophic lateral sclerosis"))*. Results were refined to
134 include only original research articles written in English and published in 2005, 2007, 2009, 2011,
135 2013 and 2015. Years of publication were selected to include papers reporting research planned
136 and carried out prior to and after the publication of guidelines for ALS research in 2007³⁹ and
137 2010², resulting from two international meetings held in 2006 and 2009, respectively (Figure 1).

138

139

140 **Figure 1. Timeline of relevant events.** The bottom arrows signal the years for which papers in our sample
141 were retrieved and the top arrows indicate the years when workshops on best practice in ALS animal
142 research were held, as well as when guidelines stemming from these were published. The grey bars
143 illustrate the 1-4 year period over which ALS animal studies reported in 2005 were likely to have been
144 designed and carried out, an estimation that can also be applied for the other years reviewed (2007, 2009,
145 2011, 2013, and 2015).

146

147

148 The choice to focus on SOD-1 mice was based on the predominant role of this model in animal-
149 based research into ALS (see Supplementary Figure 1).

150

151

152 **Supplementary Figure 1 - Trends in animal model chosen in ALS research, based on the number of hits**
153 **from an *Clarivate Analytics Web of Science*[®] advanced search applying the search queries: a) *TS=*(("ALS"
154 OR "amyotrophic lateral sclerosis") AND "SOD1" AND ("mouse" OR "mice")); b) *TS=*(("ALS" OR
155 "amyotrophic lateral sclerosis") AND "TDP-43" AND ("mouse" OR "mice")); and c) *TS=*(("ALS" OR
156 "amyotrophic lateral sclerosis") AND "FUS" AND ("mouse" OR "mice"))**

157

158

159 The search was performed in February 2013 for scientific articles from 2009 and 2011 (after the
160 first and second conferences, respectively), in August 2013 for scientific articles from 2005
161 (before the two conferences), in September 2014 for scientific articles from 2013, in November
162 2016 for scientific articles from 2015, and in February 2017 for scientific articles from 2007. After
163 the triage process, illustrated in Figure 2, 382 full-text articles remained for analysis: 77 from
164 2005, 81 from 2007, 84 from 2009, 106 from 2011, 115 from 2013, and 106 from 2015.

165

166 **Figure 2. Triage process.** The first triage step involved reading each of the 1306 abstracts and excluding
167 all papers that were not related to ALS. The second triage step excluded all papers that did not report
168 original research with SOD1 models of the disease.

169

170 2.2 Data collection

171 Each published study was categorised as either a ‘preclinical’ (i.e., carried out “to evaluate a
172 drug for use in humans”) or ‘proof-of-concept’ (i.e., aiming “to elucidate the mechanism of the
173 disease”), according to the suggested classification for animal studies on ALS^{2 39}. Table 1
174 describes the information retrieved regarding regulatory compliance, animal models,
175 experimental design and animal welfare. This information was retrieved through careful reading
176 of the full papers.

177

178 **Table 1. Data retrieved.** A description of the information collected from revised papers is presented for
179 each item.

Category	Items	Description/Observations
Regulatory compliance	Ethical approval	Studies explicitly reported to be approved by a committee / authority.
	Guideline compliance	Articles that did not report having experimental protocols ethically approved by an institution or national entity, but reported that some kind of guidelines for use and care of laboratory animals was followed.
Animal models	Genetic background	When available.
	Sex	Four options: Male, female, both or not reported. For <i>both</i> , information on whether studies were balanced for gender was retrieved.
	Number of transgene copies	When available.
Experimental design	Group size	Mean group size, based on the available information
	Randomization	Studies explicitly reporting assigning animals to groups randomly
	Blinding	Studies explicitly reporting blinding of observers to experimental groups
	Non-transgenic littermate control	Studies explicitly reporting the use of non-transgenic littermates as control.
	Splitting littermates into groups	Studies explicitly reporting that littermates were split into groups.
	Housing and husbandry conditions	Reporting information regarding temperature, humidity, light of the room where animals were kept, and cage size and number of animals per cage.
Animal Welfare/ Procedures	Severity	Described in table 2.
	Refinement	Relevant refinements to minimise suffering and distress, such as housing adaptations.
	Euthanasia method	Euthanasia methods were divided into the following categories: “Under anaesthesia” (including anaesthetic overdose); “CO ₂ asphyxiation”; “Other”; “Not reported” and “Not performed”.

180

181 For severity assessment, a scale was devised based on the specific characteristics of the ALS
182 models and their progressive disease phenotype (Table 2). The ALS models used in the reviewed
183 studies express diverse mutant forms of the *SOD1* gene. The onset of disease for these models
184 is generally characterised by weakness and tremors of the hind limbs, together with a mild loss
185 of body weight. Disease progression leads to paralysis of hind limbs, followed by complete
186 paralysis (example, Figure 3 in ⁴⁰, accompanied by increased difficulty to eat, drink and swallow
187 ^{41 42}. Mice die of respiratory failure due to paralysis of the diaphragm ⁹. Age of onset and death,
188 as well as the interval between them, vary depending on the mutation of the amino-acid and
189 codon e.g. ⁴³, number of copies of transgene e.g. ⁴⁴, and genetic background ⁴. For instance, the
190 over-expressing SOD1G93A Line Gur 1H (B6SJL hybrid) presents with an early onset of overt
191 motor symptoms (3-4 months) and moderate rate of progression (3 weeks from onset to death)
192 ⁴⁵, whereas the highly expressing SOD1G85R Line 148 presents with later onset (7.5 months)
193 with faster disease progression (2 weeks from onset to death) ⁴⁶. Also, such factors as the animal
194 supplier (e.g. ^{47 48}), in-house breeding ⁴⁹ and crosses with other non-SOD1 models (e.g. SOD1
195 mice crossed with gene-specific knockout mice ⁵⁰) are sources of variability.

196 Maximum estimated severity was classified according to a five-level scale (Table 2). Scoring was
197 based on the estimated clinical state of animals at the most advanced stage of disease
198 progression they were allowed to reach. Studies in which information was insufficient to draw
199 conclusions about the level of severity were classified as 'undetermined". This severity scale was
200 developed building upon previous work from members of this team (NF, AO) developed for
201 classifying studies on mouse models of Huntington's disease (table 2 in ⁵¹), together with our
202 own (AG) experience with mutant SOD1 mouse models and literature. For purposes of statistical
203 analysis, the severity scale was reduced to a binary scale, ("low"= severity up to level 4; "high"=
204 level 5 severity. The choice for above level-4 severity as a cut-off point, was based on its status
205 as a "standard endpoint" in published ALS guidelines ^{2 39}, whereas full paralysis or spontaneous
206 death exceeds this standard endpoint, as well as the legally recommended endpoints in many
207 countries, including the EU Member States.

208

209

210

211

212

213

214 **Table 2.** Severity scale for ALS studies on transgenic mice with a mutant SOD1 gene. Each severity level
 215 exemplified from the most commonly used B6.Cg-TgN-(SOD1G93A) G1H mouse. Classification was based
 216 on the most severe endpoint used in each publication.
 217

Severity	Description	Welfare issues during this stage
Level 1	Animals euthanized prior to disease onset, which is characterised by progressive weight loss or hind limb tremors	No overt motor dysfunction. Phenotype is subclinical. Loss of motor function can be detected using rotarod or running wheels, but does not interfere with normal behaviour
Level 2	Studies terminated at an early stage of disease: animals present trembling and weakness in hind limbs (by approx. 75d) and mild body weight loss.	Minor. Loss of motor function can be detected using rotarod or running wheels, but has little interference with normal behaviour.
Level 3	Experiments terminated when animals are no longer able to reach food hopper or bottle spout. This occurs when animals reach a moderate (gait abnormalities and weakness) to severe (hind limb paralysis) stage of motor impairment (usually at 120-125d)	Medium. Loss of motor function and body weight can be detected by monitoring (e.g. by a clinical score sheet) and by checking self-righting ability. Refinement measures to address these welfare issues include provision of softer bedding material (e.g. sawdust), elongated bottle spouts and mashed food on the cage floor.
Level 4	Animals euthanized after losing the ability to right themselves within 10-30 seconds after being laid on either side (one or both) or when percentage of weight loss reaches 15-20% of peak body weight (usually at 130-140d)	Major. Animals show severe locomotor impairment. Refinement as described for level 3
Level 5	Animals are euthanized when reaching a moribund stage (complete paralysis) or allowed to die spontaneously	Severe. At this stage animals are unable to move, eat or drink. Animals which are not euthanized will die as a result of respiratory failure.

218

219 **2.3. Methodological Standards Reporting (MSR) and Regulatory Compliance Reporting (RCR)**
 220 **indexes**

221 For each reviewed publication, data were collected on a number of items which all contributed
 222 with information about the reporting quality of the paper. For the analysis, we brought these
 223 items together into two indexes, hence generating for each paper two comprehensive measures
 224 for reporting quality, one on methodological standards and one on regulatory compliance. We
 225 then used regression analysis to investigate how the two indexes (dependent variables) were
 226 related to publication year, type of study, country of origin and journal (explanatory or predictor
 227 variables), as outlined in detail in the following. Based on the regression models it is possible to
 228 predict how the dependent variables would have changed with changes in the explanatory
 229 variables. In contrast to, for example, correlation the regression analysis takes into account all
 230 the explanatory variables that were included in the models, i.e. the estimated association

231 between an index and one of the explanatory variables is independent of the values of the other
232 explanatory variables considered. In that way, spurious associations caused by relationships
233 between the explanatory variables in the data can be avoided.

234 The two indexes were formed as weighted sums of separate sets of items, adjusted after
235 checking for internal consistency. The Methodological Standards Reporting (MSR) index was
236 constructed from the items *sampsize*, *climate*, *cagesize*, *nmice*, *sex*, *copies*, *genetic* (which refer
237 to important research parameters in animal experimentation and in ALS research in particular)
238 and the items *random*, *blinded*, *control*, *sibsplit*, and *exclus* (associated with general good
239 practices in the design of animal experiments and published recommendations for ALS studies).
240 Greater weight (1.5 versus 1) was attributed to items which are also part of the ALS guidelines.
241 Table 3 describes these items, their attributed weight in the MSR index and the absolute number
242 and percentage of papers reporting this information, divided by type of study.

243 The Regulatory Compliance Reporting (RCR) index was originally constructed from the items
244 *comply*, *protocol*, *severity* (turned into a binary classification) and *refine*.

245

246 **Table 3. List of items integrated in the MSR and the RCR indexes for preclinical (n=108) and proof-of-**
247 **concept (n=461) animal studies on ALS reporting this information.** The index score for each variable is
248 provided (MSR index score ranging from 0 to 12.5, and RCR score ranging from 0 to 3). Greater weight
249 (1.5 versus 1) was attributed to items which are also part of the ALS guidelines. The internal consistency
250 reliability of each index was checked using Cronbach's alpha ⁵², estimating the degree to which the index
251 measures a latent construct. For MSR, alpha was 0.58, which indicates poor internal consistency, which
252 however does not disqualify the index. Omission of items *exclus*, *cagesize* and *blinded* increased alpha
253 only slightly, thus the original MSR index was retained. For RCR, alpha was 0.30, but increased to 0.42
254 when item *severe* was omitted. For purposes of statistical modelling, RCR (only including items *comply*,
255 *protocol* and *refine*) was later simplified to a binary variable RCRb coded as 1 for RCR values 2-3 and as 0
256 for RCR values 0-1.

257

258

259

260

261

262

263

264

265

Reported information	MSR index		'Proof-of-Concept' (n=461)		'Preclinical' (n=108)	
	Index item	Index weight	Absolute number	%	Absolute number	%
Relevant animal research variables						
Group size	<i>sampsize</i>	1.5	368	79.8	106	98.1
Environment: light, temp., humidity (fully or partially reported)	<i>climate</i>	1	123	26.7	42	38.9
Cage size	<i>cagesize</i>	1	1	0.2	2	1.9
Mice per cage	<i>nmice</i>	1	26	5.6	15	13.9
Sex of the animals	<i>sex</i>	1.5	223	48.4	71	65.7
Number of transgene copies	<i>copies</i>	1.5	286	62.0	80	74.1
Genetic background	<i>genetic</i>	1.5	349	75.7	92	85.2
Measures to reduce 'noise' and bias in experiments						
Animals randomised to treatment groups	<i>random</i>	1	28	6.1	47	43.5
Observers blinded to treatment	<i>blinded</i>	1.5	94	20.4	52	48.1
Non-transgenic littermate controls used	<i>control</i>	1	150	32.5	39	36.1
Splitting littermates into groups	<i>Sibsplit</i>	1	28	6.1	31	28.7
Reason for exclusion of animals is reported	<i>exclus</i>	1	2	0.4	6	5.6

Reported information	RCR index		'Proof-of-Concept' (n=461)		'Preclinical' (n=108)	
	Index item	Index weight	Absolute number	%	Absolute number	%
Self-reported compliance with laws and regulations	<i>comply</i>	1	98	21.3	28	25.9
Project approval reported	<i>protocol</i>	1	315	68.3	66	61.1
Refinement measures (e.g. to aid feed and hydrate)	<i>refine</i>	1	29	6.3	14	13

266

267

268 MSR and RCRb were modelled, estimating the effects of publication year (2005, 2007, 2009,
269 2011, 2013 or 2015) and study type (preclinical or proof-of-concept) while accounting for
270 possible confounding by country of origin (15 categories), journal (17 categories) and severity
271 (low or high). Countries contributing with less than twelve papers, and journals contributing with
272 less than 6 papers, were combined into separate categories, denoted 'Other'. MSR was
273 modelled using linear regression and RCR b by logistic regression. All first-order interaction
274 effects were tested and included if significant.

275 Predictive marginal means were calculated for all significant model effects. Both models were
276 checked using the Pregibon link test⁵³, and by examining standardised residuals. The MSR model
277 was also checked with the Breusch-Pagan/Cook-Weisberg test for heteroscedasticity⁵⁴, the
278 Ramsey regression specification-error test for omitted variables⁵⁵, and the RCR b model by
279 examining delta-betas to identify influential observations. The proportion of the total variation
280 in MSR and RCR b that could be explained by differences between countries or journals was
281 determined by running empty mixed models with country and journal, respectively, as a random
282 effect, and calculating the intra-class correlation coefficients.

283 The association between MSR and RCR indexes was estimated using Spearman rank correlation.

284 A total of 490 observations could be used. Overall MSR mean \pm SD was 5.69 ± 2.39 . RCR assumed
285 values 0 (n=48), 1 (n=103), 2 (n=309) or 3 (n=30), resulting in 69% of the observations having
286 values above 1. The number of observations per level of year, study type, country, journal and
287 severity is shown in Supplementary Table 1.

288 The data were analysed in Stata/IC v. 13.1 and IBM SPSS 23.0. Each article was regarded as the
289 experimental unit and the level of significance for all tests was 0.05.

290

291 **3. Results**

292 **3.1. Quality of research and reporting**

293 The quality of methodological standards and of reporting is crucial to avoid bias and achieve
294 reliable, repeatable and translatable research results. We measured this through the
295 Methodological Standards Reporting index and also looked at specific research parameters
296 individually.

297 **3.1.1 Methodological Standards Reporting index**

298 The 12 items that comprise the Methodological Standards Reporting Index represent seven
299 relevant experimental variables and five measures for reducing bias in animal experiments, most
300 of which are covered by ALS guidelines. Higher scores mean better reporting and
301 implementation of good practices in the design of ALS animal studies.

302 MSR was significantly affected by year and study type (joint F-test $p=0.0015$ and <0.0001 ,
303 respectively). Compared to 2007, the logistic regression model predicted a higher MSR for the
304 subsequent years (2009, 2011, 2013 and 2015) as well as for 2005 (Figure 3). It also predicted a
305 higher MSR for preclinical studies than for proof-of-concept studies (marginal mean 7.28 and
306 5.26 respectively). Supplementary Table 2 shows the MSR model results.

307

308 **Figure 3. Predictive marginal means of publication year (left panel) and country (right panel) based on**
309 **a model of a Methodological Standards Reporting (MSR) Index in 487 ALS studies (predicted index**
310 **values).** According to the linear regression model, MSR could be expected to be 0.74, 1.33, 1.50 and 1.18
311 units higher in 2009, 2011, 2013 and 2015 ($p=0.047$, 0.001 , 0.000 and 0.002), respectively, and 0.98 units
312 higher in 2005 ($p=0.011$). Sweden appeared to have comparatively low MSR values, while Germany, Italy
313 and Spain had somewhat high values. No significant interactions were found (e.g. between country and
314 year). According to the R-square statistic the model explained 25% of the total variation in MSR.

315

316 3.1.2. Reporting of relevant research parameters

317 Some research parameters were very seldom reported, for example: numbers of animals per
318 cage (7.2%, 41/569); cage size (0.5%, 3/569) and exclusion of animals (1.4%, 8/569). Measures
319 in guideline recommendations to reduce bias in ALS research were mostly not reported,
320 including: splitting littermates to treatment groups (10.4%, 59/569); use of non-transgenic
321 littermates as controls (33.2%, 189/569); as well as measures of broader application, such as
322 random assignment of animals to treatments (13.2%, 75/569) or blinding of observers (25.7%,
323 146/569). By contrast, numbers of transgene copies and genetic backgrounds of animals were
324 reported in the majority of papers.

325

326 Of papers reporting sex (n=294), 54.8% (161/294) described studies using mice of both sexes,
327 while 28.9% (57/194) used only males and 16.3% (35/194) used only females. Reporting of sex
328 rose steadily from 2005 (39.0%, 30/77) to 2015 69.8% (74, 106), $\chi^2(5) = 30,831$, $p < 0.0001$,
329 linear-by-linear association=27.802, $p < 0.0001$).

330 Regarding the chosen genetic background of animals used for preclinical studies (n= 108), 76%
331 (70/92) of those reporting this parameter generated experimental animals using a cross
332 between mice hemizygous for the SOD1 mutant gene and C57/SJL outbred strains.

333

334 Only ten studies (6 proof-of-concept studies and 4 preclinical studies) from 2007, 2009, 2011,
335 2013, and 2015 justified the number of animals used per group. However, of these, only six gave
336 clear justifications (five justified the group size by a power analysis and the other by the size of
337 groups proposed in ALS guidelines^{2 39}. On the other hand, group size was reported in 83.3%
338 (474/569) of ALS papers, and more so in the preclinical studies sub-sample (Fig. 4).

339

340

341 **Figure 4. Group size.** Histogram of mean group size in 105 preclinical studies reporting this parameter
342 (left) and for each of the years analysed (yearly mean \pm 1 standard deviation) (right).

343

344

345 Of the 569 papers reviewed, 38% (214/569) did not report the method for killing animals despite
346 the fact that in 91% (195/214) of these, terminal procedures requiring anaesthesia for ethical
347 and practical reasons were identified (e.g. transcatheter perfusion fixation). The most commonly
348 used euthanasia method – of the papers reporting this information – was anaesthetic overdose
349 or the use of another method under anaesthesia (86%; 317/367) while other methods such as

350 CO₂ asphyxiation (7%; 26/367) or others such as decapitation or cervical dislocation (7%; 24/367)
351 were seldom used. Very few studies (15 out of 569) were identified as not performing
352 euthanasia of any kind. The remaining 21 articles were deemed “inconclusive”, for neither
353 reporting euthanizing animals at any point nor reporting deaths.

354

355 3.2 Regulatory compliance and estimated severity

356 For public confidence in research, it is important that research with animals is carried out
357 according to standards set by legislation and in line with the principles of the 3Rs. We measured
358 such compliance through the Regulatory Compliance Reporting index and also looked at specific
359 research parameters individually.

360

361 3.2.1. Regulatory compliance reporting index (RCR)

362 The Regulatory Compliance Reporting (RCR) index, which measures to what extent compliance
363 with legislation and approval of animal experiments are reported in published papers, shows an
364 overall improvement in the reporting over the time period under study (joint Chi-square
365 $p < 0.001$, Figure 5). RCR did not differ between journals or between proof-of-concept and
366 preclinical studies but was affected by country (Figure 5). Studies with high severity seemed to
367 have higher odds of high RCR values ($p = 0.027$). Supplementary Table 3 shows the RCR model
368 results.

369

370

371

372 **Figure 5. Predictive marginal means of publication year (left panel) and country (right panel) based on**
373 **a model of a Regulatory Compliance Reporting (RCR) index in 490 ALS studies (predicted probabilities**
374 **of index values >1).** The odds of an RCR index above 1 was 3.43 and 7.07 times higher in 2013 and 2015
375 ($p = 0.003$ and 0.000), respectively, than in 2005. China, France, Italy and South Korea appeared to have
376 comparatively low probabilities, while for example Spain, Belgium and Canada had somewhat high
377 probabilities. No significant interactions were found. The pseudo R-square statistic indicated that the
378 model explained 16% of the total variation in the data.

379

380

381 Over the entire period, most papers (67.0%; 381/569) reported that studies had been appraised
382 and approved by a third party (e.g. ethics committee, competent authority) with only 10.9%
383 (62/382) not reporting any kind of regulatory compliance. By 2015, all papers were found to

384 have some type of statement on regulatory compliance, most of which (83%) referring to prior
385 ethical approval of research protocols.

386

387 The correlation between MSR and RCR was weak, but highly significant ($\rho=0.21$; $p<0.0001$)
388 indicating that papers with high scores for methodological standards were somewhat more
389 likely to also score highly for regulatory standards.

390 3.2.2 Severity and refinement measures

391 We have found in previous systematic reviews⁶⁵⁶ that self-reported compliance with regulations
392 may not necessarily affect the severity of the experiments being conducted. To test whether
393 actual experimental practice has changed over the study period, we classified the severity of
394 each study according to the criteria in Table 2. The majority of publications (60.7%) (346/569)
395 included experiments at level-4 severity (Figure 6-A). Of the 64 studies classified as Level 5
396 (allowing animals to die of disease progression or to reach complete paralysis), 89% reported
397 regulatory compliance (70% ethical approval from a national authority or institutional ethics
398 committee and 19% compliance with relevant legislation or animal use guidelines). However,
399 between those studies that reported regulatory compliance and those that did not, there was
400 no difference in the proportion that were Level 5 ($\chi^2(5) = 2.855$, $p = 0.722$) (Figure 6-B).

401 On the other hand, we did observe a difference between preclinical and proof-of-concept
402 studies: preclinical studies included a higher proportion of studies within the highest severity
403 categories (77.9% (81/104) classified as level 4 and 19.2% (20/104) as level 5) than did proof-of-
404 concept studies (68.7% (265/386) classified as level 4 and 11.4% (44/386) as level 5). Moreover,
405 no preclinical studies were given a level 1 or level 2 severity ($\chi^2(5) = 19.593$, $p = 0.001$) (Figure
406 6-C).

407

408

409 **Figure 6. Severity classification of studies (N=569).** Figure 6-A illustrates the percentage of studies, by
410 year, classified into each of the 5-levels of our severity scale, as well as those of "undetermined" severity
411 due to insufficient information ($n = 77$ in 2005; $n=81$ in 2007; $n = 84$ in 2009; $n = 106$ in 2011; $n = 115$ in
412 2013; $n= 106$ in 2015. Figures 6-B and 6-C show percentage of studies classified into each of the 5-levels,
413 according to, respectively, reported regulatory compliance status ($n = 62$, not reported; $n = 126$, guidelines
414 followed; $n = 381$, protocol approval), and type of study ($n = 461$, proof-of-concept studies; $n = 108$,
415 preclinical studies).

416

417 Of studies classified between levels 3 and 5 severity (i.e. from which it could be ascertained
418 animals presented overt locomotor impairments), only 9.1% (42/456) described any refinement
419 measures to alleviate suffering (e.g. provision of mashed food and adaptation of bedding in later
420 stages of disease progression), which occurred almost exclusively (39/42) in Level 4 studies.
421 Differences in the regulatory landscape between countries imply that *how* animals are treated
422 in biomedical research may depend on *where* these experiments are carried out. The proportion
423 of high-severity (Level-5) studies differed significantly ($\chi^2(13) = 35,561, p=0.001$) between the 14
424 most represented countries in our sample, ranging from 40% (8/20) and 41% (7/17) – in South
425 Korea and Israel, respectively – to 4% in Canada and China and even none in Belgium (0/14) and
426 the UK (0/23).

427

428 **4. Discussion**

429 Our analysis, the first of its kind to use specially devised indexes encompassing both
430 methodological standards and regulatory compliance reporting (MSR and RCR, respectively)
431 over a 10 year period, suggest three main findings: The first is an overall improvement in both
432 regulatory compliance and methodological and reporting quality across the period assessed.
433 Also, and somewhat as expected, studies classified as 'preclinical' scored higher for
434 methodological and reporting quality as compared with more 'proof-of-concept' studies. The
435 third finding is that scores for both indexes varied widely according to the country in which the
436 first author was based, but not according to the journal publishing the paper.

437 The improved reporting of regulatory compliance, as expressed in the increase in RCR score
438 across time, is an indicator of widespread increase in reported adherence to animal welfare
439 regulatory requirements. However, this was not reflected in any significant change in the
440 proportion of highly severe (Level-5 in our classification scheme) studies or the reporting of
441 refinement measures (in studies where animals showed overt clinical signs). This is in agreement
442 with results from previous systematic reviews of animal research on Huntington's disease
443 (papers published 1997-2009)⁵¹ and tuberculosis (1997-2011)⁵⁶. Also, while 'preclinical' studies
444 were more likely to be classified in the higher severity categories, there was no relation between
445 the level of severity and whether papers reported approval of protocols or compliance with
446 regulations, the latter also reflecting previous findings^{6,57}.

447

448 Only 11.2% of ALS studies were classified at the highest severity level (level 5, i.e. including
449 experiments with spontaneous death or euthanasia at a near-death stage, i.e. complete

450 paralysis), which is much lower than that found in research using mouse models of Huntington's
451 Disease (38%)⁵¹ and Tuberculosis (66%)⁵⁶. Moreover, most endpoints applied in ALS studies
452 adhered to the same basic criterion for euthanizing animals, namely the point at which animals
453 are unable to resume their position if laid recumbent within 10-30 seconds. This is the primary
454 endpoint proposed in existing guidelines for preclinical ALS^{2 39} and the ALS Treatment
455 Development Institute's recommendations²⁹ (level-4 severity on our scale), suggesting
456 widespread compliance to published guidance in this respect. However, this endpoint was
457 already broadly used before the publication of the guidelines suggesting that these reflect
458 common practice at the time of publication.

459

460 Applying predefined endpoints is important to prevent the loss of biological samples from
461 animals found dead and for which time of death therefore cannot be defined⁶, hence
462 maintaining numbers of animals and avoiding loss of statistical power and subsequent
463 inconclusive results. However, from an animal welfare perspective, the current standard
464 endpoint for ALS studies corresponds to an end-stage where euthanasia may prevent deaths
465 from respiratory failure, but since they seldom anticipate death by more than a day, or even just
466 a few hours, late stage endpoints only curtail a small part of animal suffering⁸. Very late
467 endpoints increase the likelihood that at least some animals will die unsupervised (e.g.
468 overnight), while the confounding effect of starvation and dehydration in survival data increases
469 as animals become progressively less able to reach the bottle spout or the food hopper⁵. At
470 advanced clinical stages, refinements such as providing mashed food on the cage floor, long-
471 spouted water bottles or fluid administration are therefore crucial to avoid unnecessary animal
472 suffering and to improve validity by bringing the model closer to the clinical setting, where late-
473 stage human patients are provided palliative care⁵⁸. Defining endpoints also needs to take the
474 research purpose into account. In ALS, the mechanisms operating at different stages of the
475 disease are known to be different, principally affecting distal axons at the onset of symptoms,
476 but developing an immune/inflammatory phenotype during the end stages⁵⁹. Therefore,
477 endpoints relevant to the treatment strategies must be used, particularly when targeting
478 neuroinflammation.

479

480 Methodological standards reporting improved over the time period under study. Studies
481 classified as 'preclinical' reported methodology in more detail than those deemed 'proof-of-
482 concept', consistent with the view that a more rigorous design and execution should be
483 demanded for preclinical studies⁶⁰. Nevertheless, the checklist provided in the 2010 edition of
484 the guidelines for ALS research sets high methodological standards for both types of studies².

485 Throughout the period under study, the MSR scores remain below 50% of the maximum score,
486 showing that the overall level of reporting of methodological detail did not change and remain
487 substantially below the recommendations in the guidelines.

488

489 Only three parameters (genetic background, number of transgene copies, and group size) were
490 reported in more than half of the sample, whereas other relevant information, such as housing
491 conditions, randomisation of animals into treatment groups or blinding of researchers was
492 absent in well over two thirds of the papers analysed, in line with previous reviews of animal
493 research in the neurosciences^{5 51 61}. Other biological and methodological parameters, such as
494 sex (only reported in the majority of papers in the "preclinical studies" sub-sample) and method
495 of choice for euthanizing animals were also largely under-reported. The method used for
496 euthanizing animals has both animal welfare implications and scientific relevance, as the
497 method affects biological and histological parameters differently, which can impact the *post*
498 *mortem* data collected^{62 63}. The increase in the proportion of articles in our sample reporting
499 sex of the animals is positive, as sex differences^{4 64-66} in the phenotype or response to
500 therapeutical drugs may influence results and be of clinical relevance. However, although ALS
501 guidelines propose the use of both male and female mice, little over half of the studies providing
502 this information reported doing so. Overall, making these and other details on animals and
503 protocol available is central to allowing an adequate interpretation of results and a critical
504 evaluation of their validity, as well as allowing study replication and proper integration of results
505 in systematic reviews and meta-analyses^{32 67}.

506

507 Sample size was generally well reported, but of those reporting this parameter, only a small
508 minority used the 24 per group recommended in the 2010 guidelines². Furthermore, only three
509 studies clearly justified group size, in agreement with previous reports that this is frequently
510 overlooked e.g.^{32 68}. Adequate sample size is paramount to ensure that animals, time and
511 resources are not wasted as a result of underpowering experiments by using too few animals⁶⁹
512⁷⁰. Noise reduction by genetic standardisation could also help reduce the number of animals
513 needed per study, as the reduced inter-individual variability of isogenic strains allows increasing
514 power without requiring more animals⁷¹ and is indeed mentioned in the 2007 guidelines as a
515 way of reducing variability in drug testing³⁹. Mead and colleagues⁷², for instance, have shown
516 great consistency of results by using SOD1G93A transgenic mice on an inbred C57BL/6 genetic
517 background, with the added advantage of presenting early indicators of disease progress,
518 allowing for faster and more humane drug screening. Only 11% of the preclinical studies
519 reviewed, however, used a fully inbred background. The use of a single well characterised model

520 for initial studies can be supported further by independent replication studies in a different
521 disease model.

522

523 Most articles did not report random assignment of animals to groups or blinded outcome
524 assessment. This reflects similar data from reviews on the methodological quality of preclinical
525 research on ALS^{29 61 73} and other fields^{7 32 34 74 75}. This lack of attention to measures to avoid noise
526 and biases in animal experiments is cause for concern, given their role in improving the reliability
527 of results, as well as the translational value of preclinical research^{17 25 34 70 74}. While it cannot be
528 excluded that in some cases blinding and randomisation were applied but not reported, one
529 might expect that researchers carrying out well thought-out and planned experiments would
530 state such measures, since this strengthens their results and conclusions. There is ample
531 evidence for ALS^{29 76} and other areas^{34 75 77-79} that published studies which do not report
532 measures to minimise bias (i.e. blinding, randomisation and allocation concealment) tend to
533 present an exaggerated estimate of the therapeutic effect of experimental drugs. This is
534 particularly relevant in the light of the ongoing discussion of why promising pre-clinical results
535 of candidate drugs for ALS have not translated into the clinic. Although the disappointing
536 outcomes of clinical trials apparently contradict the promising preclinical results that elicited
537 them, they may actually mirror the results obtained from adequately designed animal studies
538 carried out to high methodological standards^{29 76}.

539

540 Methodological standards reporting and regulatory compliance reporting scores were not
541 influenced by the journal in which the results were published. Other researchers who have
542 investigated the effect of journal on methodological standards and reporting quality have found
543 a statistically significant but very small effect of whether or not the journal had endorsed the
544 ARRIVE guidelines.^{80 81}

545

546 By contrast, we found an improvement in these scores over time. However, it is difficult to say
547 to what extent this is the result of the field-specific guidelines, as there is an increasing trend
548 overall in the methodological standards and regulatory compliance reporting scores. Our study,
549 of course, is limited to the period and model under study and some improvements may have
550 occurred as a result of the informal discussion leading up to the formal workshops and guidelines
551 (and more recently the appearance of other transgenic models means that the study does not
552 cover the entire field of ALS research for later years). Also, a surprisingly low number of papers
553 (1/84 in 2009, 10/106 in 2011, 10/115 in 2013 and 14/106 in 2015) referred to the Ludolph *et*

554 *al.* guidelines^{2 39}. Given the slow adoption of the ARRIVE guidelines⁸², it seems likely it may also
555 take some time for the ALS guidelines to have a detectable effect.

556 While reporting of relevant parameters such as blinding and randomisation was higher in our
557 'preclinical' subsample than what has been reported in other systematic reviews^{17 32 80 82-85},
558 results for the overall sample were generally comparable. Also, and similarly to what was found
559 in these systematic reviews, justification for sample size was rarely reported.

560

561 One way of addressing the problems with study quality could be for preclinical researchers to
562 adopt the standards of randomised controlled trials in humans⁸⁶⁻⁸⁹, including trial pre-
563 registration^{90 91}. Compliance with existing guidelines would seem a more readily achievable goal,
564 however other self-regulatory mechanisms may be warranted to improve compliance, such as
565 changes to the publishing requirements of biomedical journals⁹²⁻⁹⁴ or more demanding
566 requirements by science funders, both of which are clearly on the horizon^{31 95}.

567

568 **5. Conclusion**

569 The ALS research community pioneered the development of field-specific guidelines, setting
570 science community-based standards for animal research methodology and reporting^{2 39}.

571 Whereas we found significant improvement over time, it is less clear to what extent this is linked
572 to the guidelines, which are rarely referred to. Animal research in the field of ALS does however
573 differ from comparable research in other reviewed fields in one aspect: the implementation of
574 predefined endpoints in studies of advanced disease stages. This practice is important both for
575 research quality and animal welfare and is indeed coherent with the field-specific guidelines.
576 We propose that future guidelines should address measures to raise standards in the design,
577 conduct and reporting of experiments as well as to reduce the impact on animal welfare, as part
578 of a concerted effort to make biomedical research using animals more ethically and socially
579 acceptable and effective.

580

581 **Acknowledgements**

582 We thank Gilly Griffin for her input on current practice regarding humane endpoints in Canada.

583

584 **Funding**

585 NHF was a recipient of a Post-doctoral Research Fellowship from the Portuguese Foundation for
586 Science and Technology (FCT), grant reference SFRH/BPD/85978/2012. The research leading to

587 these results has received funding from the European Union Seventh Framework Programme
588 [FP7-HEALTH-2013-INNOVATION-1] under grant agreement n. ° 602616 [Project ANIMPACT]
589 and from the project Norte-01-0145-FEDER-000008 - Porto Neurosciences and Neurologic
590 Disease Research Initiative at I3S , supported by Norte Portugal Regional Operational
591 Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, through the
592 European Regional Development Fund (FEDER)

593

594 **Competing Interests**

595 We have read and understood BMJ policy on declaration of interests and declare that we have
596 no competing interests.

597

598 **Role of Authors and Contributors**

599

600 Original idea for this study: NF, AO

601 Conception and design of the work: NF, AO, AJF, AJG

602 Data collection: JF, NF

603 Data analysis and interpretation: NF, JF, JH, AJF, AJG, AO

604 Drafting the article: NF, JF

605 Critical revision of the article: AJG, AJF, JH, AO

606 Final approval of the version to be published: NF, JF, AJG, AJF, JH, AO

607

608 **Data access statement**

609 Dataset will be made available upon publication, at the University of Porto repository.

610

611

612 **References**

- 613 1. Miller RG, Mitchell JD, Lyon M, et al. Riluzole for amyotrophic lateral sclerosis (ALS)/motor
614 neuron disease (MND). Amyotrophic lateral sclerosis and other motor neuron disorders
615 : official publication of the World Federation of Neurology, Research Group on Motor
616 Neuron Diseases 2003;**4**(3):191-206.
- 617 2. Ludolph AC, Bendotti C, Blaugrund E, et al. Guidelines for preclinical animal research in
618 ALS/MND: A consensus meeting. Amyotroph Lateral Scler 2010;**11**(1-2):38-45.
- 619 3. Shibata N. Transgenic mouse model for familial amyotrophic lateral sclerosis with superoxide
620 dismutase-1 mutation. Neuropathology 2001;**21**(1):82-92.

- 621 4. Heiman-Patterson TD, Deitch JS, Blankenhorn EP, et al. Background and gender effects on
622 survival in the TgN(SOD1-G93A)1Gur mouse model of ALS. *Journal of the Neurological*
623 *Sciences* 2005;**236**(1–2):1-7.
- 624 5. Olsson IAS, Hansen AK, Sandoe P. Animal welfare and the refinement of neuroscience
625 research methods - a case study of Huntington's disease models. *Lab Anim*
626 2008;**42**(3):277-83.
- 627 6. Franco NH, Olsson IA. "How sick must your mouse be? " - An analysis of the use of animal
628 models in Huntington's disease research. *Alternatives to laboratory animals : ATLA*
629 2012;**40**(5):271-83.
- 630 7. Bara M, Joffe AR. The methodological quality of animal research in critical care: the public
631 face of science. *Annals of intensive care* 2014;**4**(1):1-9.
- 632 8. Franco NH, Correia-Neves M, Olsson IAS. How "humane" is your endpoint?—Refining the
633 science-driven approach for termination of animal studies of chronic infection. *PLoS*
634 *pathogens* 2012;**8**(1):e1002399.
- 635 9. Solomon JA, Tarnopolsky MA, Hamadeh MJ. One universal common endpoint in mouse
636 models of amyotrophic lateral sclerosis. *PLoS one* 2011;**6**(6):e20582.
- 637 10. Morton DB. Humane endpoints in animal experimentation for biomedical research: ethical,
638 legal and practical aspects: London: Royal Society of Medicine Press, 1999:5-12.
- 639 11. Sawiak S, Wood N, Williams G, et al. Use of magnetic resonance imaging for anatomical
640 phenotyping of the R6/2 mouse model of Huntington's disease. *Neurobiology of Disease*
641 2009;**33**(1):12-19.
- 642 12. Hockly E, Woodman B, Mahal A, et al. Standardization and statistical approaches to
643 therapeutic trials in the R6/2 mouse. *Brain research bulletin* 2003;**61**(5):469-79.
- 644 13. Menalled L, Brunner D. Animal models of Huntington's disease for translation to the clinic:
645 Best practices. *Movement Disorders* 2014;**29**(11):1375-90.
- 646 14. van der Worp HB, Howells DW, Sena ES, et al. Can animal models of disease reliably inform
647 human studies? *PLoS Medicine* 2010;**7**(3):e1000245.
- 648 15. Ioannidis JP. Why most published research findings are false. *PLoS Medicine* 2005;**2**(8):e124.
- 649 16. Schnabel J. Neuroscience: standard model. *Nature News* 2008;**454**(7205):682-85.
- 650 17. Ioannidis JPA, Greenland S, Hlatky MA, et al. Increasing value and reducing waste in research
651 design, conduct, and analysis. *The Lancet* 2014;**383**(9912):166-75.
- 652 18. Festing MFW. Randomized Block Experimental Designs Can Increase the Power and
653 Reproducibility of Laboratory Animal Experiments. *ILAR Journal* 2014;**55**(3):472-76.
- 654 19. Garner JP. The Significance of Meaning: Why Do Over 90% of Behavioral Neuroscience
655 Results Fail to Translate to Humans, and What Can We Do to Fix It? *ILAR Journal*
656 2014;**55**(3):438-56.
- 657 20. Lapchak PA. Scientific Rigor Recommendations for Optimizing the Clinical Applicability of
658 Translational Research. *Journal of neurology & neurophysiology* 2012;**3**.
- 659 21. Steward O, Balice-Gordon R. Rigor or Mortis: Best Practices for Preclinical Research in
660 Neuroscience. *Neuron* 2014;**84**(3):572-81.
- 661 22. Tsilidis KK, Panagiotou OA, Sena ES, et al. Evaluation of excess significance bias in animal
662 studies of neurological diseases. *PLoS biology* 2013;**11**(7):e1001609.
- 663 23. Button KS, Ioannidis JPA, Mokrysz C, et al. Power failure: why small sample size undermines
664 the reliability of neuroscience. *Nat Rev Neurosci* 2013;**14**(5):365-76.
- 665 24. Vesterinen HM, Sena ES, French-Constant C, et al. Improving the translational hit of
666 experimental treatments in multiple sclerosis. *Multiple Sclerosis Journal*
667 2010;**16**(9):1044-55.
- 668 25. Sena ES, van der Worp HB, Bath PMW, et al. Publication Bias in Reports of Animal Stroke
669 Studies Leads to Major Overstatement of Efficacy. *PLoS Biol* 2010;**8**(3):e1000344.
- 670 26. Steward O, Popovich PG, Dietrich WD, et al. Replication and reproducibility in spinal cord
671 injury research. *Exp Neurol* 2012;**233**(2):597-605.

- 672 27. Shineman DW, Basi GS, Bizon JL, et al. Accelerating drug discovery for Alzheimer's disease:
673 best practices for preclinical animal studies. *Alzheimers Res Ther* 2011;**3**(5):28.
- 674 28. Kimmelman J, London AJ, Ravina B, et al. Launching invasive, first-in-human trials against
675 Parkinson's disease: Ethical considerations. *Movement Disorders* 2009;**24**(13):1893-
676 901.
- 677 29. Scott S, Kranz JE, Cole J, et al. Design, power, and interpretation of studies in the standard
678 murine model of ALS. *Amyotroph Lateral Scler* 2008;**9**(1):4-15.
- 679 30. Collins FS, Tabak LA. NIH plans to enhance reproducibility. *Nature* 2014;**505**(7485):612.
- 680 31. Cressey D. UK funders demand strong statistics for animal studies. *Nature*
681 2015;**520**(7547):271.
- 682 32. Kilkeny C, Parsons N, Kadyszewski E, et al. Survey of the Quality of Experimental Design,
683 Statistical Analysis and Reporting of Research Using Animals. *PloS one*
684 2009;**4**(11):e7824.
- 685 33. Crossley NA, Sena E, Goehler J, et al. Empirical evidence of bias in the design of experimental
686 stroke studies: a metaepidemiologic approach. *Stroke; a journal of cerebral circulation*
687 2008;**39**(3):929-34.
- 688 34. Hirst JA, Howick J, Aronson JK, et al. The need for randomization in animal trials: an overview
689 of systematic reviews. *PloS one* 2014;**9**(6):e98856.
- 690 35. van der Worp HB, Sena ES, Donnan GA, et al. Hypothermia in animal models of acute
691 ischaemic stroke: a systematic review and meta-analysis. *Brain : a journal of neurology*
692 2007;**130**(Pt 12):3063-74.
- 693 36. von Roten FC. Public perceptions of animal experimentation across Europe. *Public*
694 *Understanding of Science* 2012.
- 695 37. Lund TB, Mørkbak MR, Lassen J, et al. Painful dilemmas: A study of the way the public's
696 assessment of animal research balances costs to animals against human benefits. *Public*
697 *Understanding of Science* 2014;**23**(4):428-44.
- 698 38. Rollin BE. *Science and Ethics*: Cambridge University Press, 2006.
- 699 39. Ludolph AC, Bendotti C, Blaugrund E, et al. Guidelines for the preclinical in vivo evaluation of
700 pharmacological active drugs for ALS/MND: report on the 142nd ENMC international
701 workshop. *Amyotrophic lateral sclerosis : official publication of the World Federation of*
702 *Neurology Research Group on Motor Neuron Diseases* 2007;**8**(4):217-23.
- 703 40. Kanning KC, Kaplan A, Henderson CE. Motor neuron diversity in development and disease.
704 *Annual review of neuroscience* 2010;**33**:409-40.
- 705 41. Lever TE, Gorsek A, Cox KT, et al. An animal model of oral dysphagia in amyotrophic lateral
706 sclerosis. *Dysphagia* 2009;**24**(2):180-95.
- 707 42. Lever TE, Simon E, Cox KT, et al. A mouse model of pharyngeal dysphagia in amyotrophic
708 lateral sclerosis. *Dysphagia* 2010;**25**(2):112-26.
- 709 43. Boylan K, Yang C, Crook J, et al. Immunoreactivity of the phosphorylated axonal
710 neurofilament H subunit (pNF-H) in blood of ALS model rodents and ALS patients:
711 evaluation of blood pNF-H as a potential ALS biomarker. *Journal of neurochemistry*
712 2009;**111**(5):1182-91.
- 713 44. Kato S, Kato M, Abe Y, et al. Redox system expression in the motor neurons in amyotrophic
714 lateral sclerosis (ALS): immunohistochemical studies on sporadic ALS, superoxide
715 dismutase 1 (SOD1)-mutated familial ALS, and SOD1-mutated ALS animal models. *Acta*
716 *neuropathologica* 2005;**110**(2):101-12.
- 717 45. Gurney ME, Pu HF, Chiu AY, et al. Motor-neuron degeneration in mice that express a human
718 Cu,Zn superoxide-dismutase mutation. *Science* 1994;**264**(5166):1772-75.
- 719 46. Buijn LI, Becher MW, Lee MK, et al. ALS-linked SOD1 mutant G85R mediates damage to
720 astrocytes and promotes rapidly progressive disease with SOD1-containing inclusions.
721 *Neuron* 1997;**18**(2):327-38.

- 722 47. Marcuzzo S, Zucca I, Mastropietro A, et al. Hind limb muscle atrophy precedes cerebral
723 neuronal degeneration in G93A-SOD1 mouse model of amyotrophic lateral sclerosis: a
724 longitudinal MRI study. *Exp Neurol* 2011;**231**(1):30-7.
- 725 48. Neymotin A, Calingasan NY, Wille E, et al. Neuroprotective effect of Nrf2/ARE activators,
726 CDDO ethylamide and CDDO trifluoroethylamide, in a mouse model of amyotrophic
727 lateral sclerosis. *Free Radical Biology and Medicine* 2011;**51**(1):88-96.
- 728 49. Del Signore SJ, Amante DJ, Kim J, et al. Combined riluzole and sodium phenylbutyrate therapy
729 in transgenic amyotrophic lateral sclerosis mice. *Amyotrophic lateral sclerosis : official
730 publication of the World Federation of Neurology Research Group on Motor Neuron
731 Diseases* 2009;**10**(2):85-94.
- 732 50. Tada S, Okuno T, Yasui T, et al. Deleterious effects of lymphocytes at the early stage of
733 neurodegeneration in an animal model of amyotrophic lateral sclerosis. *Journal of
734 neuroinflammation* 2011;**8**(1):19.
- 735 51. Franco NH, Olsson I. " How sick must your mouse be?"-An analysis of the use of animal
736 models in Huntington's disease research. *Alternatives to laboratory animals: ATLA*
737 2012;**40**(5):271-83.
- 738 52. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika*
739 1951;**16**(3):297-334.
- 740 53. Pregibon D. Goodness of link tests for generalized linear models. *Applied statistics* 1980:15-
741 14.
- 742 54. Breusch TS, Pagan AR. A simple test for heteroscedasticity and random coefficient variation.
743 *Econometrica: Journal of the Econometric Society* 1979:1287-94.
- 744 55. Ramsey JB. Tests for specification errors in classical linear least-squares regression analysis.
745 *Journal of the Royal Statistical Society Series B (Methodological)* 1969:350-71.
- 746 56. Franco NH, Correia-Neves M, Olsson IAS. Animal Welfare in Studies on Murine Tuberculosis:
747 Assessing Progress over a 12-Year Period and the Need for Further Improvement. *PLoS
748 one* 2012;**7**(10):e47723.
- 749 57. Franco NH, Olsson I. Is the ethical appraisal of protocols enough to ensure best practice in
750 animal research? *Alternatives to laboratory animals: ATLA* 2013;**41**(1):P5-7.
- 751 58. Lilley E, Hawkins P, Jennings M. A 'road map' toward ending severe suffering of animals used
752 in research and testing. *ATLA - Alternatives to Laboratory Animals* 2014;**42**(4):267-72.
- 753 59. Boillée S, Yamanaka K, Lobsiger CS, et al. Onset and progression in inherited ALS determined
754 by motor neurons and microglia. *Science* 2006;**312**(5778):1389-92.
- 755 60. Kimmelman J, Mogil JS, Dirnagl U. Distinguishing between Exploratory and Confirmatory
756 Preclinical Research Will Improve Translation. *PLoS biology* 2014;**12**(5):e1001863.
- 757 61. Benatar M. Lost in translation: Treatment trials in the SOD1 mouse and in human ALS.
758 *Neurobiology of Disease* 2007;**26**(1):1-13.
- 759 62. Reilly J, Blackshaw AW. *Euthanasia of animals used for scientific purposes*: ANZCCART, 2001.
- 760 63. Artwohl J, Brown P, Corning B, et al. Report of the ACLAM Task Force on Rodent Euthanasia.
761 *Journal of the American Association for Laboratory Animal Science* 2006;**45**(1):98-105.
- 762 64. Bame M, Pentiak PA, Needleman R, et al. Effect of Sex on Lifespan, Disease Progression, and
763 the Response to Methionine Sulfoximine in the SOD1 G93A Mouse Model for ALS.
764 *Gender Medicine* 2012;**9**(6):524-35.
- 765 65. McCombe PA, Henderson RD. Effects of gender in amyotrophic lateral sclerosis. *Gender
766 Medicine* 2010;**7**(6):557-70.
- 767 66. Alves CJ, de Santana LP, Santos AJDd, et al. Early motor and electrophysiological changes in
768 transgenic mouse model of amyotrophic lateral sclerosis and gender differences on
769 clinical outcome. *Brain Research* 2011;**1394**(0):90-104.
- 770 67. Hooijmans CR, Leenaars M, Ritskes-Hoitinga M. A gold standard publication checklist to
771 improve the quality of animal studies, to fully integrate the Three Rs, and to make
772 systematic reviews more feasible. *Alternatives to laboratory animals : ATLA*
773 2010;**38**(2):167-82.

- 774 68. Banwell V, Sena ES, Macleod MR. Systematic review and stratified meta-analysis of the
775 efficacy of interleukin-1 receptor antagonist in animal models of stroke. *Journal of*
776 *stroke and cerebrovascular diseases : the official journal of National Stroke Association*
777 2009;**18**(4):269-76.
- 778 69. Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research
779 evidence. *Lancet* 2009;**374**(9683):86-9.
- 780 70. Festing MF, Altman DG. Guidelines for the design and statistical analysis of experiments using
781 laboratory animals. *ILAR journal / National Research Council, Institute of Laboratory*
782 *Animal Resources* 2002;**43**(4):244-58.
- 783 71. Festing MFW. Warning: the use of heterogeneous mice may seriously damage your research.
784 *Neurobiology of Aging* 1999;**20**(2):237-44.
- 785 72. Mead RJ, Bennett EJ, Kennerley AJ, et al. Optimised and Rapid Pre-clinical Screening in the
786 SOD1^{G93A} Transgenic Mouse Model of Amyotrophic Lateral Sclerosis
787 (ALS). *PloS one* 2011;**6**(8):e23244.
- 788 73. Perrin S. Preclinical research: Make mouse studies work. *Nature* 2014;**507**:423-25.
- 789 74. Landis SC, Amara SG, Asadullah K, et al. A call for transparent reporting to optimize the
790 predictive value of preclinical research. *Nature* 2012;**490**(7419):187-91.
- 791 75. Howells DW, Macleod MR. Evidence-based Translational Medicine. *Stroke; a journal of*
792 *cerebral circulation* 2013;**44**(5):1466-71.
- 793 76. Perrin S. Preclinical research: Make mouse studies work. *Nature* 2014;**507**(7493):423-25.
- 794 77. Bebarta V, Luyten D, Heard K. Emergency medicine animal research: does use of
795 randomization and blinding affect the results? *Academic Emergency Medicine*
796 2003;**10**(6):684-87.
- 797 78. Macleod MR, van der Worp HB, Sena ES, et al. Evidence for the efficacy of NXY-059 in
798 experimental focal cerebral ischaemia is confounded by study quality. *Stroke; a journal*
799 *of cerebral circulation* 2008;**39**(10):2824-29.
- 800 79. Crossley NA, Sena E, Goehler J, et al. Empirical Evidence of Bias in the Design of Experimental
801 Stroke Studies A Metaepidemiologic Approach. *Stroke* 2008;**39**(3):929-34.
- 802 80. Avey MT, Moher D, Sullivan KJ, et al. The Devil Is in the Details: Incomplete Reporting in
803 Preclinical Animal Research. *PloS one* 2016;**11**(11):e0166733.
- 804 81. Vogt L, Reichlin TS, Nathues C, et al. Authorization of Animal Experiments Is Based on
805 Confidence Rather than Evidence of Scientific Rigor. *PLoS biology*
806 2016;**14**(12):e2000598.
- 807 82. Baker D, Lidster K, Sottomayor A, et al. Two years later: journals are not yet enforcing the
808 ARRIVE guidelines on reporting standards for pre-clinical animal studies. *PLoS biology*
809 2014;**12**(1):e1001756.
- 810 83. Gulin JEN, Rocco DM, García-Bournissen F. Quality of Reporting and Adherence to ARRIVE
811 Guidelines in Animal Studies for Chagas Disease Preclinical Drug Research: A Systematic
812 Review. *PLOS Neglected Tropical Diseases* 2015;**9**(11):e0004194.
- 813 84. Macleod MR, Lawson McLean A, Kyriakopoulou A, et al. Risk of Bias in Reports of In Vivo
814 Research: A Focus for Improvement. *PLoS biology* 2015;**13**(10):e1002273.
- 815 85. Ting KH, Hill CL, Whittle SL. Quality of reporting of interventional animal studies in
816 rheumatology: a systematic review using the ARRIVE guidelines. *International Journal of*
817 *Rheumatic Diseases* 2015;**18**(5):488-94.
- 818 86. Muhlhausler BS, Bloomfield FH, Gillman MW. Whole animal experiments should be more
819 like human randomized controlled trials. *PLoS biology* 2013;**11**(2):e1001481.
- 820 87. McGonigle P, Ruggeri B. Animal models of human disease: Challenges in enabling translation.
821 *Biochemical Pharmacology* 2014;**87**(1):162-71.
- 822 88. Hackam DG. Translating animal research into clinical benefit. *BMJ: British Medical Journal*
823 2007;**334**(7586):163.

- 824 89. de Vries RBM, Wever KE, Avey MT, et al. The Usefulness of Systematic Reviews of Animal
825 Experiments for the Design of Preclinical and Clinical Studies. *ILAR Journal*
826 2014;**55**(3):427-37.
- 827 90. Jansen of Lorkeers SJ, Doevendans PA, Chamuleau SAJ. All preclinical trials should be
828 registered in advance in an online registry. *European Journal of Clinical Investigation*
829 2014;**44**(9):891-92.
- 830 91. Dal-Ré R, Ioannidis JP, Bracken MB, et al. Making prospective registration of observational
831 research a reality. *Science translational medicine* 2014;**6**(224):224cm1-24cm1.
- 832 92. Rollin BE. Animal Research, Animal Welfare, and the Three R's *The Journal of Philosophy,*
833 *Science & Law* 2010;**10**.
- 834 93. Osborne NJ, Phillips BJ, Westwood K. Journal editorial policies as a driver for change - animal
835 welfare and the 3R. *New Paradigms In Laboratory Animal Science - Proceedings of the*
836 *Eleventh FELASA symposium and the 40th Scand-LAS Symposium. Helsinki, 2010:18-23.*
- 837 94. Martins A, Franco N. A Critical Look at Biomedical Journals' Policies on Animal Research by
838 Use of a Novel Tool: The EXEMPLAR Scale. *Animals* 2015;**5**(2):315-31.
- 839 95. Editorial. Checklists work to improve science. *Nature* 2018;**556**:273-74.

840

841

842

1 **Methodological standards, quality of reporting, and regulatory compliance in**
2 **animal research on amyotrophic lateral sclerosis: a systematic review**

3 Joana G Fernandes^{1,2¶}, Nuno H Franco^{1,2¶}, Andrew J Grierson³, Jan Hultgren⁴, Andrew JW
4 Furley^{5&}, I Anna S Olsson^{1,2&*}

5 ¹ Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Portugal

6 ² IBMC-Instituto de Biologia Molecular e Celular, Universidade do Porto, Portugal

7 ³ Sheffield Institute for Translational Neuroscience, Department of Neuroscience, University of
8 Sheffield, Sheffield, United Kingdom

9 ⁴ Department of Animal Environment and Health, Swedish University of Agricultural Sciences,
10 Skara, Sweden

11 ⁵ Department of Biomedical Science, University of Sheffield, Western Bank, Sheffield, United
12 Kingdom

13 *Corresponding author

14 E-mail: olsson@ibmc.up.pt

15 ¶ These authors share the first authorship

16 & These authors share the last authorship

17

18 **Key words:** Amyotrophic Lateral Sclerosis, ALS, Guidelines, Methodology, Reporting, Quality,
19 Compliance, Animal Welfare, Reproducibility

20

21 **Abstract**

22 **Objectives**

23 The ALS research community was one of the first to adopt methodology guidelines to improve
24 preclinical research reproducibility. We here present results of a systematic review to
25 investigate how the standards in this field changed over the ten-year period during which the
26 guidelines were first published (2007) and updated (2010).

27

28 **Methods**

29 We reviewed 569 research papers reporting research with SOD1 mice, published between 2005
30 and 2015.

31

32 **Results**

33 Reporting standards improved over time. Of papers published after the first ALS guidelines were
34 made public, fewer than 9% referred specifically to these. Of key research parameters, only
35 three (genetic background, number of transgenes and group size) were reported in >50% of the
36 papers. Information on housing conditions, randomization and blinding were absent in over two
37 thirds of papers. Group size was among the best reported parameters, but the majority reported
38 using fewer than the recommended sample size and only two studies clearly justified group size.
39

40 **Conclusions**

41 General methodological standards improved gradually over an 8- to 10-year period but
42 remained generally comparable to related fields with no specific guidelines, except with
43 regard to severity. only 11% of ALS studies were classified in the highest severity level
44 (animals allowed to reach death or moribund stages), substantially below the proportion in
45 studies of comparable neurodegenerative diseases such as Huntington's. The existence of
46 field-specific guidelines, though a welcome indication of concern, seems insufficient to
47 ensure adherence to high methodological standards. Other mechanisms may be required to
48 improve methodological and welfare standards.

49 **Strengths and limitations:**

- 50 - This systematic review is the first to assess both methodological standards and regulatory
51 compliance reporting in a research field with community guidelines on methodological
52 standards
- 53 - Our large sample (N=569 papers) includes half the total population of published papers
54 between 2005-2015
- 55 - The approach for this systematic review is unique in covering methodological quality,
56 regulatory compliance and severity / animal welfare
- 57 - We built two comprehensive scores (for methodological standard and for reporting quality)
58 which were subjected to regression analysis to investigate how these scores (dependent
59 variables) were related to publication year, type of study, country of origin and journal
60 (explanatory or predictor variables).
- 61 - While more models of ALS are now available, only studies using the SOD-1 mouse were
62 included

63 **1. Introduction**

64 Amyotrophic lateral sclerosis (ALS) is a rapidly progressing neurodegenerative disease typically
65 resulting in death two to five years after the onset of symptoms. There is no known cure and the
66 most widely used treatment– riluzole – extends survival by just two months¹. ALS research using
67 animal models focuses primarily on two main interconnected goals: understanding the
68 underlying mechanisms involved in motor neuron death in the brain and spinal cord, and
69 development and testing of potential drug therapies². This research relies substantially on
70 genetically modified animals, particularly transgenic mice expressing mutant forms of the
71 human Superoxide Dismutase 1 (SOD1) gene, which manifest several important characteristics
72 of the human disease^{3,4}.

73

74 While the use of animal models is relevant for advancing knowledge and considered essential
75 for testing putative treatments, it also presents ethical challenges and thus may be a reason for
76 public concern. As a result, a common legal requirement in many countries is that animal
77 research projects undergo an evaluation process intended to ensure that protocols are designed
78 and carried out in compliance with the 3Rs principle: *replacement* of animal use by non-animal
79 methods, *reduction* of animal numbers needed to achieve the scientific objectives, and
80 *refinement* of procedures to reduce or prevent harm to animals and improve their wellbeing.
81 Systematic reviews of animal use in both neuroscience⁵ and infection⁶ research indicate that
82 self-reported regulatory compliance – including of ethical approval of protocols – has steadily
83 increased over the last decade, but that significant progress could still be made to minimise and
84 prevent avoidable suffering of laboratory animals. One key measure for accomplishing this is the
85 termination of experiments during less severe stages of disease development where it is
86 scientifically valid to do so. Endpoints based on early obtainable and scientifically sound
87 indicators of phenotype progression can not only improve the ethical acceptability of animal
88 studies, but also prevent the confounding influence of secondary factors; in the case of animal
89 models of neurodegenerative diseases, starvation and dehydration arising from difficulties in
90 eating and drinking due to progressive motor impairment can affect the phenotype and the
91 readout of survival studies⁷⁻⁹. Simple refinements – such as adding mash food and longer bottle
92 spouts – can however help reduce the influence of such factors¹⁰⁻¹².

93

94 Of related concern are reports that a number of published animal studies fail to uphold basic
95 standards regarding experimental design – e.g. random assignment of animals to treatment
96 groups, blinding of observers – or use too few animals often leading to irreproducible results of

97 limited translational value¹³⁻¹⁸. This also holds true for neuroscience¹⁹⁻²², with concerns over the
98 overall quality and reproducibility of published results being raised for several neuroscience sub-
99 fields, including multiple sclerosis²³, stroke²⁴, spinal cord injury²⁵ Alzheimer's²⁶, Parkinson's²⁷,
100 Huntington's¹² and ALS²⁸ research. This has led major science funders, including the National
101 Institutes of Health²⁹ and Research Councils UK³⁰ to demand that future grant proposals attest
102 to the likelihood of providing reliable results, by including details of experimental design and
103 adequate justification of sample sizes. Reproducibility is further hindered by insufficient
104 provision of information on methodology in published research³¹ – including failure to account
105 for key variables such as sex, genotype, age, and weight of animals, anaesthetics used or
106 methods of euthanasia. Omitting information also makes it impossible to evaluate the study
107 quality and there is evidence that papers that do not report randomization or blinding
108 exaggerate biological effects³²⁻³⁴.

109

110 Broadly, the public conditionally approves of animal studies on the assumption that the harm
111 caused is offset by the benefits achieved and that scientists strive to minimise the former and
112 optimise the latter^{35,36}. Doing so requires scientists to critically revise their methods to maximise
113 translational relevance^{18,37}. Scientists are rightly concerned and, within the self-correcting
114 process of science, must rely on themselves to both identify the main obstacles hindering its
115 progress and find adequate solutions. To address the issue of methodological standards and
116 quality of reporting of basic and applied ALS studies, the ALS research community held two
117 meetings in 2006 and 2009, resulting in the publication of guidelines for animal studies in this
118 field^{2,38}. These guidelines aim to improve and standardise research methodology, and
119 encourage authors and journals to publish negative results in order to avoid publication bias.
120 The actual impact of such guidelines on how the ALS community carries out and reports research
121 has however not been assessed.

122

123 The present systematic review of animal studies of ALS uniquely aimed to assess, over an
124 extended period, the attention given to relevant methodological parameters (as a proxy for the
125 likely reliability of the study) and to examine how the principles of *refinement* and *reduction*
126 (measures to minimise animal harm) were considered. Both proof-of-concept and preclinical
127 studies were included in order to assess the influence of type of study.

128 2. Methods

129 2.1 Database search

130 An advanced search was conducted on the *ISI Web of Science*[®] database with the query *TS =*
131 *((mice OR mouse) SAME (ALS OR "amyotrophic lateral sclerosis"))*. Results were refined to
132 include only original research articles written in English and published in 2005, 2007, 2009, 2011,
133 2013 and 2015. Years of publication were selected to include papers reporting research planned
134 and carried out prior to and after the publication of guidelines for ALS research in 2007³⁸ and
135 2010², resulting from two international meetings held in 2006 and 2009, respectively (Figure 1).

136

137

138 **Figure 1. Timeline of relevant events.** The bottom arrows signal the years for which papers in our sample
139 were retrieved and the top arrows indicate the years when workshops on best practice in ALS animal
140 research were held, as well as when guidelines stemming from these were published. The grey bars
141 illustrate the 1-4 year period over which ALS animal studies reported in 2005 were likely to have been
142 designed and carried out, an estimation that can also be applied for the other years reviewed (2007, 2009,
143 2011, 2013, and 2015).

144

145

146 The choice to focus on SOD-1 mice was based on the predominant role of this model in animal-
147 based research into ALS (see Supplementary Figure 1).

148

149

150 **Supplementary Figure 1 - Trends in animal model chosen in ALS research, based on the number of hits**
151 **from an *Clarivate Analytics Web of Science*[®] advanced search applying the search queries: a) *TS=*(("ALS"
152 OR "amyotrophic lateral sclerosis") AND "SOD1" AND ("mouse" OR "mice")); b) *TS=*(("ALS" OR
153 "amyotrophic lateral sclerosis") AND "TDP-43" AND ("mouse" OR "mice")); and c) *TS=*(("ALS" OR
154 "amyotrophic lateral sclerosis") AND "FUS" AND ("mouse" OR "mice"))**

155

156

157 The search was performed in February 2013 for scientific articles from 2009 and 2011 (after the
158 first and second conferences, respectively), in August 2013 for scientific articles from 2005
159 (before the two conferences), in September 2014 for scientific articles from 2013, in November
160 2016 for scientific articles from 2015, and in February 2017 for scientific articles from 2007. After
161 the triage process, illustrated in Figure 2, 569 full-text articles remained for analysis: 77 from
162 2005, 81 from 2007, 84 from 2009, 106 from 2011, 115 from 2013, and 106 from 2015.

163

164 **Figure 2. Triage process.** The first triage step involved reading each of the 1993 abstracts and excluding
165 all papers that were not related to ALS. The second triage step excluded all papers that did not report
166 original research with SOD1 models of the disease.

167

168 2.2 Data collection

169 Each published study was categorised as either a ‘preclinical’ (i.e., carried out “to evaluate a
170 drug for use in humans”) or ‘proof-of-concept’ (i.e., aiming “to elucidate the mechanism of the
171 disease”), according to the suggested classification for animal studies on ALS ^{2 38}. Table 1
172 describes the information retrieved regarding regulatory compliance, animal models,
173 experimental design and animal welfare. This information was retrieved through careful reading
174 of the full papers. The review protocol was defined prior to data collection. Data extraction was
175 carried out by JGF, with support from NHF, AJG and IASO for disambiguation.

176

177 **Table 1. Data retrieved.** A description of the information collected from revised papers is presented for
178 each item.

Category	Items	Description/Observations
Regulatory compliance	Ethical approval	Studies explicitly reported to be approved by a committee / authority.
	Guideline compliance	Articles that did not report having experimental protocols ethically approved by an institution or national entity, but reported that some kind of guidelines for use and care of laboratory animals was followed.
Animal models	Genetic background	When available.
	Sex	Four options: Male, female, both or not reported. For <i>both</i> , information on whether studies were balanced for gender was retrieved.
	Number of transgene copies	When available.
Experimental design	Group size	Mean group size, based on the available information
	Randomization	Studies explicitly reporting assigning animals to groups randomly
	Blinding	Studies explicitly reporting blinding of observers to experimental groups
	Non-transgenic littermate control	Studies explicitly reporting the use of non-transgenic littermates as control.
	Splitting littermates into groups	Studies explicitly reporting that littermates were split into groups.
	Housing and husbandry conditions	Reporting information regarding temperature, humidity, light of the room where animals were kept, and cage size and number of animals per cage.
Animal Welfare/ Procedures	Severity	Described in table 2.
	Refinement	Relevant refinements to minimise suffering and distress, such as housing adaptations.
	Euthanasia method	Euthanasia methods were divided into the following categories: “Under anaesthesia” (including anaesthetic overdose); “CO ₂ asphyxiation”; “Other”; “Not reported” and “Not performed”.

179

180 For severity assessment, a scale was devised based on the specific characteristics of the ALS
181 models and their progressive disease phenotype (Table 2). The ALS models used in the reviewed
182 studies express diverse mutant forms of the *SOD1* gene. The onset of disease for these models
183 is generally characterised by weakness and tremors of the hind limbs, together with a mild loss
184 of body weight. Disease progression leads to paralysis of hind limbs, followed by complete
185 paralysis (example, Figure 3 in ³⁹), accompanied by increased difficulty to eat, drink and swallow
186 ^{40,41}. Mice die of respiratory failure due to paralysis of the diaphragm ⁸. Age of onset and death,
187 as well as the interval between them, vary depending on the mutation of the amino-acid and
188 codon e.g. ⁴², number of copies of transgene e.g. ⁴³, and genetic background ⁴. For instance, the
189 over-expressing SOD1G93A Line Gur 1H (B6SJL hybrid) presents with an early onset of overt
190 motor symptoms (3-4 months) and moderate rate of progression (3 weeks from onset to death)
191 ⁴⁴, whereas the highly expressing SOD1G85R Line 148 presents with later onset (7.5 months) and
192 faster disease progression (2 weeks from onset to death) ⁴⁵. Also, factors such as the animal
193 supplier (e.g. ^{46,47}), in-house breeding ⁴⁸ and crosses with other non-SOD1 models (e.g. SOD1
194 mice crossed with gene-specific knockout mice ⁴⁹) are further sources of variability.

195 Maximum estimated severity was classified according to a five-level scale (Table 2). Scoring was
196 based on the estimated clinical state of animals at the most advanced stage of disease
197 progression they were allowed to reach. Studies in which information was insufficient to draw
198 conclusions about the level of severity were classified as 'undetermined'. This severity scale was
199 developed building upon previous work from members of this team (NF, AO) developed for
200 classifying studies on mouse models of Huntington's disease (table 2 in ⁵), together with our own
201 (AG) experience with mutant SOD1 mouse models and literature. For purposes of statistical
202 analysis, the severity scale was reduced to a binary scale, ("low"= severity up to level 4; "high"=
203 level 5 severity. The choice for above level-4 severity as a cut-off point, was based on its status
204 as a "standard endpoint" in published ALS guidelines ^{2,38}, whereas full paralysis or spontaneous
205 death exceeds this standard endpoint, as well as the legally recommended endpoints in many
206 countries, including the EU Member States.

207

208

209

210

211

212

213 **Table 2.** Severity scale for ALS studies on transgenic mice with a mutant SOD1 gene. Each severity level
 214 exemplified from the most commonly used B6.Cg-TgN-(SOD1G93A) G1H mouse. Classification was based
 215 on the most severe endpoint used in each publication.
 216

Severity	Description	Welfare issues during this stage
Level 1	Animals euthanized prior to disease onset, which is characterised by progressive weight loss or hind limb tremors	No overt motor dysfunction. Phenotype is subclinical. Loss of motor function can be detected using rotarod or running wheels, but does not interfere with normal behaviour
Level 2	Studies terminated at an early stage of disease: animals present trembling and weakness in hind limbs (by approx. 75d) and mild body weight loss.	Minor. Loss of motor function can be detected using rotarod or running wheels, but has little interference with normal behaviour.
Level 3	Experiments terminated when animals are no longer able to reach food hopper or bottle spout. This occurs when animals reach a moderate (gait abnormalities and weakness) to severe (hind limb paralysis) stage of motor impairment (usually at 120-125d)	Medium. Loss of motor function and body weight can be detected by monitoring (e.g. by a clinical score sheet) and by checking self-righting ability. Refinement measures to address these welfare issues include provision of softer bedding material (e.g. sawdust), elongated bottle spouts and mashed food on the cage floor.
Level 4	Animals euthanized after losing the ability to right themselves within 10-30 seconds after being laid on either side (one or both) or when percentage of weight loss reaches 15-20% of peak body weight (usually at 130-140d)	Major. Animals show severe locomotor impairment. Refinement as described for level 3
Level 5	Animals are euthanized when reaching a moribund stage (complete paralysis) or allowed to die spontaneously	Severe. At this stage animals are unable to move, eat or drink. Animals which are not euthanized will die as a result of respiratory failure.

217

218 **2.3. Methodological Standards Reporting (MSR) and Regulatory Compliance Reporting (RCR)**
 219 **scores**

220 For each reviewed publication, data were collected on a number of items which all contributed
 221 with information about the reporting quality of the paper. For the analysis, we brought these
 222 items together into two scores, hence generating for each paper two comprehensive measures
 223 for reporting quality, one on methodological standards and one on regulatory compliance. We
 224 then used regression analysis to investigate how the two scores (dependent variables) were
 225 related to publication year, type of study, country of origin and journal (explanatory or predictor
 226 variables), as outlined in detail in the following. Based on the regression models it is possible to
 227 predict how the dependent variables would have changed with changes in the explanatory
 228 variables. In contrast to, for example, correlation, the regression analysis takes into account all
 229 the explanatory variables that were included in the models, i.e. the estimated association

230 between a score and one of the explanatory variables is independent of the values of the other
231 explanatory variables considered. In that way, spurious associations caused by relationships
232 between the explanatory variables in the data can be avoided.

233 The two scores were formed as weighted sums of separate sets of items. The Methodological
234 Standards Reporting (MSR) score was constructed from the items *sampsize*, *climate*, *cagesize*,
235 *nmice*, *sex*, *copies*, *genetic* (which refer to important research parameters in animal
236 experimentation and in ALS research in particular) and the items *random*, *blinded*, *control*,
237 *sibsplit*, and *exclus* (associated with general good practices in the design of animal experiments
238 and published recommendations for ALS studies). Greater weight (1.5 versus 1) was attributed
239 to items which are also part of the ALS guidelines. Table 3 describes these items, their attributed
240 weight in the MSR score and the absolute number and percentage of papers reporting this
241 information, divided by type of study.

242 The Regulatory Compliance Reporting (RCR) score was originally constructed from the items
243 *comply*, *protocol*, *severity* (turned into a binary classification) and *refine*; the final version (RCRb)
244 included *comply*, *protocol* and *refine*.

245

246 **Table 3. List of items integrated in the MSR and the RCR scores for preclinical (n=108) and proof-of-**
247 **concept (n=461) animal studies on ALS reporting this information.** The score for each variable is provided
248 (MSR score ranging from 0 to 12.5, and RCR score ranging from 0 to 3). Greater weight (1.5 versus 1) was
249 attributed to items which are also part of the ALS guidelines. The internal consistency reliability of each
250 score was checked using Cronbach's alpha⁵⁰, estimating the degree to which the items measured a latent
251 construct. For MSR, alpha was 0.58, which indicates poor internal consistency, which however does not
252 disqualify the score. Omission of items *exclus*, *cagesize* and *blinded* increased alpha only slightly, thus the
253 original MSR score was retained. For RCR, alpha was 0.30, but increased to 0.42 when item *severity* was
254 omitted. For purposes of statistical modelling, RCR (only including items *comply*, *protocol* and *refine*) was
255 later simplified to a binary variable RCRb coded as 1 for RCR values 2-3 and as 0 for RCR values 0-1.

256

257

258

259

260

261

262

263

264

265

Reported information	MSR score		'Proof-of-Concept' (n=461)		'Preclinical' (n=108)	
	Score item	Score weight	Absolute number	%	Absolute number	%
Relevant animal research variables						
Group size	<i>sampsize</i>	1.5	368	79.8	106	98.1
Environment: light, temp., humidity (fully or partially reported)	<i>climate</i>	1	123	26.7	42	38.9
Cage size	<i>cagesize</i>	1	1	0.2	2	1.9
Mice per cage	<i>nmice</i>	1	26	5.6	15	13.9
Sex of the animals	<i>sex</i>	1.5	223	48.4	71	65.7
Number of transgene copies	<i>copies</i>	1.5	286	62.0	80	74.1
Genetic background	<i>genetic</i>	1.5	349	75.7	92	85.2
Measures to reduce 'noise' and bias in experiments						
Animals randomised to treatment groups	<i>random</i>	1	28	6.1	47	43.5
Observers blinded to treatment	<i>blinded</i>	1.5	94	20.4	52	48.1
Non-transgenic littermate controls used	<i>control</i>	1	150	32.5	39	36.1
Splitting littermates into groups	<i>Sibsplit</i>	1	28	6.1	31	28.7
Reason for exclusion of animals is reported	<i>exclus</i>	1	2	0.4	6	5.6

Reported information	RCR score		'Proof-of-Concept' (n=461)		'Preclinical' (n=108)	
	Score item	Score weight	Absolute number	%	Absolute number	%
Self-reported compliance with laws and regulations	<i>comply</i>	1	98	21.3	28	25.9
Project approval reported	<i>protocol</i>	1	315	68.3	66	61.1
Refinement measures (e.g. to aid feed and hydrate) to aid feed and hydrate)	<i>refine</i>	1	29	6.3	14	13

266

267

268 MSR and RCRb were modelled, estimating the effects of publication year (2005, 2007, 2009,
269 2011, 2013 or 2015) and study type (preclinical or proof-of-concept) while accounting for
270 possible confounding by country of origin (15 categories), journal (17 categories) and severity
271 (low or high). Countries contributing with less than twelve papers, and journals contributing with
272 less than 6 papers, were combined into separate categories, denoted 'Other'. MSR was
273 modelled using linear regression and RCRb by logistic regression. All first-order interaction
274 effects were tested and included if significant.

275 Predictive marginal means were calculated for all independent variables. The marginal means
276 showed the values of MSR and the probabilities of RCR >1, respectively, that the models
277 predicted for different values of each one of the significant independent variables, assuming
278 that the remaining variables in the models had their observed values. Both models were checked
279 using the Pregibon link test ⁵¹, and by examining standardised residuals. The MSR model was
280 also checked with the Breusch-Pagan/Cook-Weisberg test for heteroscedasticity ⁵², the Ramsey
281 regression specification-error test for omitted variables ⁵³, and the RCRb model by examining
282 delta-betas to identify influential observations. The proportion of the total variation in MSR and
283 RCRb that could be explained by differences between countries or journals was determined by
284 running empty mixed models with country and journal, respectively, as a random effect, and

285 calculating the intra-class correlation coefficients. The justification for weighting the items
286 composing MSR was checked by modelling an alternative score formed without weighting. The
287 differences between years and countries remained virtually unchanged, although the score
288 values were now generally lower.

289 The association between MSR and RCR scores was estimated using Spearman rank correlation.
290 A total of 490 observations could be used. Overall MSR mean \pm SD was 5.69 ± 2.39 . RCR assumed
291 values 0 (n=48), 1 (n=103), 2 (n=309) or 3 (n=30), resulting in 69% of the observations having
292 values above 1. The number of observations per level of year, study type, country, journal and
293 severity is shown in Supplementary Table 1.

294 The data were analysed in Stata/IC v. 13.1 and IBM SPSS 23.0. Each article was regarded as the
295 experimental unit and the level of significance for all tests was 0.05.

296

297 **3. Results**

298 **3.1. Quality of research and reporting**

299 The quality of methodological standards and of reporting is crucial to avoid bias and achieve
300 reliable, repeatable and translatable research results. We measured this through the
301 Methodological Standards Reporting Score and also looked at specific research parameters
302 individually.

303 **3.1.1 Methodological Standards Reporting Score**

304 The 12 items that comprise the Methodological Standards Reporting Score represent seven
305 relevant experimental variables and five measures for reducing bias in animal experiments.
306 Higher scores mean better reporting and implementation of good practices in the design of ALS
307 animal studies.

308 MSR was significantly affected by year and study type (joint F-test $p=0.0015$ and <0.0001 ,
309 respectively). Compared to 2007, the logistic regression model predicted a higher MSR for the
310 subsequent years (2009, 2011, 2013 and 2015) as well as for 2005 (Figure 3). It also predicted a
311 higher MSR for preclinical studies than for proof-of-concept studies (marginal mean 7.28 and
312 5.26 respectively). Model diagnostics showed that linear regression was justified and the model
313 fit was excellent. Supplementary Table 2 shows the MSR model results.

314

315 **Figure 3. Predictive marginal means \pm 95% confidence interval of publication year (left panel) and**
316 **country (right panel) based on a model of a Methodological Standards Reporting (MSR) Score in 487**
317 **ALS studies (predicted score values).** According to the linear regression model, MSR could be expected
318 to be 0.74, 1.33, 1.50 and 1.18 units higher in 2009, 2011, 2013 and 2015 ($p=0.047$, 0.001, 0.000 and
319 0.002), respectively, and 0.98 units higher in 2005 ($p=0.011$). No significant interactions were found (e.g.
320 between country and year). According to the R-square statistic the model explained 25% of the total
321 variation in MSR.

322

323 3.1.2. Reporting of relevant research parameters

324 Some research parameters were very seldom reported, for example: numbers of animals per
325 cage (7.2%, 41/569); cage size (0.5%, 3/569) and exclusion of animals (1.4%, 8/569). Measures
326 in guideline recommendations to reduce bias in ALS research were mostly not reported,
327 including: splitting littermates to treatment groups (10.4%, 59/569); use of non-transgenic
328 littermates as controls (33.2%, 189/569); as well as measures of broader application, such as
329 random assignment of animals to treatments (13.2%, 75/569) or blinding of observers (25.7%,
330 146/569). By contrast, numbers of transgene copies and genetic backgrounds of animals were
331 reported in the majority of papers.

332

333 Of papers reporting sex ($n=297$), 54.2% (161/297) described studies using mice of both sexes,
334 while 29.0% (86/297) used only males and 16.8% (50/297) used only females. Reporting of sex
335 rose steadily from 2005 (39.0%, 30/77) to 2015 69.8% (74/106), $\chi^2(5) = 30,831$, $p < 0.0001$,
336 linear-by-linear association=27.802, $p < 0.0001$).

337 Regarding the chosen genetic background of animals used for preclinical studies ($n= 108$), 76%
338 (70/92) of those reporting this parameter generated experimental animals using a cross
339 between mice hemizygous for the SOD1 mutant gene and C57/SJL outbred strains.

340

341 Only ten studies (6 proof-of-concept studies and 4 preclinical studies) from 2007, 2009, 2011,
342 2013, and 2015 justified the number of animals used per group. However, of these, only six gave
343 clear justifications (five justified the group size by a power analysis and the other by the size of
344 groups proposed in ALS guidelines^{2 38}. On the other hand, group size was reported in 83.3%
345 (474/569) of ALS papers, and more so in the preclinical studies sub-sample (Figure 4).

346

347

348 **Figure 4. Group size.** Histogram of mean group size in 105 preclinical studies reporting this parameter
349 (left) and for each of the years analysed (yearly mean ± 1 standard deviation) (right).

350

351

352 Of the 569 papers reviewed, 38% (214/569) did not report the method for killing animals despite
353 the fact that in 91% (195/214) of these, terminal procedures requiring anaesthesia for ethical
354 and practical reasons were identified (e.g. transcardial perfusion fixation). The most commonly
355 used euthanasia method – of the papers reporting this information – was anaesthetic overdose
356 or the use of another method under anaesthesia (86%; 317/367) while other methods such as
357 CO₂ asphyxiation (7%; 26/367) or others such as decapitation or cervical dislocation (7%; 24/367)
358 were seldom used. Very few studies (15 out of 569) were identified as not performing
359 euthanasia of any kind. The remaining 21 articles were deemed “inconclusive”, for neither
360 reporting euthanizing animals at any point nor reporting deaths.

361

362 **3.2 Regulatory compliance and estimated severity**

363 For public confidence in research, it is important that research with animals is carried out
364 according to standards set by legislation and in line with the principles of the 3Rs. We measured
365 such compliance through the Regulatory Compliance Reporting score and also looked at specific
366 research parameters individually.

367

368 **3.2.1. Regulatory Compliance Reporting Score (RCR)**

369 The Regulatory Compliance Reporting (RCR) Score, which measures to what extent compliance
370 with legislation and approval of animal experiments are reported in published papers, shows an
371 overall improvement in the reporting over the time period under study (joint Chi-square
372 $p < 0.001$, Figure 5). RCR did not differ between journals or between proof-of-concept and
373 preclinical studies but was affected by country (Figure 5). Studies with high severity seemed to
374 have higher odds of high RCR values ($p = 0.027$). Model diagnostics showed that logistic
375 regression was justified. Supplementary Table 3 shows the RCR model results.

376

377 **Figure 5. Predictive marginal means \pm 95% confidence interval of publication year (left panel) and**
378 **country (right panel) based on a model of a Regulatory Compliance Reporting (RCR) Score in 490 ALS**
379 **studies (predicted probabilities of S values > 1).** The odds of an RCR score above 1 was 3.43 and 7.07 times
380 higher in 2013 and 2015 ($p = 0.003$ and 0.000), respectively, than in 2005. China, France, Italy and South
381 Korea appeared to have comparatively low probabilities, while for example Spain, Belgium and Canada

382 had somewhat high probabilities. No significant interactions were found. The pseudo R-square statistic
383 indicated that the model explained 16% of the total variation in the data.

384

385 Over the entire period, most papers (67.0%; 381/569) reported that studies had been appraised
386 and approved by a third party (e.g. ethics committee, competent authority) with only 10.9%
387 (62/569) not reporting any kind of regulatory compliance. By 2015, all papers were found to
388 have some type of statement on regulatory compliance, most of which (83%) referring to prior
389 ethical approval of research protocols.

390

391 The correlation between MSR and RCR was weak, but highly significant ($\rho=0.21$; $p<0.0001$)
392 indicating that papers with high scores for methodological standards were somewhat more
393 likely to also score highly for regulatory standards.

394 3.2.2 Severity and refinement measures

395 We have found in previous systematic reviews^{5 6 54} that self-reported compliance with
396 regulations may not necessarily affect the severity of the experiments being conducted. To test
397 whether actual experimental practice has changed over the study period, we classified the
398 severity of each study according to the criteria in Table 2. The majority of publications (60.7%)
399 (346/569) included experiments at level-4 severity (Figure 6-A). Of the 64 studies classified as
400 Level 5 (allowing animals to die of disease progression or to reach complete paralysis), 89%
401 reported regulatory compliance (70% ethical approval from a national authority or institutional
402 ethics committee and 19% compliance with relevant legislation or animal use guidelines).
403 However, between those studies that reported regulatory compliance and those that did not,
404 there was no difference in the proportion that were Level 5 ($\chi^2(5) = 2.855$, $p = 0.722$) (Figure 6-
405 B).

406 On the other hand, we did observe a difference between preclinical and proof-of-concept
407 studies: preclinical studies included a higher proportion of studies within the highest severity
408 categories (77.9% (81/104) classified as level 4 and 19.2% (20/104) as level 5) than did proof-of-
409 concept studies (68.7% (265/386) classified as level 4 and 11.4% (44/386) as level 5). Moreover,
410 no preclinical studies were given a level 1 or level 2 severity ($\chi^2(5) = 19.593$, $p = 0.001$) (Figure
411 6-C).

412

413

414 **Figure 6. Severity classification of studies (N=569).** Figure 6-A illustrates the percentage of studies, by
415 year, classified into each of the 5-levels of our severity scale, as well as those of "undetermined" severity

416 due to insufficient information ($n = 77$ in 2005; $n=81$ in 2007; $n = 84$ in 2009; $n = 106$ in 2011; $n = 115$ in
417 2013; $n= 106$ in 2015. Figures 6-B and 6-C show percentage of studies classified into each of the 5-levels,
418 according to, respectively, reported regulatory compliance status ($n = 62$, not reported; $n = 126$, guidelines
419 followed; $n = 381$, protocol approval), and type of study ($n = 461$, proof-of-concept studies; $n = 108$,
420 preclinical studies).

421

422 Of studies classified between levels 3 and 5 severity (i.e. from which it could be ascertained
423 animals presented overt locomotor impairments), only 9.1% (42/456) described any refinement
424 measures to alleviate suffering (e.g. provision of mashed food and adaptation of bedding in later
425 stages of disease progression), which occurred almost exclusively (39/42) in Level 4 studies.

426 Differences in the regulatory landscape between countries imply that *how* animals are treated
427 in biomedical research may depend on *where* these experiments are carried out. The proportion
428 of high-severity (Level-5) studies differed significantly ($\chi(13) = 35,561$, $p=0.001$) between the 14
429 most represented countries in our sample, ranging from 40% (8/20) and 41% (7/17) – in South
430 Korea and Israel, respectively – to 4% in Canada and China and even none in Belgium (0/14) and
431 the UK (0/23).

432

433 **4. Discussion**

434 Our analysis, the first of its kind to use specially devised scores encompassing both
435 methodological standards and regulatory compliance reporting (MSR and RCR, respectively)
436 over a 10 year period, suggest three main findings: The first is an overall improvement in both
437 regulatory compliance and methodological and reporting quality across the period assessed.
438 Also, and somewhat as expected, studies classified as 'preclinical' scored higher for
439 methodological and reporting quality as compared with more 'proof-of-concept' studies. The
440 third finding is that scores for both scores varied widely according to the country in which the
441 first author was based, but not according to the journal publishing the paper.

442 The improved reporting of regulatory compliance, as expressed in the increase in RCR score
443 across time, is an indicator of widespread increase in reported adherence to animal welfare
444 regulatory requirements. However, this was not reflected in any significant change in the
445 proportion of highly severe (Level-5 in our classification scheme) studies or the reporting of
446 refinement measures (in studies where animals showed overt clinical signs). This is in agreement
447 with results from previous systematic reviews of animal research on Huntington's disease
448 (papers published 1997-2009) ⁵ and tuberculosis (1997-2011) ⁶. Also, while 'preclinical' studies

449 were more likely to be classified in the higher severity categories, there was no relation between
450 the level of severity and whether papers reported approval of protocols or compliance with
451 regulations, the latter also reflecting previous findings^{5 54}.

452

453 Only 11.2% of ALS studies were classified at the highest severity level (level 5, i.e. including
454 experiments with spontaneous death or euthanasia at a near-death stage, i.e. complete
455 paralysis), which is much lower than that found in research using mouse models of Huntington's
456 Disease (38%)⁵ and Tuberculosis (66%)⁶. Moreover, most endpoints applied in ALS studies
457 adhered to the same basic criterion for euthanizing animals, namely the point at which animals
458 are unable to resume their position if laid recumbent within 10-30 seconds. This is the primary
459 endpoint proposed in existing guidelines for preclinical ALS^{2 38} and the ALS Treatment
460 Development Institute's recommendations²⁸ (level-4 severity on our scale), suggesting
461 researchers to a great extent act in accordance with published guidance published guidance in
462 this respect. However, this endpoint was already broadly used before the publication of the
463 guidelines suggesting that these reflect common practice at the time of publication.

464

465 Applying predefined endpoints is important to prevent the loss of biological samples from
466 animals found dead and for which time of death therefore cannot be defined⁵ hence maintaining
467 numbers of animals and avoiding loss of statistical power and subsequent inconclusive results.
468 However, from an animal welfare perspective, the current standard endpoint for ALS studies
469 corresponds to an end-stage where euthanasia may prevent deaths from respiratory failure, but
470 since they seldom anticipate death by more than a day, or even just a few hours, late stage
471 endpoints only curtail a small part of animal suffering⁷. Very late endpoints increase the
472 likelihood that at least some animals will die unsupervised (e.g. overnight), while the
473 confounding effect of starvation and dehydration in survival data increases as animals become
474 progressively less able to reach the bottle spout or the food hopper⁵⁵. At advanced clinical
475 stages, refinements such as providing mashed food on the cage floor, long-spouted water
476 bottles or fluid administration are therefore crucial to avoid unnecessary animal suffering and
477 to improve validity by bringing the model closer to the clinical setting, where late-stage human
478 patients are provided palliative care⁵⁶. Defining endpoints also needs to take the research
479 purpose into account. In ALS, the mechanisms operating at different stages of the disease are
480 known to be different, principally affecting distal axons at the onset of symptoms, but
481 developing an immune/inflammatory phenotype during the end stages⁵⁷. Therefore, endpoints
482 relevant to the treatment strategies must be used, particularly when targeting
483 neuroinflammation.

484

485 Methodological standards reporting improved over the time period under study. Studies
486 classified as 'preclinical' reported methodology in more detail than those deemed 'proof-of-
487 concept', consistent with the view that a more rigorous design and execution should be
488 demanded for preclinical studies⁵⁸. Nevertheless, the checklist provided in the 2010 edition of
489 the guidelines for ALS research sets high methodological standards for both types of studies².
490 Throughout the period under study, the MSR scores remain below 50% of the maximum score,
491 showing that the overall level of reporting of methodological detail remain substantially below
492 the recommendations in the guidelines.

493

494 Only three parameters (genetic background, number of transgene copies, and group size) were
495 reported in more than half of the sample, whereas other relevant information, such as housing
496 conditions, randomisation of animals into treatment groups or blinding of researchers was
497 absent in well over two thirds of the papers analysed, in line with previous reviews of animal
498 research in the neurosciences^{5 55 59}. Other biological and methodological parameters, such as
499 sex (only reported in the majority of papers in the "preclinical studies" sub-sample) and method
500 of choice for euthanizing animals were also largely under-reported. The method used for
501 euthanizing animals has both animal welfare implications and scientific relevance, as the
502 method affects biological and histological parameters differently, which can impact the *post*
503 *mortem* data collected^{60 61}. The increase in the proportion of articles in our sample reporting
504 sex of the animals is positive, as sex differences^{4 62-64} in the phenotype or response to
505 therapeutical drugs may influence results and be of clinical relevance. However, although ALS
506 guidelines propose the use of both male and female mice, little over half of the studies providing
507 this information reported doing so. Overall, making these and other details on animals and
508 protocol available is central to allowing an adequate interpretation of results and a critical
509 evaluation of their validity, as well as allowing study replication and proper integration of results
510 in systematic reviews and meta-analyses^{31 65}.

511

512 Sample size was generally well reported, but of those reporting this parameter, only a small
513 minority used the 24 per group recommended in the 2010 guidelines². Furthermore, only three
514 studies clearly justified group size, in agreement with previous reports that this is frequently
515 overlooked e.g.^{31 66}. Adequate sample size is paramount to ensure that animals, time and
516 resources are not wasted as a result of underpowering experiments by using too few animals⁶⁷
517⁶⁸. Noise reduction by genetic standardisation could also help reduce the number of animals
518 needed per study, as the reduced inter-individual variability of isogenic strains allows increasing

519 power without requiring more animals⁶⁹ and is indeed mentioned in the 2007 guidelines as a
520 way of reducing variability in drug testing³⁸. Mead and colleagues⁷⁰, for instance, have shown
521 great consistency of results by using SOD1G93A transgenic mice on an inbred C57BL/6 genetic
522 background, with the added advantage of presenting early indicators of disease progress,
523 allowing for faster and more humane drug screening. Only 11% of the preclinical studies
524 reviewed, however, used a fully inbred background. The use of a single well characterised model
525 for initial studies can be supported further by independent replication studies in a different
526 disease model.

527

528 Most articles did not report random assignment of animals to groups or blinded outcome
529 assessment. This reflects similar data from reviews on the methodological quality of preclinical
530 research on ALS^{28 59 71} and other fields^{31 33 72-74}. This lack of attention to measures to avoid noise
531 and biases in animal experiments is cause for concern, given their role in improving the reliability
532 of results, as well as the translational value of preclinical research^{16 24 33 68 72}. While it cannot be
533 excluded that in some cases blinding and randomisation were applied but not reported, one
534 might expect that researchers carrying out well thought-out and planned experiments would
535 state such measures, since this strengthens their results and conclusions. There is ample
536 evidence for many areas^{32 33 74-76} that published studies which do not report measures to
537 minimise bias (i.e. blinding, randomisation and allocation concealment) tend to present an
538 exaggerated estimate of the therapeutic effect of experimental drugs. This is particularly
539 relevant in the light of the ongoing discussion of why promising pre-clinical results of candidate
540 drugs for ALS have not translated into the clinic. Although the disappointing outcomes of clinical
541 trials apparently contradict the promising preclinical results that elicited them, they may actually
542 mirror the results obtained from adequately designed animal studies carried out to high
543 methodological standards^{28 77}.

544

545 Methodological standards reporting and regulatory compliance reporting scores were not
546 influenced by the journal in which the results were published. Other researchers who have
547 investigated the effect of journal on methodological standards and reporting quality have found
548 a statistically significant but very small effect of whether or not the journal had endorsed the
549 ARRIVE guidelines.^{78 79}

550

551 In contrast to previous research, this study indicated a gradual improvement in the
552 methodological standards and regulatory compliance reporting scores over time. However, it is
553 difficult to say to what extent this is the result of field-specific guidelines, as there is an overall

554 increasing trend in these score. Our study, of course, is limited to the period and model under
555 study and some improvements may have occurred as a result of the informal discussion leading
556 up to the formal workshops and guidelines (and more recently the appearance of other
557 transgenic models means that the study does not cover the entire field of ALS research for later
558 years). Also, a surprisingly low number of papers (1/84 in 2009, 10/106 in 2011, 10/115 in 2013
559 and 14/106 in 2015) referred to the Ludolph *et al.* guidelines^{2 38}. Given the slow adoption of the
560 ARRIVE guidelines⁸⁰, it seems likely it may also take some time for the ALS guidelines to have a
561 detectable effect.

562 While reporting of relevant parameters such as blinding and randomisation was higher in our
563 'preclinical' subsample than what has been reported in other systematic reviews^{16 31 78 80-83},
564 results for the overall sample were generally comparable. Also, and similarly to what was found
565 in these systematic reviews, justification for sample size was rarely reported.

566
567 One way of addressing the problems with study quality could be for preclinical researchers to
568 adopt the standards of randomised controlled trials in humans⁸⁴⁻⁸⁷, including trial pre-
569 registration^{88 89}. Compliance with existing guidelines would seem a more readily achievable goal,
570 however other self-regulatory mechanisms may be warranted to improve compliance, such as
571 changes to the publishing requirements of biomedical journals⁹⁰⁻⁹² or more demanding
572 requirements by science funders, both of which are clearly on the horizon^{30 93}.

573

574 **5. Conclusion**

575 The ALS research community pioneered the development of field-specific guidelines, setting
576 science community-based standards for animal research methodology and reporting^{2 38}.
577 Whereas we found significant improvement over time, it is less clear to what extent this is linked
578 to the guidelines, which are rarely referred to. Animal research in the field of ALS does however
579 differ from comparable research in other reviewed fields in one aspect: the implementation of
580 predefined endpoints in studies of advanced disease stages. This practice is important both for
581 research quality and animal welfare and is indeed coherent with the field-specific guidelines.
582 We propose that future guidelines should address measures to raise standards in the design,
583 conduct and reporting of experiments as well as to reduce the impact on animal welfare, as part
584 of a concerted effort to make biomedical research using animals more ethically and socially
585 acceptable and effective.

586

587 **Acknowledgements**

588 We thank Gilly Griffin for her input on current practice regarding humane endpoints in Canada.
589

590 **Funding**

591 NHF was a recipient of a Post-doctoral Research Fellowship from the Portuguese Foundation for
592 Science and Technology (FCT), grant reference SFRH/BPD/85978/2012. The research leading to
593 these results has received funding from the European Union Seventh Framework Programme
594 [FP7-HEALTH-2013-INNOVATION-1] under grant agreement n. ° 602616 [Project ANIMPACT].
595 Analysis and revision was supported by the project Norte-01-0145-FEDER-000008 - Porto
596 Neurosciences and Neurologic Disease Research Initiative at I3S , supported by Norte Portugal
597 Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership
598 Agreement, through the European Regional Development Fund (FEDER) and FEDER - Fundo
599 Europeu de Desenvolvimento Regional funds through the COMPETE 2020 - Operational
600 Programme for Competitiveness and Internationalisation (POCI), Portugal 2020, and by
601 Portuguese funds through FCT - Fundação para a Ciência e a Tecnologia/Ministério da Ciência,
602 Tecnologia e Ensino Superior in the framework of the project "Institute for Research and
603 Innovation in Health Sciences" (POCI-01-0145-FEDER-007274).

604

605 **Competing Interests**

606 We have read and understood BMJ policy on declaration of interests and declare that we have
607 no competing interests.

608

609 **Role of Authors and Contributors**

610

611 Original idea for this study: NF, AO

612 Conception and design of the work: NF, AO, AJF, AJG

613 Data collection: JF, NF

614 Data analysis and interpretation: NF, JF, JH, AJF, AJG, AO

615 Drafting the article: NF, JF

616 Critical revision of the article: AJG, AJF, JH, AO

617 Final approval of the version to be published: NF, JF, AJG, AJF, JH, AO

618

619 **Data access statement**

620 Dataset will be made available upon publication, at the University of Porto data repository.

621

622

623 **References**

- 624 1. Miller RG, Mitchell JD, Lyon M, et al. Riluzole for amyotrophic lateral sclerosis (ALS)/motor
625 neuron disease (MND). Amyotrophic lateral sclerosis and other motor neuron
626 disorders : official publication of the World Federation of Neurology, Research Group
627 on Motor Neuron Diseases 2003;**4**(3):191-206.
- 628 2. Ludolph AC, Bendotti C, Blaugrund E, et al. Guidelines for preclinical animal research in
629 ALS/MND: A consensus meeting. Amyotroph Lateral Scler 2010;**11**(1-2):38-45.
- 630 3. Shibata N. Transgenic mouse model for familial amyotrophic lateral sclerosis with
631 superoxide dismutase-1 mutation. Neuropathology 2001;**21**(1):82-92.
- 632 4. Heiman-Patterson TD, Deitch JS, Blankenhorn EP, et al. Background and gender effects on
633 survival in the TgN(SOD1-G93A)1Gur mouse model of ALS. Journal of the Neurological
634 Sciences 2005;**236**(1-2):1-7.
- 635 5. Franco NH, Olsson IAS. "How sick must your mouse be?"-An analysis of the use of animal
636 models in Huntington's disease research. Alternatives to laboratory animals: ATLA
637 2012;**40**(5):271-83.
- 638 6. Franco NH, Correia-Neves M, Olsson IAS. Animal Welfare in Studies on Murine Tuberculosis:
639 Assessing Progress over a 12-Year Period and the Need for Further Improvement. PLoS
640 one 2012;**7**(10):e47723.
- 641 7. Franco NH, Correia-Neves M, Olsson IAS. How "humane" is your endpoint?—Refining the
642 science-driven approach for termination of animal studies of chronic infection. PLoS
643 pathogens 2012;**8**(1):e1002399.
- 644 8. Solomon JA, Tarnopolsky MA, Hamadeh MJ. One universal common endpoint in mouse
645 models of amyotrophic lateral sclerosis. PLoS one 2011;**6**(6):e20582.
- 646 9. Morton DB. Humane endpoints in animal experimentation for biomedical research: ethical,
647 legal and practical aspects: London: Royal Society of Medicine Press, 1999:5-12.
- 648 10. Sawiak S, Wood N, Williams G, et al. Use of magnetic resonance imaging for anatomical
649 phenotyping of the R6/2 mouse model of Huntington's disease. Neurobiology of
650 Disease 2009;**33**(1):12-19.
- 651 11. Hockly E, Woodman B, Mahal A, et al. Standardization and statistical approaches to
652 therapeutic trials in the R6/2 mouse. Brain research bulletin 2003;**61**(5):469-79.
- 653 12. Menalled L, Brunner D. Animal models of Huntington's disease for translation to the clinic:
654 Best practices. Movement Disorders 2014;**29**(11):1375-90.
- 655 13. van der Worp HB, Howells DW, Sena ES, et al. Can animal models of disease reliably inform
656 human studies? PLoS Medicine 2010;**7**(3):e1000245.
- 657 14. Ioannidis JP. Why most published research findings are false. PLoS Medicine
658 2005;**2**(8):e124.
- 659 15. Schnabel J. Neuroscience: standard model. Nature News 2008;**454**(7205):682-85.
- 660 16. Ioannidis JPA, Greenland S, Hlatky MA, et al. Increasing value and reducing waste in
661 research design, conduct, and analysis. The Lancet 2014;**383**(9912):166-75.
- 662 17. Festing MFW. Randomized Block Experimental Designs Can Increase the Power and
663 Reproducibility of Laboratory Animal Experiments. ILAR Journal 2014;**55**(3):472-76.
- 664 18. Garner JP. The Significance of Meaning: Why Do Over 90% of Behavioral Neuroscience
665 Results Fail to Translate to Humans, and What Can We Do to Fix It? ILAR Journal
666 2014;**55**(3):438-56.
- 667 19. Lapchak PA. Scientific Rigor Recommendations for Optimizing the Clinical Applicability of
668 Translational Research. Journal of neurology & neurophysiology 2012;**3**.

- 669 20. Steward O, Balice-Gordon R. Rigor or Mortis: Best Practices for Preclinical Research in
670 Neuroscience. *Neuron* 2014;**84**(3):572-81.
- 671 21. Tsilidis KK, Panagiotou OA, Sena ES, et al. Evaluation of excess significance bias in animal
672 studies of neurological diseases. *PLoS biology* 2013;**11**(7):e1001609.
- 673 22. Button KS, Ioannidis JPA, Mokrysz C, et al. Power failure: why small sample size undermines
674 the reliability of neuroscience. *Nat Rev Neurosci* 2013;**14**(5):365-76.
- 675 23. Vesterinen HM, Sena ES, French-Constant C, et al. Improving the translational hit of
676 experimental treatments in multiple sclerosis. *Multiple Sclerosis Journal*
677 2010;**16**(9):1044-55.
- 678 24. Sena ES, van der Worp HB, Bath PMW, et al. Publication Bias in Reports of Animal Stroke
679 Studies Leads to Major Overstatement of Efficacy. *PLoS Biol* 2010;**8**(3):e1000344.
- 680 25. Steward O, Popovich PG, Dietrich WD, et al. Replication and reproducibility in spinal cord
681 injury research. *Exp Neurol* 2012;**233**(2):597-605.
- 682 26. Shineman DW, Basi GS, Bizon JL, et al. Accelerating drug discovery for Alzheimer's disease:
683 best practices for preclinical animal studies. *Alzheimers Res Ther* 2011;**3**(5):28.
- 684 27. Kimmelman J, London AJ, Ravina B, et al. Launching invasive, first-in-human trials against
685 Parkinson's disease: Ethical considerations. *Movement Disorders* 2009;**24**(13):1893-
686 901.
- 687 28. Scott S, Kranz JE, Cole J, et al. Design, power, and interpretation of studies in the standard
688 murine model of ALS. *Amyotroph Lateral Scler* 2008;**9**(1):4-15.
- 689 29. Collins FS, Tabak LA. NIH plans to enhance reproducibility. *Nature* 2014;**505**(7485):612.
- 690 30. Cressey D. UK funders demand strong statistics for animal studies. *Nature*
691 2015;**520**(7547):271.
- 692 31. Kilkenny C, Parsons N, Kadyszewski E, et al. Survey of the Quality of Experimental Design,
693 Statistical Analysis and Reporting of Research Using Animals. *PLoS one*
694 2009;**4**(11):e7824.
- 695 32. Crossley NA, Sena E, Goehler J, et al. Empirical evidence of bias in the design of
696 experimental stroke studies: a metaepidemiologic approach. *Stroke; a journal of*
697 *cerebral circulation* 2008;**39**(3):929-34.
- 698 33. Hirst JA, Howick J, Aronson JK, et al. The need for randomization in animal trials: an
699 overview of systematic reviews. *PLoS one* 2014;**9**(6):e98856.
- 700 34. van der Worp HB, Sena ES, Donnan GA, et al. Hypothermia in animal models of acute
701 ischaemic stroke: a systematic review and meta-analysis. *Brain : a journal of neurology*
702 2007;**130**(Pt 12):3063-74.
- 703 35. von Roten FC. Public perceptions of animal experimentation across Europe. *Public*
704 *Understanding of Science* 2012.
- 705 36. Lund TB, Mørkbak MR, Lassen J, et al. Painful dilemmas: A study of the way the public's
706 assessment of animal research balances costs to animals against human benefits.
707 *Public Understanding of Science* 2014;**23**(4):428-44.
- 708 37. Rollin BE. *Science and Ethics*: Cambridge University Press, 2006.
- 709 38. Ludolph AC, Bendotti C, Blaugrund E, et al. Guidelines for the preclinical in vivo evaluation
710 of pharmacological active drugs for ALS/MND: report on the 142nd ENMC
711 international workshop. *Amyotrophic lateral sclerosis : official publication of the World*
712 *Federation of Neurology Research Group on Motor Neuron Diseases* 2007;**8**(4):217-23.
- 713 39. Kanning KC, Kaplan A, Henderson CE. Motor neuron diversity in development and disease.
714 *Annual review of neuroscience* 2010;**33**:409-40.
- 715 40. Lever TE, Gorsek A, Cox KT, et al. An animal model of oral dysphagia in amyotrophic lateral
716 sclerosis. *Dysphagia* 2009;**24**(2):180-95.
- 717 41. Lever TE, Simon E, Cox KT, et al. A mouse model of pharyngeal dysphagia in amyotrophic
718 lateral sclerosis. *Dysphagia* 2010;**25**(2):112-26.
- 719 42. Boylan K, Yang C, Crook J, et al. Immunoreactivity of the phosphorylated axonal
720 neurofilament H subunit (pNF-H) in blood of ALS model rodents and ALS patients:

- 721 evaluation of blood pNF-H as a potential ALS biomarker. Journal of neurochemistry
722 2009;**111**(5):1182-91.
- 723 43. Kato S, Kato M, Abe Y, et al. Redox system expression in the motor neurons in amyotrophic
724 lateral sclerosis (ALS): immunohistochemical studies on sporadic ALS, superoxide
725 dismutase 1 (SOD1)-mutated familial ALS, and SOD1-mutated ALS animal models. Acta
726 neuropathologica 2005;**110**(2):101-12.
- 727 44. Gurney ME, Pu HF, Chiu AY, et al. Motor-neuron degeneration in mice that express a
728 human Cu,Zn superoxide-dismutase mutation. Science 1994;**264**(5166):1772-75.
- 729 45. Bruijn LI, Becher MW, Lee MK, et al. ALS-linked SOD1 mutant G85R mediates damage to
730 astrocytes and promotes rapidly progressive disease with SOD1-containing inclusions.
731 Neuron 1997;**18**(2):327-38.
- 732 46. Marcuzzo S, Zucca I, Mastropietro A, et al. Hind limb muscle atrophy precedes cerebral
733 neuronal degeneration in G93A-SOD1 mouse model of amyotrophic lateral sclerosis: a
734 longitudinal MRI study. Exp Neurol 2011;**231**(1):30-7.
- 735 47. Neymotin A, Calingasan NY, Wille E, et al. Neuroprotective effect of Nrf2/ARE activators,
736 CDDO ethylamide and CDDO trifluoroethylamide, in a mouse model of amyotrophic
737 lateral sclerosis. Free Radical Biology and Medicine 2011;**51**(1):88-96.
- 738 48. Del Signore SJ, Amante DJ, Kim J, et al. Combined riluzole and sodium phenylbutyrate
739 therapy in transgenic amyotrophic lateral sclerosis mice. Amyotrophic lateral sclerosis :
740 official publication of the World Federation of Neurology Research Group on Motor
741 Neuron Diseases 2009;**10**(2):85-94.
- 742 49. Tada S, Okuno T, Yasui T, et al. Deleterious effects of lymphocytes at the early stage of
743 neurodegeneration in an animal model of amyotrophic lateral sclerosis. Journal of
744 neuroinflammation 2011;**8**(1):19.
- 745 50. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika
746 1951;**16**(3):297-334.
- 747 51. Pregibon D. Goodness of link tests for generalized linear models. Applied statistics 1980:15-
748 14.
- 749 52. Breusch TS, Pagan AR. A simple test for heteroscedasticity and random coefficient
750 variation. Econometrica: Journal of the Econometric Society 1979:1287-94.
- 751 53. Ramsey JB. Tests for specification errors in classical linear least-squares regression analysis.
752 Journal of the Royal Statistical Society Series B (Methodological) 1969:350-71.
- 753 54. Franco NH, Olsson IAS. Is the ethical appraisal of protocols enough to ensure best practice
754 in animal research? Alternatives to laboratory animals: ATLA 2013;**41**(1):P5-7.
- 755 55. Olsson IAS, Hansen AK, Sandoe P. Animal welfare and the refinement of neuroscience
756 research methods - a case study of Huntington's disease models. Lab Anim
757 2008;**42**(3):277-83.
- 758 56. Lilley E, Hawkins P, Jennings M. A 'road map' toward ending severe suffering of animals
759 used in research and testing. ATLA - Alternatives to Laboratory Animals
760 2014;**42**(4):267-72.
- 761 57. Boillée S, Yamanaka K, Lobsiger CS, et al. Onset and progression in inherited ALS
762 determined by motor neurons and microglia. Science 2006;**312**(5778):1389-92.
- 763 58. Kimmelman J, Mogil JS, Dirnagl U. Distinguishing between Exploratory and Confirmatory
764 Preclinical Research Will Improve Translation. PLoS biology 2014;**12**(5):e1001863.
- 765 59. Benatar M. Lost in translation: Treatment trials in the SOD1 mouse and in human ALS.
766 Neurobiology of Disease 2007;**26**(1):1-13.
- 767 60. Reilly J, Blackshaw AW. *Euthanasia of animals used for scientific purposes*: ANZCCART,
768 2001.
- 769 61. Artwohl J, Brown P, Corning B, et al. Report of the ACLAM Task Force on Rodent
770 Euthanasia. Journal of the American Association for Laboratory Animal Science
771 2006;**45**(1):98-105.

- 772 62. Bame M, Pentia PA, Needleman R, et al. Effect of Sex on Lifespan, Disease Progression,
773 and the Response to Methionine Sulfoximine in the SOD1 G93A Mouse Model for ALS.
774 *Gender Medicine* 2012;**9**(6):524-35.
- 775 63. McCombe PA, Henderson RD. Effects of gender in amyotrophic lateral sclerosis. *Gender*
776 *Medicine* 2010;**7**(6):557-70.
- 777 64. Alves CJ, de Santana LP, Santos AJDd, et al. Early motor and electrophysiological changes in
778 transgenic mouse model of amyotrophic lateral sclerosis and gender differences on
779 clinical outcome. *Brain Research* 2011;**1394**(0):90-104.
- 780 65. Hooijmans CR, Leenaars M, Ritskes-Hoitinga M. A gold standard publication checklist to
781 improve the quality of animal studies, to fully integrate the Three Rs, and to make
782 systematic reviews more feasible. *Alternatives to laboratory animals : ATLA*
783 2010;**38**(2):167-82.
- 784 66. Banwell V, Sena ES, Macleod MR. Systematic review and stratified meta-analysis of the
785 efficacy of interleukin-1 receptor antagonist in animal models of stroke. *Journal of*
786 *stroke and cerebrovascular diseases : the official journal of National Stroke Association*
787 2009;**18**(4):269-76.
- 788 67. Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research
789 evidence. *Lancet* 2009;**374**(9683):86-9.
- 790 68. Festing MF, Altman DG. Guidelines for the design and statistical analysis of experiments
791 using laboratory animals. *ILAR journal / National Research Council, Institute of*
792 *Laboratory Animal Resources* 2002;**43**(4):244-58.
- 793 69. Festing MFW. Warning: the use of heterogeneous mice may seriously damage your
794 research. *Neurobiology of Aging* 1999;**20**(2):237-44.
- 795 70. Mead RJ, Bennett EJ, Kennerley AJ, et al. Optimised and Rapid Pre-clinical Screening in the
796 SOD1^{G93A} Transgenic Mouse Model of Amyotrophic Lateral Sclerosis
797 (ALS). *PloS one* 2011;**6**(8):e23244.
- 798 71. Perrin S. Preclinical research: Make mouse studies work. *Nature* 2014;**507**:423-25.
- 799 72. Landis SC, Amara SG, Asadullah K, et al. A call for transparent reporting to optimize the
800 predictive value of preclinical research. *Nature* 2012;**490**(7419):187-91.
- 801 73. Bara M, Joffe AR. The methodological quality of animal research in critical care: the public
802 face of science. *Annals of intensive care* 2014;**4**(1):1-9.
- 803 74. Howells DW, Macleod MR. Evidence-based Translational Medicine. *Stroke; a journal of*
804 *cerebral circulation* 2013;**44**(5):1466-71.
- 805 75. Bebarta V, Luyten D, Heard K. Emergency medicine animal research: does use of
806 randomization and blinding affect the results? *Academic Emergency Medicine*
807 2003;**10**(6):684-87.
- 808 76. Macleod MR, van der Worp HB, Sena ES, et al. Evidence for the efficacy of NXY-059 in
809 experimental focal cerebral ischaemia is confounded by study quality. *Stroke; a journal*
810 *of cerebral circulation* 2008;**39**(10):2824-29.
- 811 77. Perrin S. Preclinical research: Make mouse studies work. *Nature* 2014;**507**(7493):423-25.
- 812 78. Avey MT, Moher D, Sullivan KJ, et al. The Devil Is in the Details: Incomplete Reporting in
813 Preclinical Animal Research. *PloS one* 2016;**11**(11):e0166733.
- 814 79. Vogt L, Reichlin TS, Nathues C, et al. Authorization of Animal Experiments Is Based on
815 Confidence Rather than Evidence of Scientific Rigor. *PLoS biology*
816 2016;**14**(12):e2000598.
- 817 80. Baker D, Lidster K, Sottomayor A, et al. Two years later: journals are not yet enforcing the
818 ARRIVE guidelines on reporting standards for pre-clinical animal studies. *PLoS biology*
819 2014;**12**(1):e1001756.
- 820 81. Gulin JEN, Rocco DM, García-Bournissen F. Quality of Reporting and Adherence to ARRIVE
821 Guidelines in Animal Studies for Chagas Disease Preclinical Drug Research: A
822 Systematic Review. *PLOS Neglected Tropical Diseases* 2015;**9**(11):e0004194.

- 823 82. Macleod MR, Lawson McLean A, Kyriakopoulou A, et al. Risk of Bias in Reports of In Vivo
824 Research: A Focus for Improvement. *PLoS biology* 2015;**13**(10):e1002273.
- 825 83. Ting KH, Hill CL, Whittle SL. Quality of reporting of interventional animal studies in
826 rheumatology: a systematic review using the ARRIVE guidelines. *International Journal*
827 *of Rheumatic Diseases* 2015;**18**(5):488-94.
- 828 84. Muhlhausler BS, Bloomfield FH, Gillman MW. Whole animal experiments should be more
829 like human randomized controlled trials. *PLoS biology* 2013;**11**(2):e1001481.
- 830 85. McGonigle P, Ruggeri B. Animal models of human disease: Challenges in enabling
831 translation. *Biochemical Pharmacology* 2014;**87**(1):162-71.
- 832 86. Hackam DG. Translating animal research into clinical benefit. *BMJ: British Medical Journal*
833 2007;**334**(7586):163.
- 834 87. de Vries RBM, Wever KE, Avey MT, et al. The Usefulness of Systematic Reviews of Animal
835 Experiments for the Design of Preclinical and Clinical Studies. *ILAR Journal*
836 2014;**55**(3):427-37.
- 837 88. Jansen of Lorkeers SJ, Doevendans PA, Chamuleau SAJ. All preclinical trials should be
838 registered in advance in an online registry. *European Journal of Clinical Investigation*
839 2014;**44**(9):891-92.
- 840 89. Dal-Ré R, Ioannidis JP, Bracken MB, et al. Making prospective registration of observational
841 research a reality. *Science translational medicine* 2014;**6**(224):224cm1-24cm1.
- 842 90. Rollin BE. Animal Research, Animal Welfare, and the Three R's *The Journal of Philosophy,*
843 *Science & Law* 2010;**10**.
- 844 91. Osborne NJ, Phillips BJ, Westwood K. Journal editorial policies as a driver for change -
845 animal welfare and the 3R. *New Paradigms In Laboratory Animal Science - Proceedings*
846 *of the Eleventh FELASA symposium and the 40th Scand-LAS Symposium. Helsinki,*
847 2010:18-23.
- 848 92. Martins AR, Franco NH. A Critical Look at Biomedical Journals' Policies on Animal Research
849 by Use of a Novel Tool: The EXEMPLAR Scale. *Animals* 2015;**5**(2):315-31.
- 850 93. Editorial. Checklists work to improve science. *Nature* 2018;**556**:273-74.

851

852

1 **Methodological standards, quality of reporting, and regulatory compliance in**
2 **animal research on amyotrophic lateral sclerosis: a systematic review**

3 Joana G Fernandes^{1,2¶}, Nuno H Franco^{1,2¶}, Andrew J Grierson³, Jan Hultgren⁴, Andrew JW
4 Furley^{5&}, I Anna S Olsson^{1,2&*}

5 ¹ Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Portugal

6 ² IBMC-Instituto de Biologia Molecular e Celular, Universidade do Porto, Portugal

7 ³ Sheffield Institute for Translational Neuroscience, Department of Neuroscience, University of
8 Sheffield, Sheffield, United Kingdom

9 ⁴ Department of Animal Environment and Health, Swedish University of Agricultural Sciences,
10 Skara, Sweden

11 ⁵ Department of Biomedical Science, University of Sheffield, Western Bank, Sheffield, United
12 Kingdom

13 *Corresponding author

14 E-mail: olsson@ibmc.up.pt

15 ¶ These authors share the first authorship

16 & These authors share the last authorship

17

18 **Key words:** Amyotrophic Lateral Sclerosis, ALS, Guidelines, Methodology, Reporting, Quality,
19 Compliance, Animal Welfare, Reproducibility

20

21 **Abstract**

22 **Objectives**

23 The ALS research community was one of the first to adopt methodology guidelines to improve
24 preclinical research reproducibility. We here present results of a systematic review to
25 investigate how the standards in this field changed over the ten-year period during which the
26 guidelines were first published (2007) and updated (2010).

27

28 **Methods**

29 We searched for papers reporting ALS research on SOD1 mice published between 2005 and 2015
30 on the ISI Web of Science® database, resulting in a sample of 569 papers to review, after triage.

31 Two scores – one for methodological quality, one for regulatory compliance – were built from

32 weighted sums of separate sets of items, and subjected to multivariable regression analysis, to
33 assess how these related to publication year, type of study, country of origin and journal.

34

35 **Results**

36 Reporting standards improved over time. Of papers published after the first ALS guidelines were
37 made public, fewer than 9% referred specifically to these. Of key research parameters, only
38 three (genetic background, number of transgenes and group size) were reported in >50% of the
39 papers. Information on housing conditions, randomization and blinding were absent in over two
40 thirds of papers. Group size was among the best reported parameters, but the majority reported
41 using fewer than the recommended sample size and only two studies clearly justified group size.

42

43 **Conclusions**

44 General methodological standards improved gradually over an 8- to 10-year period but
45 remained generally comparable to related fields with no specific guidelines, except with
46 regard to severity. only 11% of ALS studies were classified in the highest severity level
47 (animals allowed to reach death or moribund stages), substantially below the proportion in
48 studies of comparable neurodegenerative diseases such as Huntington's. The existence of
49 field-specific guidelines, though a welcome indication of concern, seems insufficient to
50 ensure adherence to high methodological standards. Other mechanisms may be required to
51 improve methodological and welfare standards.

52 **Strengths and limitations:**

53 - The approach for this systematic review is unique in covering methodological quality,
54 regulatory compliance, and severity or animal welfare.

55 - We built two comprehensive scores (for methodological standards and for regulatory
56 compliance) which were subjected to multivariable regression analysis to investigate how these
57 scores were related to publication year, type of study, country of origin and journal,
58 simultaneously accounting for all these factors.

59 - Our large sample (N=569 papers) included half the total population of published papers
60 between 2005 and 2015.

61 - While more models of ALS are now available, only studies using the SOD-1 mouse were
62 included.

63 - The protocol was defined prior to data collection but was not registered prior to the study.

64 - Information retrieval and assessment was not blinded.

65 **1. Introduction**

66 Amyotrophic lateral sclerosis (ALS) is a rapidly progressing neurodegenerative disease typically
67 resulting in death two to five years after the onset of symptoms. There is no known cure and the
68 most widely used treatment– riluzole – extends survival by just two months¹. ALS research using
69 animal models focuses primarily on two main interconnected goals: understanding the
70 underlying mechanisms involved in motor neuron death in the brain and spinal cord, and
71 development and testing of potential drug therapies². This research relies substantially on
72 genetically modified animals, particularly transgenic mice expressing mutant forms of the
73 human Superoxide Dismutase 1 (SOD1) gene, which manifest several important characteristics
74 of the human disease^{3,4}.

75

76 While the use of animal models is relevant for advancing knowledge and considered essential
77 for testing putative treatments, it also presents ethical challenges and thus may be a reason for
78 public concern. As a result, a common legal requirement in many countries is that animal
79 research projects undergo an evaluation process intended to ensure that protocols are designed
80 and carried out in compliance with the 3Rs principle: *replacement* of animal use by non-animal
81 methods, *reduction* of animal numbers needed to achieve the scientific objectives, and
82 *refinement* of procedures to reduce or prevent harm to animals and improve their wellbeing.
83 Systematic reviews of animal use in both neuroscience⁵ and infection⁶ research indicate that
84 self-reported regulatory compliance – including of ethical approval of protocols – has steadily
85 increased over the last decade, but that significant progress could still be made to minimise and
86 prevent avoidable suffering of laboratory animals. One key measure for accomplishing this is the
87 termination of experiments during less severe stages of disease development where it is
88 scientifically valid to do so. Endpoints based on early obtainable and scientifically sound
89 indicators of phenotype progression can not only improve the ethical acceptability of animal
90 studies, but also prevent the confounding influence of secondary factors; in the case of animal
91 models of neurodegenerative diseases, starvation and dehydration arising from difficulties in
92 eating and drinking due to progressive motor impairment can affect the phenotype and the
93 readout of survival studies⁷⁻⁹. Simple refinements – such as adding mash food and longer bottle
94 spouts – can however help reduce the influence of such factors¹⁰⁻¹².

95

96 Of related concern are reports that a number of published animal studies fail to uphold basic
97 standards regarding experimental design – e.g. random assignment of animals to treatment
98 groups, blinding of observers – or use too few animals often leading to irreproducible results of

99 limited translational value¹³⁻¹⁸. This also holds true for neuroscience¹⁹⁻²², with concerns over the
100 overall quality and reproducibility of published results being raised for several neuroscience sub-
101 fields, including multiple sclerosis²³, stroke²⁴, spinal cord injury²⁵ Alzheimer's²⁶, Parkinson's²⁷,
102 Huntington's¹² and ALS²⁸ research. This has led major science funders, including the National
103 Institutes of Health²⁹ and Research Councils UK³⁰ to demand that future grant proposals attest
104 to the likelihood of providing reliable results, by including details of experimental design and
105 adequate justification of sample sizes. Reproducibility is further hindered by insufficient
106 provision of information on methodology in published research³¹ – including failure to account
107 for key variables such as sex, genotype, age, and weight of animals, anaesthetics used or
108 methods of euthanasia. Omitting information also makes it impossible to evaluate the study
109 quality and there is evidence that papers that do not report randomization or blinding
110 exaggerate biological effects³²⁻³⁴.

111

112 Broadly, the public conditionally approves of animal studies on the assumption that the harm
113 caused is offset by the benefits achieved and that scientists strive to minimise the former and
114 optimise the latter^{35,36}. Doing so requires scientists to critically revise their methods to maximise
115 translational relevance^{18,37}. Scientists are rightly concerned and, within the self-correcting
116 process of science, must rely on themselves to both identify the main obstacles hindering its
117 progress and find adequate solutions. To address the issue of methodological standards and
118 quality of reporting of basic and applied ALS studies, the ALS research community held two
119 meetings in 2006 and 2009, resulting in the publication of guidelines for animal studies in this
120 field^{2,38}. These guidelines aim to improve and standardise research methodology, and
121 encourage authors and journals to publish negative results in order to avoid publication bias.
122 The actual impact of such guidelines on how the ALS community carries out and reports research
123 has however not been assessed.

124

125 The present systematic review of animal studies of ALS uniquely aimed to assess, over an
126 extended period, the attention given to relevant methodological parameters (as a proxy for the
127 likely reliability of the study) and to examine how the principles of *refinement* and *reduction*
128 (measures to minimise animal harm) were considered. Both proof-of-concept and preclinical
129 studies were included in order to assess the influence of type of study.

130 2. Methods

131 2.1 Database search

132 An advanced search was conducted on the *ISI Web of Science*[®] database with the query *TS =*
133 *((mice OR mouse) SAME (ALS OR "amyotrophic lateral sclerosis"))*. The database choice followed
134 the protocol established for our previous reviews^{5 6}, based on considerations of access, search
135 function and wide coverage of life sciences research. Results were refined to include only
136 original research articles written in English and published in 2005, 2007, 2009, 2011, 2013 and
137 2015. Years of publication were selected to include papers reporting research planned and
138 carried out prior to and after the publication of guidelines for ALS research in 2007³⁸ and 2010
139 ², resulting from two international meetings held in 2006 and 2009, respectively (Figure 1).

140

141

142 **Figure 1. Timeline of relevant events.** The bottom arrows signal the years for which papers in our sample
143 were retrieved and the top arrows indicate the years when workshops on best practice in ALS animal
144 research were held, as well as when guidelines stemming from these were published. The grey bars
145 illustrate the 1-4 year period over which ALS animal studies reported in 2005 were likely to have been
146 designed and carried out, an estimation that can also be applied for the other years reviewed (2007, 2009,
147 2011, 2013, and 2015).

148

149

150 The choice to focus on SOD-1 mice was based on the predominant role of this model in animal-
151 based research into ALS (see Supplementary Figure 1).

152

153

154 **Supplementary Figure 1** - Trends in animal model chosen in ALS research, based on the number of hits
155 from an *Clarivate Analytics Web of Science*[®] advanced search applying the search queries: a) *TS = (("ALS"*
156 *OR "amyotrophic lateral sclerosis") AND "SOD1" AND ("mouse" OR "mice"))*; b) *TS = (("ALS" OR*
157 *"amyotrophic lateral sclerosis") AND "TDP-43" AND ("mouse" OR "mice"))*; and c) *TS = (("ALS" OR*
158 *"amyotrophic lateral sclerosis") AND "FUS" AND ("mouse" OR "mice"))*

159

160

161 The search was performed in February 2013 for scientific articles from 2009 and 2011 (after the
162 first and second conferences, respectively), in August 2013 for scientific articles from 2005
163 (before the two conferences), in September 2014 for scientific articles from 2013, in November
164 2016 for scientific articles from 2015, and in February 2017 for scientific articles from 2007. After

165 the triage process, illustrated in Figure 2, 569 full-text articles remained for analysis: 77 from
 166 2005, 81 from 2007, 84 from 2009, 106 from 2011, 115 from 2013, and 106 from 2015.

167

168 **Figure 2. Triage process.** The first triage step involved reading each of the 1993 abstracts and excluding
 169 all papers that were not related to ALS. The second triage step excluded all papers that did not report
 170 original research with SOD1 models of the disease.

171

172 2.2 Data collection

173 Each published study was categorised as either a ‘preclinical’ (i.e., carried out “to evaluate a
 174 drug for use in humans”) or ‘proof-of-concept’ (i.e., aiming “to elucidate the mechanism of the
 175 disease”), according to the suggested classification for animal studies on ALS^{2 38}. Thus, papers
 176 reporting outcomes of drug tests in animal models to inform of their therapeutic value for
 177 humans were classified as ‘preclinical’, whereas those reporting studies which primary goal was
 178 to decipher a mechanism of the disease without an immediate application to therapeutic
 179 approaches in humans – regardless of using a drug as an investigational tool – were classified as
 180 ‘proof-of-concept’. Table 1 describes the information retrieved regarding regulatory
 181 compliance, animal models, experimental design and animal welfare. This information was
 182 retrieved through careful reading of the full papers, and logged into a spreadsheet.

183

184 The review protocol was defined prior to data collection. No modifications to data collection
 185 methods were made during the research, but the period to be covered was extended to include
 186 publication year 2015. Data extraction was carried out by JGF, with support from NHF, AJG and
 187 IASO for disambiguation. Blinding was not possible as access to the full paper was required in
 188 order to retrieve information.

189

190 **Table 1. Data retrieved.** A description of the information collected from revised papers is presented for
 191 each item.

Category	Items	Description/Observations
Regulatory compliance	Ethical approval	Studies explicitly reported to be approved by a committee / authority.
	Guideline compliance	Articles that did not report having experimental protocols ethically approved by an institution or national entity, but reported that some kind of guidelines for use and care of laboratory animals was followed.
Animal models	Genetic background	When available.
	Sex	Four options: Male, female, both or not reported. For <i>both</i> , information on whether studies were balanced for gender was retrieved.
	Number of transgene copies	When available.

Experimental design	Group size	Mean group size, based on the available information
	Randomization	Studies explicitly reporting assigning animals to groups randomly
	Blinding	Studies explicitly reporting blinding of observers to experimental groups
	Non-transgenic littermate control	Studies explicitly reporting the use of non-transgenic littermates as control.
	Splitting littermates into groups	Studies explicitly reporting that littermates were split into groups.
	Housing and husbandry conditions	Reporting information regarding temperature, humidity, light of the room where animals were kept, and cage size and number of animals per cage.
Animal Welfare/ Procedures	Severity	Described in table 2.
	Refinement	Relevant refinements to minimise suffering and distress, such as housing adaptations.
	Euthanasia method	Euthanasia methods were divided into the following categories: "Under anaesthesia" (including anaesthetic overdose); "CO ₂ asphyxiation"; "Other"; "Not reported" and "Not performed".

192

193 For severity assessment, a scale was devised based on the specific characteristics of the ALS
 194 models and their progressive disease phenotype (Table 2). The ALS models used in the reviewed
 195 studies express diverse mutant forms of the *SOD1* gene. The onset of disease for these models
 196 is generally characterised by weakness and tremors of the hind limbs, together with a mild loss
 197 of body weight. Disease progression leads to paralysis of hind limbs, followed by complete
 198 paralysis (example, Figure 3 in ³⁹), accompanied by increased difficulty to eat, drink and swallow
 199 ^{40 41}. Mice die of respiratory failure due to paralysis of the diaphragm ⁸. Age of onset and death,
 200 as well as the interval between them, vary depending on the mutation of the amino-acid and
 201 codon e.g. ⁴², number of copies of transgene e.g. ⁴³, and genetic background ⁴. For instance, the
 202 over-expressing SOD1G93A Line Gur 1H (B6SJL hybrid) presents with an early onset of overt
 203 motor symptoms (3-4 months) and moderate rate of progression (3 weeks from onset to death)
 204 ⁴⁴, whereas the highly expressing SOD1G85R Line 148 presents with later onset (7.5 months) and
 205 faster disease progression (2 weeks from onset to death) ⁴⁵. Also, factors such as the animal
 206 supplier (e.g. ^{46 47}), in-house breeding ⁴⁸ and crosses with other non-SOD1 models (e.g. SOD1
 207 mice crossed with gene-specific knockout mice ⁴⁹) are further sources of variability.

208 Maximum estimated severity was classified according to a five-level scale (Table 2). Scoring was
 209 based on the estimated clinical state of animals at the most advanced stage of disease
 210 progression they were allowed to reach. Studies in which information was insufficient to draw
 211 conclusions about the level of severity were classified as 'undetermined'. This severity scale was
 212 developed building upon previous work from members of this team (NF, AO) developed for
 213 classifying studies on mouse models of Huntington's disease (table 2 in ⁵), together with our own
 214 (AG) experience with mutant SOD1 mouse models and literature. For purposes of statistical
 215 analysis, the severity scale was reduced to a binary scale, ("low"= severity up to level 4; "high"=
 216 level 5 severity. The choice for above level-4 severity as a cut-off point, was based on its status

217 as a "standard endpoint" in published ALS guidelines ^{2 38}, whereas full paralysis or spontaneous
 218 death exceeds this standard endpoint, as well as the legally recommended endpoints in many
 219 countries, including the EU Member States.

220 **Table 2.** Severity scale for ALS studies on transgenic mice with a mutant SOD1 gene. Each severity level
 221 exemplified from the most commonly used B6.Cg-TgN-(SOD1G93A) G1H mouse. Classification was based
 222 on the most severe endpoint used in each publication.
 223

Severity	Description	Welfare issues during this stage
Level 1	Animals euthanized prior to disease onset, which is characterised by progressive weight loss or hind limb tremors	No overt motor dysfunction. Phenotype is subclinical. Loss of motor function can be detected using rotarod or running wheels, but does not interfere with normal behaviour
Level 2	Studies terminated at an early stage of disease: animals present trembling and weakness in hind limbs (by approx. 75d) and mild body weight loss.	Minor. Loss of motor function can be detected using rotarod or running wheels, but has little interference with normal behaviour.
Level 3	Experiments terminated when animals are no longer able to reach food hopper or bottle spout. This occurs when animals reach a moderate (gait abnormalities and weakness) to severe (hind limb paralysis) stage of motor impairment (usually at 120-125d)	Medium. Loss of motor function and body weight can be detected by monitoring (e.g. by a clinical score sheet) and by checking self-righting ability. Refinement measures to address these welfare issues include provision of softer bedding material (e.g. sawdust), elongated bottle spouts and mashed food on the cage floor.
Level 4	Animals euthanized after losing the ability to right themselves within 10-30 seconds after being laid on either side (one or both) or when percentage of weight loss reaches 15-20% of peak body weight (usually at 130-140d)	Major. Animals show severe locomotor impairment. Refinement as described for level 3
Level 5	Animals are euthanized when reaching a moribund stage (complete paralysis) or allowed to die spontaneously	Severe. At this stage, animals are unable to move, eat or drink. Animals which are not euthanized will die as a result of respiratory failure.

224

225 **2.3. Methodological Standards Reporting (MSR) and Regulatory Compliance Reporting (RCR)**
 226 **scores**

227 For each reviewed publication, data were collected on a number of items which all contributed
 228 with information about the reporting quality of the paper. For the analysis, we brought these
 229 items together into two scores, hence generating for each paper two comprehensive measures
 230 for reporting quality, one on methodological standards and one on regulatory compliance. We
 231 then used regression analysis to investigate how the two scores (dependent variables) were
 232 related to publication year, type of study, country of origin and journal (explanatory or predictor
 233 variables), as outlined in detail in the following. Based on the regression models it is possible to

234 predict how the dependent variables would have changed with changes in the explanatory
235 variables. In contrast to, for example, correlation, the regression analysis takes into account all
236 the explanatory variables that were included in the models, i.e. the estimated association
237 between a score and one of the explanatory variables is independent of the values of the other
238 explanatory variables considered. In that way, spurious associations caused by relationships
239 between the explanatory variables in the data can be avoided.

240 The two scores were formed as weighted sums of separate sets of items. The Methodological
241 Standards Reporting (MSR) score was constructed as the weighted sum of the items *sampsize*,
242 *climate*, *cagesize*, *nmice*, *sex*, *copies*, *genetic* (which refer to important research parameters in
243 animal experimentation and in ALS research in particular) and the items *random*, *blinded*,
244 *control*, *sibsplit*, and *exclus* (associated with general good practices in the design of animal
245 experiments and published recommendations for ALS studies). Greater weight (1.5 versus 1)
246 was attributed to items which are also part of the ALS guidelines. Table 3 describes these items,
247 their attributed weight in the MSR score and the absolute number and percentage of papers
248 reporting this information, divided by type of study.

249 The Regulatory Compliance Reporting (RCR) score was originally constructed from the items
250 *comply*, *protocol*, *severity* (turned into a binary classification) and *refine*. For purposes of
251 statistical modelling, the final version of this score (RCRb) included *comply*, *protocol* and *refine*
252 and was coded as 1 when the sum of these was 2-3, and as 0 when the sum was 0-1.

253

254 **Table 3. List of items integrated in the MSR and the RCR scores for preclinical (n=108) and proof-of-**
255 **concept (n=461) animal studies on ALS reporting this information.** The score for each variable is provided
256 (MSR score ranging from 0 to 12.5, and RCR score ranging from 0 to 3). Greater weight (1.5 versus 1) was
257 attributed to items which are also part of the ALS guidelines. For purposes of statistical modelling, RCR
258 (only including items *comply*, *protocol* and *refine*) was later simplified to a binary variable RCRb coded as
259 1 for RCR values 2-3 and as 0 for RCR values 0-1.

260

261

262

263

264

265

266

267

268

269
270
271

Reported information	MSR score		'Proof-of-Concept' (n=461)		'Preclinical' (n=108)	
	Score item	Score weight	Absolute number	%	Absolute number	%
Relevant animal research variables						
Group size	<i>sampsize</i>	1.5	368	79.8	106	98.1
Environment: light, temp., humidity (fully or partially reported)	<i>climate</i>	1	123	26.7	42	38.9
Cage size	<i>cagesize</i>	1	1	0.2	2	1.9
Mice per cage	<i>nmice</i>	1	26	5.6	15	13.9
Sex of the animals	<i>sex</i>	1.5	223	48.4	71	65.7
Number of transgene copies	<i>copies</i>	1.5	286	62.0	80	74.1
Genetic background	<i>genetic</i>	1.5	349	75.7	92	85.2
Measures to reduce 'noise' and bias in experiments						
Animals randomised to treatment groups	<i>random</i>	1	28	6.1	47	43.5
Observers blinded to treatment	<i>blinded</i>	1.5	94	20.4	52	48.1
Non-transgenic littermate controls used	<i>control</i>	1	150	32.5	39	36.1
Splitting littermates into groups	<i>Sibsplit</i>	1	28	6.1	31	28.7
Reason for exclusion of animals is reported	<i>exclus</i>	1	2	0.4	6	5.6

Reported information	RCR score		'Proof-of-Concept' (n=461)		'Preclinical' (n=108)	
	Score item	Score weight	Absolute number	%	Absolute number	%
Self-reported compliance with laws and regulations	<i>comply</i>	1	98	21.3	28	25.9
Project approval reported	<i>protocol</i>	1	315	68.3	66	61.1
Refinement measures (e.g. to aid feed and hydrate) to aid feed and hydrate)	<i>refine</i>	1	29	6.3	14	13

272
273

274 MSR and RCRb were modelled statistically to estimate the effects of publication year (2005,
275 2007, 2009, 2011, 2013 or 2015), study type (preclinical or proof-of-concept), country of origin
276 (15 categories), journal (17 categories) and severity (low or high), simultaneously accounting for
277 all the explanatory variables in the models. Countries contributing with less than twelve papers,
278 and journals contributing with less than 6 papers, were combined into separate categories,
279 denoted 'Other'. MSR was modelled using linear regression and RCRb by logistic regression.
280 Logistic regression is appropriate for binary dependent variables (assuming a linear relationship
281 of the log-odds of the dependent variable with the explanatory variables). The results of a
282 logistic regression can be expressed as the odds of a positive value of the dependent variable at
283 one level of a categorical explanatory variable relative to the odds at another level (the odds
284 ratio), or the probability of a positive dependent variable at any given level of the explanatory
285 variables. All first-order interaction effects (combined effects of two explanatory variables at a
286 time) were tested and included if significant.

287 Predictive marginal means were calculated, showing the values of MSR and probabilities of RCR
288 >1, respectively, that the models predicted for different publication years, study types and
289 countries of origin, in each case assuming that the remaining variables in the models had their
290 observed values. Both models were checked using the Pregibon link test ⁵⁰, and by examining
291 standardised residuals, looking for model mis-specification and extreme values. The MSR model
292 was also checked with the Breusch-Pagan/Cook-Weisberg test for heteroscedasticity
293 ⁵¹(variability differing between parts of the data), the Ramsey regression specification-error test
294 for omitted variables ⁵², and the RCRb model by examining delta-betas to identify particularly
295 influential observations. The proportion of the total variation in MSR and RCRb that could be
296 explained by differences between countries or journals was determined by running empty mixed
297 models with country and journal, respectively, as a random effect, and calculating the intra-class
298 correlation coefficients. The justification for weighting the items composing MSR was checked
299 by modelling an alternative score formed without weighting. The differences between years and
300 countries remained virtually unchanged, although the unweighted score values were generally
301 lower.

302

303 The association between MSR and RCR scores was estimated using Spearman rank correlation,
304 which is suitable for non-Normally distributed data. A total of 490 observations could be used.
305 Overall MSR mean \pm SD was 5.69 ± 2.39 . RCR assumed values 0 (n=48), 1 (n=103), 2 (n=309) or
306 3 (n=30), resulting in 69% of the observations having values above 1. The number of
307 observations per level of year, study type, country, journal and severity is shown in
308 Supplementary Table 1.

309

310 The data were analysed in Stata/IC v. 13.1 and IBM SPSS 23.0. Each article was regarded as the
311 experimental unit and the level of significance for all tests was 0.05.

312

313 **3. Results**

314 **3.1. Quality of research and reporting**

315 The quality of methodological standards and of reporting is crucial to avoid bias and achieve
316 reliable, repeatable and translatable research results. We measured this through the
317 Methodological Standards Reporting Score and also looked at specific research parameters
318 individually.

319 3.1.1 Methodological Standards Reporting Score

320 The 12 items that comprise the Methodological Standards Reporting Score represent seven
321 relevant experimental variables and five measures for reducing bias in animal experiments.
322 Higher scores mean better reporting and implementation of good practices in the design of ALS
323 animal studies.

324 MSR was significantly affected by year and study type (joint F-test $p=0.0015$ and <0.0001 ,
325 respectively). Compared to 2005, the logistic regression model predicted a lower MSR for 2007.
326 However, the subsequent years (2009, 2011, 2013 and 2015) were all predicted to be higher
327 than 2007, with a consistent and unbroken increasing trend until 2013 (Figure 3). In 2013, MSR
328 was predicted to be 1.5 units higher than in 2007 ($p<0.0001$). The model also predicted a higher
329 MSR for preclinical studies than for proof-of-concept studies (marginal mean 7.28 and 5.26
330 respectively). Model diagnostics showed that linear regression was justified and the model fit
331 was excellent. Supplementary Table 2 shows the complete MSR model results.

332

333 **Figure 3. Predictive marginal means (predicted score values) \pm 95% confidence interval of publication**
334 **year (left panel) and country (right panel) based on a model of a Methodological Standards Reporting**
335 **(MSR) Score in 487 ALS studies.** According to the linear regression model, MSR could be expected to be
336 lower in 2007 than 2005, but higher in 2009, 2011, 2013 and 2015 than 2007. No significant interactions
337 were found (e.g. between country and year). According to the R-square statistic the model explained 25%
338 of the total variation in MSR.

339

340 3.1.2. Reporting of relevant research parameters

341 Some research parameters were very seldom reported, for example: numbers of animals per
342 cage (7.2%, 41/569); cage size (0.5%, 3/569) and exclusion of animals (1.4%, 8/569). Measures
343 in guideline recommendations to reduce bias in ALS research were mostly not reported,
344 including: splitting littermates to treatment groups (10.4%, 59/569); use of non-transgenic
345 littermates as controls (33.2%, 189/569); as well as measures of broader application, such as
346 random assignment of animals to treatments (13.2%, 75/569) or blinding of observers (25.7%,
347 146/569). By contrast, numbers of transgene copies and genetic backgrounds of animals were
348 reported in the majority of papers.

349

350 Of papers reporting sex ($n=297$), 54.2% (161/297) described studies using mice of both sexes,
351 while 29.0% (86/297) used only males and 16.8% (50/297) used only females. Reporting of sex
352 rose steadily from 2005 (39.0%, 30/77) to 2015 69.8% (74/106).

353

354

355 Regarding the chosen genetic background of animals used for preclinical studies (n= 108), 76%
356 (70/92) of those reporting this parameter generated experimental animals using a cross
357 between mice hemizygous for the SOD1 mutant gene and C57/SJL outbred strains.

358

359 Only ten studies (6 proof-of-concept studies and 4 preclinical studies) from 2007, 2009, 2011,
360 2013, and 2015 justified the number of animals used per group. However, of these, only six gave
361 clear justifications (five justified the group size by a power analysis and the other by the size of
362 groups proposed in ALS guidelines^{2 38}. On the other hand, group size was reported in 83.3%
363 (474/569) of ALS papers, and more so in the preclinical studies sub-sample (Figure 4).

364

365

366 **Figure 4. Group size.** Histogram of mean group size in 105 preclinical studies reporting this parameter
367 (left) and for each of the years analysed (yearly mean \pm 1 standard deviation) (right).

368

369

370 Of the 569 papers reviewed, 38% (214/569) did not report the method for killing animals despite
371 the fact that in 91% (195/214) of these, terminal procedures requiring anaesthesia for ethical
372 and practical reasons were identified (e.g. transcardial perfusion fixation). The most commonly
373 used euthanasia method – of the papers reporting this information – was anaesthetic overdose
374 or the use of another method under anaesthesia (86%; 317/367) while other methods such as
375 CO₂ asphyxiation (7%; 26/367) or others such as decapitation or cervical dislocation (7%; 24/367)
376 were seldom used. Very few studies (15 out of 569) were identified as not performing
377 euthanasia of any kind. The remaining 21 articles were deemed “inconclusive”, for neither
378 reporting euthanizing animals at any point nor reporting deaths.

379

380 **3.2 Regulatory compliance and estimated severity**

381 For public confidence in research, it is important that research with animals is carried out
382 according to standards set by legislation and in line with the principles of the 3Rs. We measured
383 such compliance through the Regulatory Compliance Reporting score and also looked at specific
384 research parameters individually.

385

386 3.2.1. Regulatory Compliance Reporting Score (RCR)

387 The Regulatory Compliance Reporting (RCR) Score, which measures to what extent compliance
388 with legislation and approval of animal experiments are reported in published papers, shows an
389 overall improvement in the reporting over the time period under study (joint Chi-square
390 $p < 0.001$, Figure 5). The estimated odds of RCR > 1 was 7.1 times higher in 2015 than 2005
391 ($p < 0.0001$). RCR did not differ between journals or between proof-of-concept and preclinical
392 studies but was affected by country (Figure 5). Studies with high severity seemed to have higher
393 odds of high RCR values ($p = 0.027$). Model diagnostics showed that logistic regression was
394 justified. Supplementary Table 3 shows the RCR model results.

395

396 **Figure 5. Predictive marginal means (predicted probabilities of values > 1) \pm 95% confidence interval of**
397 **publication year (left panel) and country (right panel) based on a model of a Regulatory Compliance**
398 **Reporting (RCR) Score in 490 ALS studies.** The probability of an RCR score above 1 was higher in 2013 and
399 2015 than 2005. China, France, Italy and South Korea appeared to have comparatively low probabilities,
400 while for example Spain, Belgium and Canada had somewhat high probabilities. No significant interactions
401 were found. The pseudo R-square statistic indicated that the model explained 16% of the total variation
402 in the data.

403

404 Over the entire period, most papers (67.0%; 381/569) reported that studies had been appraised
405 and approved by a third party (e.g. ethics committee, competent authority) with only 10.9%
406 (62/569) not reporting any kind of regulatory compliance. By 2015, all papers were found to
407 have some type of statement on regulatory compliance, most of which (83%) referring to prior
408 ethical approval of research protocols.

409

410 The correlation between MSR and RCR was weak, but highly significant (Spearman $\rho = 0.21$;
411 $p < 0.0001$) indicating that papers with high scores for methodological standards were somewhat
412 more likely to also score highly for regulatory standards.

413 3.2.2 Severity and refinement measures

414 We have found in previous systematic reviews ^{5 6 53} that self-reported compliance with
415 regulations may not necessarily affect the severity of the experiments being conducted. To test
416 whether actual experimental practice has changed over the study period, we classified the
417 severity of each study according to the criteria in Table 2. The majority of publications (60.7%)
418 (346/569) included experiments at level-4 severity (Figure 6-A). Of the 64 studies classified as
419 Level 5 (allowing animals to die of disease progression or to reach complete paralysis), 89%

420 reported regulatory compliance (70% ethical approval from a national authority or institutional
421 ethics committee and 19% compliance with relevant legislation or animal use guidelines).
422 However, between those studies that reported regulatory compliance and those that did not,
423 there was no difference in the proportion that were Level 5 (Chi-square (5 d.f.)=2.855, p=0.722)
424 (Figure 6-B).

425 On the other hand, we did observe a difference between preclinical and proof-of-concept
426 studies: preclinical studies included a higher proportion of studies within the highest severity
427 categories (77.9% (81/104) classified as level 4 and 19.2% (20/104) as level 5) than did proof-of-
428 concept studies (68.7% (265/386) classified as level 4 and 11.4% (44/386) as level 5). Moreover,
429 no preclinical studies were given a level 1 or level 2 severity (Chi-square (5 d.f.)=19.593, p=0.001)
430 (Figure 6-C).

431

432

433 **Figure 6. Severity classification of studies (N=569).** Figure 6-A illustrates the percentage of studies, by
434 year, classified into each of the 5-levels of our severity scale, as well as those of "undetermined" severity
435 due to insufficient information ($n = 77$ in 2005; $n=81$ in 2007; $n = 84$ in 2009; $n = 106$ in 2011; $n = 115$ in
436 2013; $n= 106$ in 2015). Figures 6-B and 6-C show percentage of studies classified into each of the 5-levels,
437 according to, respectively, reported regulatory compliance status ($n = 62$, not reported; $n = 126$, guidelines
438 followed; $n = 381$, protocol approval), and type of study ($n = 461$, proof-of-concept studies; $n = 108$,
439 preclinical studies).

440

441 Of studies classified between level-3 and level-5 severity (i.e. from which it could be ascertained
442 animals presented overt locomotor impairments), only 9.1% (42/456) described any refinement
443 measures to alleviate suffering (e.g. provision of mashed food and adaptation of bedding in later
444 stages of disease progression), which occurred almost exclusively (39/42) in Level 4 studies.

445 Differences in the regulatory landscape between countries imply that *how* animals are treated
446 in biomedical research may depend on *where* these experiments are carried out. The proportion
447 of high-severity (Level-5) studies differed significantly (Chi-square (13 d.f.)=35,561, p=0.001)
448 between the 14 most represented countries in our sample, ranging from 40% (8/20) and 41%
449 (7/17) – in South Korea and Israel, respectively – to 4% in Canada and China and even none in
450 Belgium (0/14) and the UK (0/23).

451

452 **4. Discussion**

453 Our analysis, the first of its kind to use specially devised scores encompassing both
454 methodological standards and regulatory compliance reporting (MSR and RCR, respectively)
455 over a 10-year period, suggest three main findings: The first is an overall improvement in both
456 regulatory compliance and methodological and reporting quality across the period assessed.
457 Also, and somewhat as expected, studies classified as 'preclinical' scored higher for
458 methodological and reporting quality as compared with more 'proof-of-concept' studies. The
459 third finding is that these scores varied widely according to the country in which the first author
460 was based, but not according to the journal publishing the paper.

461 The improved reporting of regulatory compliance, as expressed in the increase in RCR score
462 across time, is an indicator of widespread increase in reported adherence to animal welfare
463 regulatory requirements. However, this was not reflected in any significant change in the
464 proportion of highly severe (Level-5 in our classification scheme) studies or the reporting of
465 refinement measures (in studies where animals showed overt clinical signs). This is in agreement
466 with results from previous systematic reviews of animal research on Huntington's disease
467 (papers published 1997-2009) ⁵ and tuberculosis (1997-2011) ⁶. Also, while 'preclinical' studies
468 were more likely to be classified in the higher severity categories, there was no relation between
469 the level of severity and whether papers reported approval of protocols or compliance with
470 regulations, the latter also reflecting previous findings ^{5 53}.

471

472 Only 11.2% of ALS studies were classified at the highest severity level (level 5, i.e. including
473 experiments with spontaneous death or euthanasia at a near-death stage, i.e. complete
474 paralysis), which is much lower than that found in research using mouse models of Huntington's
475 Disease (38%) ⁵ and Tuberculosis (66%) ⁶. Moreover, most endpoints applied in ALS studies
476 adhered to the same basic criterion for euthanizing animals, namely the point at which animals
477 are unable to resume their position if laid recumbent within 10-30 seconds. This is the primary
478 endpoint proposed in existing guidelines for preclinical ALS ^{2 38} and the ALS Treatment
479 Development Institute's recommendations ²⁸ (level-4 severity on our scale), suggesting
480 researchers to a great extent act in accordance with published guidance published guidance in
481 this respect. However, this endpoint was already broadly used before the publication of the
482 guidelines suggesting that these reflect common practice at the time of publication.

483

484 Applying predefined endpoints is important to prevent the loss of biological samples from
485 animals found dead and for which time of death therefore cannot be defined⁵ hence maintaining

486 numbers of animals and avoiding loss of statistical power and subsequent inconclusive results.
487 However, from an animal welfare perspective, the current standard endpoint for ALS studies
488 corresponds to an end-stage where euthanasia may prevent deaths from respiratory failure, but
489 since they seldom anticipate death by more than a day, or even just a few hours, late stage
490 endpoints only curtail a small part of animal suffering ⁷. Very late endpoints increase the
491 likelihood that at least some animals will die unsupervised (e.g. overnight), while the
492 confounding effect of starvation and dehydration in survival data increases as animals become
493 progressively less able to reach the bottle spout or the food hopper ⁵⁴. At advanced clinical
494 stages, refinements such as providing mashed food on the cage floor, long-spouted water
495 bottles or fluid administration are therefore crucial to avoid unnecessary animal suffering and
496 to improve validity by bringing the model closer to the clinical setting, where late-stage human
497 patients are provided palliative care ⁵⁵. Defining endpoints also needs to take the research
498 purpose into account. In ALS, the mechanisms operating at different stages of the disease are
499 known to be different, principally affecting distal axons at the onset of symptoms, but
500 developing an immune/inflammatory phenotype during the end stages ⁵⁶. Therefore, endpoints
501 relevant to the treatment strategies must be used, particularly when targeting
502 neuroinflammation.

503

504 Methodological standards reporting improved over the time period under study. Studies
505 classified as 'preclinical' reported methodology in more detail than those deemed 'proof-of-
506 concept', consistent with the view that a more rigorous design and execution should be
507 demanded for preclinical studies ⁵⁷. Nevertheless, the checklist provided in the 2010 edition of
508 the guidelines for ALS research sets high methodological standards for both types of studies ².
509 Throughout the period under study, the MSR scores remain below 50% of the maximum score,
510 showing that the overall level of reporting of methodological detail remain substantially below
511 the recommendations in the guidelines.

512

513 Only three parameters (genetic background, number of transgene copies, and group size) were
514 reported in more than half of the sample, whereas other relevant information, such as housing
515 conditions, randomisation of animals into treatment groups or blinding of researchers was
516 absent in well over two thirds of the papers analysed, in line with previous reviews of animal
517 research in the neurosciences ^{5 54 58}. Other biological and methodological parameters, such as
518 sex (only reported in the majority of papers in the "preclinical studies" sub-sample) and method
519 of choice for euthanizing animals were also largely under-reported. The method used for
520 euthanizing animals has both animal welfare implications and scientific relevance, as the

521 method affects biological and histological parameters differently, which can impact the *post*
522 *mortem* data collected^{59 60}. The increase in the proportion of articles in our sample reporting
523 sex of the animals is positive, as sex differences^{4 61-63} in the phenotype or response to
524 therapeutic drugs may influence results and be of clinical relevance. However, although ALS
525 guidelines propose the use of both male and female mice, little over half of the studies providing
526 this information reported doing so. Overall, making these and other details on animals and
527 protocol available is central to allowing an adequate interpretation of results and a critical
528 evaluation of their validity, as well as allowing study replication and proper integration of results
529 in systematic reviews and meta-analyses^{31 64}.

530

531 Sample size was generally well reported, but of those reporting this parameter, only a small
532 minority used the 24 per group recommended in the 2010 guidelines². Furthermore, only three
533 studies clearly justified group size, in agreement with previous reports that this is frequently
534 overlooked e.g.^{31 65}. Adequate sample size is paramount to ensure that animals, time and
535 resources are not wasted as a result of underpowering experiments by using too few animals⁶⁶
536⁶⁷. Noise reduction by genetic standardisation could also help reduce the number of animals
537 needed per study, as the reduced inter-individual variability of isogenic strains allows increasing
538 power without requiring more animals⁶⁸ and is indeed mentioned in the 2007 guidelines as a
539 way of reducing variability in drug testing³⁸. Mead and colleagues⁶⁹, for instance, have shown
540 great consistency of results by using SOD1G93A transgenic mice on an inbred C57BL/6 genetic
541 background, with the added advantage of presenting early indicators of disease progress,
542 allowing for faster and more humane drug screening. Only 11% of the preclinical studies
543 reviewed, however, used a fully inbred background. The use of a single well characterised model
544 for initial studies can be supported further by independent replication studies in a different
545 disease model.

546

547 Most articles did not report random assignment of animals to groups or blinded outcome
548 assessment. This reflects similar data from reviews on the methodological quality of preclinical
549 research on ALS^{28 58 70} and other fields^{31 33 71-73}. This lack of attention to measures to avoid noise
550 and biases in animal experiments is cause for concern, given their role in improving the reliability
551 of results, as well as the translational value of preclinical research^{16 24 33 67 71}. While it cannot be
552 excluded that in some cases blinding and randomisation were applied but not reported, one
553 might expect that researchers carrying out well thought-out and planned experiments would
554 state such measures, since this strengthens their results and conclusions. There is ample
555 evidence for many areas^{32 33 73-75} that published studies which do not report measures to

556 minimise bias (i.e. blinding, randomisation and allocation concealment) tend to present an
557 exaggerated estimate of the therapeutic effect of experimental drugs. This is particularly
558 relevant in the light of the ongoing discussion of why promising pre-clinical results of candidate
559 drugs for ALS have not translated into the clinic. Although the disappointing outcomes of clinical
560 trials apparently contradict the promising preclinical results that elicited them, they may actually
561 mirror the results obtained from adequately designed animal studies carried out to high
562 methodological standards^{28 76}.

563

564 Methodological standards reporting and regulatory compliance reporting scores were not
565 influenced by the journal in which the results were published. Other researchers who have
566 investigated the effect of journal on methodological standards and reporting quality have found
567 a statistically significant but very small effect of whether or not the journal had endorsed the
568 ARRIVE guidelines.^{77 78}.

569

570 In contrast to previous research, this study indicated a gradual improvement in the
571 methodological standards and regulatory compliance reporting scores over time. However, it is
572 difficult to say to what extent this is the result of field-specific guidelines, as there is an overall
573 increasing trend in these score. Our study, of course, is limited to the period and model under
574 study and some improvements may have occurred as a result of the informal discussion leading
575 up to the formal workshops and guidelines (and more recently the appearance of other
576 transgenic models means that the study does not cover the entire field of ALS research for later
577 years). Also, a surprisingly low number of papers (1/84 in 2009, 10/106 in 2011, 10/115 in 2013
578 and 14/106 in 2015) referred to the Ludolph *et al.* guidelines^{2 38}. Given the slow adoption of the
579 ARRIVE guidelines⁷⁹, it seems likely it may also take some time for the ALS guidelines to have a
580 detectable effect.

581 While reporting of relevant parameters such as blinding and randomisation was higher in our
582 'preclinical' subsample than what has been reported in other systematic reviews^{16 31 77 79-82},
583 results for the overall sample were generally comparable. Also, and similarly to what was found
584 in these systematic reviews, justification for sample size was rarely reported.

585

586 One way of addressing the problems with study quality could be for preclinical researchers to
587 adopt the standards of randomised controlled trials in humans⁸³⁻⁸⁶, including trial pre-
588 registration^{87 88}. Compliance with existing guidelines would seem a more readily achievable goal,
589 however other self-regulatory mechanisms may be warranted to improve compliance, such as

590 changes to the publishing requirements of biomedical journals⁸⁹⁻⁹¹ or more demanding
591 requirements by science funders, both of which are clearly on the horizon^{30,92}.

592

593 **5. Conclusion**

594 The ALS research community pioneered the development of field-specific guidelines, setting
595 science community-based standards for animal research methodology and reporting^{2,38}.

596 Whereas we found significant improvement over time, it is less clear to what extent this is linked
597 to the guidelines, which are rarely referred to. Animal research in the field of ALS does however
598 differ from comparable research in other reviewed fields in one aspect: the implementation of
599 predefined endpoints in studies of advanced disease stages. This practice is important both for
600 research quality and animal welfare and is indeed coherent with the field-specific guidelines.
601 We propose that future guidelines should address measures to raise standards in the design,
602 conduct and reporting of experiments as well as to reduce the impact on animal welfare, as part
603 of a concerted effort to make biomedical research using animals more ethically and socially
604 acceptable and effective.

605

606 **Acknowledgements**

607 We thank Gilly Griffin for her input on current practice regarding humane endpoints in Canada.

608

609 **Funding**

610 NHF was a recipient of a Post-Doctoral Research Fellowship from the Portuguese Foundation for
611 Science and Technology (FCT), grant reference SFRH/BPD/85978/2012. The research leading to
612 these results has received funding from the European Union Seventh Framework Programme
613 [FP7-HEALTH-2013-INNOVATION-1] under grant agreement n. ° 602616 [Project ANIMPACT].
614 Analysis and revision was supported by the project Norte-01-0145-FEDER-000008 - Porto
615 Neurosciences and Neurologic Disease Research Initiative at I3S, supported by Norte Portugal
616 Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership
617 Agreement, through the European Regional Development Fund (FEDER) and FEDER - Fundo
618 Europeu de Desenvolvimento Regional funds through the COMPETE 2020 - Operational
619 Programme for Competitiveness and Internationalisation (POCI), Portugal 2020, and by
620 Portuguese funds through FCT - Fundação para a Ciência e a Tecnologia/Ministério da Ciência,
621 Tecnologia e Ensino Superior in the framework of the project "Institute for Research and
622 Innovation in Health Sciences" (POCI-01-0145-FEDER-007274).

623

624 **Competing Interests**

625 We have read and understood BMJ policy on declaration of interests and declare that we have
626 no competing interests.

627

628 **Role of Authors and Contributors**

629

630 Original idea for this study: NF, AO

631 Conception and design of the work: NF, AO, AJF, AJG

632 Data collection: JF, NF

633 Data analysis and interpretation: NF, JF, JH, AJF, AJG, AO

634 Drafting the article: NF, JF

635 Critical revision of the article: AJG, AJF, JH, AO

636 Final approval of the version to be published: NF, JF, AJG, AJF, JH, AO

637

638 **Data access statement**

639 The dataset is available at the University of Porto data repository ([https://ckan-](https://ckan-rdm.up.pt/dataset/i3s-2019-001)
640 [rdm.up.pt/dataset/i3s-2019-001](https://ckan-rdm.up.pt/dataset/i3s-2019-001))

641

642

643 **References**

- 644 1. Miller RG, Mitchell JD, Lyon M, et al. Riluzole for amyotrophic lateral sclerosis (ALS)/motor
645 neuron disease (MND). *Amyotrophic lateral sclerosis and other motor neuron disorders*
646 : official publication of the World Federation of Neurology, Research Group on Motor
647 Neuron Diseases 2003;4(3):191-206. doi: 10.1002/14651858.CD001447 [published
648 Online First: 2003/09/18]
- 649 2. Ludolph AC, Bendotti C, Blaugrund E, et al. Guidelines for preclinical animal research in
650 ALS/MND: A consensus meeting. *Amyotroph Lateral Scler* 2010;11(1-2):38-45. doi:
651 10.3109/17482960903545334
- 652 3. Shibata N. Transgenic mouse model for familial amyotrophic lateral sclerosis with
653 superoxide dismutase-1 mutation. *Neuropathology* 2001;21(1):82-92. doi:
654 10.1046/j.1440-1789.2001.00361.x
- 655 4. Heiman-Patterson TD, Deitch JS, Blankenhorn EP, et al. Background and gender effects on
656 survival in the TgN(SOD1-G93A)1Gur mouse model of ALS. *Journal of the Neurological*
657 *Sciences* 2005;236(1-2):1-7. doi: <http://dx.doi.org/10.1016/j.jns.2005.02.006>
- 658 5. Franco NH, Olsson IAS. "How sick must your mouse be?"-An analysis of the use of animal
659 models in Huntington's disease research. *Alternatives to laboratory animals: ATLA*
660 2012;40(5):271-83.

- 661 6. Franco NH, Correia-Neves M, Olsson IAS. Animal Welfare in Studies on Murine Tuberculosis:
662 Assessing Progress over a 12-Year Period and the Need for Further Improvement. *PLoS*
663 *One* 2012;7(10):e47723. doi: 10.1371/journal.pone.0047723
- 664 7. Franco NH, Correia-Neves M, Olsson IAS. How “humane” is your endpoint?—Refining the
665 science-driven approach for termination of animal studies of chronic infection. *PLoS*
666 *pathogens* 2012;8(1):e1002399. doi: 10.1371/journal.ppat.1002399
- 667 8. Solomon JA, Tarnopolsky MA, Hamadeh MJ. One universal common endpoint in mouse
668 models of amyotrophic lateral sclerosis. *PLoS One* 2011;6(6):e20582. doi:
669 10.1371/journal.pone.0020582 [published Online First: 2011/06/21]
- 670 9. Morton DB. Humane endpoints in animal experimentation for biomedical research: ethical,
671 legal and practical aspects: London: Royal Society of Medicine Press, 1999:5-12.
- 672 10. Sawiak S, Wood N, Williams G, et al. Use of magnetic resonance imaging for anatomical
673 phenotyping of the R6/2 mouse model of Huntington's disease. *Neurobiology of*
674 *Disease* 2009;33(1):12-19. doi: 10.1016/j.nbd.2008.09.017
- 675 11. Hockly E, Woodman B, Mahal A, et al. Standardization and statistical approaches to
676 therapeutic trials in the R6/2 mouse. *Brain research bulletin* 2003;61(5):469-79.
- 677 12. Menalled L, Brunner D. Animal models of Huntington's disease for translation to the clinic:
678 Best practices. *Movement Disorders* 2014;29(11):1375-90. doi: 10.1002/mds.26006
- 679 13. van der Worp HB, Howells DW, Sena ES, et al. Can animal models of disease reliably inform
680 human studies? *PLoS Medicine* 2010;7(3):e1000245. doi:
681 10.1371/journal.pmed.1000245
- 682 14. Ioannidis JP. Why most published research findings are false. *PLoS Medicine*
683 2005;2(8):e124. doi: 10.1371/journal.pmed.0020124
- 684 15. Schnabel J. Neuroscience: standard model. *Nature News* 2008;454(7205):682-85.
- 685 16. Ioannidis JPA, Greenland S, Hlatky MA, et al. Increasing value and reducing waste in
686 research design, conduct, and analysis. *The Lancet* 2014;383(9912):166-75. doi:
687 10.1016/S0140-6736(13)62227-8
- 688 17. Festing MFW. Randomized Block Experimental Designs Can Increase the Power and
689 Reproducibility of Laboratory Animal Experiments. *ILAR Journal* 2014;55(3):472-76.
690 doi: 10.1093/ilar/ilu045
- 691 18. Garner JP. The Significance of Meaning: Why Do Over 90% of Behavioral Neuroscience
692 Results Fail to Translate to Humans, and What Can We Do to Fix It? *ILAR Journal*
693 2014;55(3):438-56. doi: 10.1093/ilar/ilu047
- 694 19. Lapchak PA. Scientific Rigor Recommendations for Optimizing the Clinical Applicability of
695 Translational Research. *Journal of neurology & neurophysiology* 2012;3 doi:
696 10.4172/2155-9562.1000e111
- 697 20. Steward O, Balice-Gordon R. Rigor or Mortis: Best Practices for Preclinical Research in
698 Neuroscience. *Neuron* 2014;84(3):572-81. doi:
699 <http://dx.doi.org/10.1016/j.neuron.2014.10.042>
- 700 21. Tsilidis KK, Panagiotou OA, Sena ES, et al. Evaluation of excess significance bias in animal
701 studies of neurological diseases. *PLoS biology* 2013;11(7):e1001609. doi:
702 10.1371/journal.pbio.1001609
- 703 22. Button KS, Ioannidis JPA, Mokrysz C, et al. Power failure: why small sample size undermines
704 the reliability of neuroscience. *Nat Rev Neurosci* 2013;14(5):365-76. doi:
705 10.1038/nrn3475
- 706 23. Vesterinen HM, Sena ES, French-Constant C, et al. Improving the translational hit of
707 experimental treatments in multiple sclerosis. *Multiple Sclerosis Journal*
708 2010;16(9):1044-55. doi: 10.1177/1352458510379612
- 709 24. Sena ES, van der Worp HB, Bath PMW, et al. Publication Bias in Reports of Animal Stroke
710 Studies Leads to Major Overstatement of Efficacy. *PLoS Biol* 2010;8(3):e1000344. doi:
711 10.1371/journal.pbio.1000344

- 712 25. Steward O, Popovich PG, Dietrich WD, et al. Replication and reproducibility in spinal cord
713 injury research. *Exp Neurol* 2012;233(2):597-605. doi:
714 <http://dx.doi.org/10.1016/j.expneurol.2011.06.017>
- 715 26. Shineman DW, Basi GS, Bizon JL, et al. Accelerating drug discovery for Alzheimer's disease:
716 best practices for preclinical animal studies. *Alzheimers Res Ther* 2011;3(5):28. doi:
717 10.1186/alzrt90
- 718 27. Kimmelman J, London AJ, Ravina B, et al. Launching invasive, first-in-human trials against
719 Parkinson's disease: Ethical considerations. *Movement Disorders* 2009;24(13):1893-
720 901. doi: 10.1002/mds.22712
- 721 28. Scott S, Kranz JE, Cole J, et al. Design, power, and interpretation of studies in the standard
722 murine model of ALS. *Amyotroph Lateral Scler* 2008;9(1):4-15. doi:
723 10.1080/17482960701856300 [published Online First: 2008/02/15]
- 724 29. Collins FS, Tabak LA. NIH plans to enhance reproducibility. *Nature* 2014;505(7485):612.
- 725 30. Cressey D. UK funders demand strong statistics for animal studies. *Nature*
726 2015;520(7547):271.
- 727 31. Kilkenny C, Parsons N, Kadyszewski E, et al. Survey of the Quality of Experimental Design,
728 Statistical Analysis and Reporting of Research Using Animals. *PLoS One*
729 2009;4(11):e7824. doi: 10.1371/journal.pone.0007824
- 730 32. Crossley NA, Sena E, Goehler J, et al. Empirical evidence of bias in the design of
731 experimental stroke studies: a metaepidemiologic approach. *Stroke; a journal of*
732 *cerebral circulation* 2008;39(3):929-34. doi: 10.1161/strokeaha.107.498725 [published
733 Online First: 2008/02/02]
- 734 33. Hirst JA, Howick J, Aronson JK, et al. The need for randomization in animal trials: an
735 overview of systematic reviews. *PloS one* 2014;9(6):e98856.
- 736 34. van der Worp HB, Sena ES, Donnan GA, et al. Hypothermia in animal models of acute
737 ischaemic stroke: a systematic review and meta-analysis. *Brain : a journal of neurology*
738 2007;130(Pt 12):3063-74. doi: 10.1093/brain/awm083 [published Online First:
739 2007/05/05]
- 740 35. von Roten FC. Public perceptions of animal experimentation across Europe. *Public*
741 *Understanding of Science* 2012 doi: 10.1177/0963662511428045
- 742 36. Lund TB, Mørkbak MR, Lassen J, et al. Painful dilemmas: A study of the way the public's
743 assessment of animal research balances costs to animals against human benefits.
744 *Public Understanding of Science* 2014;23(4):428-44. doi: 10.1177/0963662512451402
- 745 37. Rollin BE. *Science and Ethics*: Cambridge University Press 2006.
- 746 38. Ludolph AC, Bendotti C, Blaugrund E, et al. Guidelines for the preclinical in vivo evaluation
747 of pharmacological active drugs for ALS/MND: report on the 142nd ENMC
748 international workshop. *Amyotrophic lateral sclerosis : official publication of the World*
749 *Federation of Neurology Research Group on Motor Neuron Diseases* 2007;8(4):217-23.
750 doi: 10.1080/17482960701292837 [published Online First: 2007/07/27]
- 751 39. Kanning KC, Kaplan A, Henderson CE. Motor neuron diversity in development and disease.
752 *Annual review of neuroscience* 2010;33:409-40. doi:
753 10.1146/annurev.neuro.051508.135722 [published Online First: 2010/04/07]
- 754 40. Lever TE, Gorsek A, Cox KT, et al. An animal model of oral dysphagia in amyotrophic lateral
755 sclerosis. *Dysphagia* 2009;24(2):180-95. doi: 10.1007/s00455-008-9190-z [published
756 Online First: 2008/12/25]
- 757 41. Lever TE, Simon E, Cox KT, et al. A mouse model of pharyngeal dysphagia in amyotrophic
758 lateral sclerosis. *Dysphagia* 2010;25(2):112-26. doi: 10.1007/s00455-009-9232-1
759 [published Online First: 2009/06/06]
- 760 42. Boylan K, Yang C, Crook J, et al. Immunoreactivity of the phosphorylated axonal
761 neurofilament H subunit (pNF-H) in blood of ALS model rodents and ALS patients:
762 evaluation of blood pNF-H as a potential ALS biomarker. *Journal of neurochemistry*

- 763 2009;111(5):1182-91. doi: 10.1111/j.1471-4159.2009.06386.x [published Online First:
764 2009/09/22]
- 765 43. Kato S, Kato M, Abe Y, et al. Redox system expression in the motor neurons in amyotrophic
766 lateral sclerosis (ALS): immunohistochemical studies on sporadic ALS, superoxide
767 dismutase 1 (SOD1)-mutated familial ALS, and SOD1-mutated ALS animal models. *Acta*
768 *neuropathologica* 2005;110(2):101-12. doi: 10.1007/s00401-005-1019-3 [published
769 Online First: 2005/06/29]
- 770 44. Gurney ME, Pu HF, Chiu AY, et al. Motor-neuron degeneration in mice that express a
771 human Cu,Zn superoxide-dismutase mutation. *Science* 1994;264(5166):1772-75. doi:
772 10.1126/science.8209258
- 773 45. Bruijn LI, Becher MW, Lee MK, et al. ALS-linked SOD1 mutant G85R mediates damage to
774 astrocytes and promotes rapidly progressive disease with SOD1-containing inclusions.
775 *Neuron* 1997;18(2):327-38. [published Online First: 1997/02/01]
- 776 46. Marcuzzo S, Zucca I, Mastropietro A, et al. Hind limb muscle atrophy precedes cerebral
777 neuronal degeneration in G93A-SOD1 mouse model of amyotrophic lateral sclerosis: a
778 longitudinal MRI study. *Exp Neurol* 2011;231(1):30-7. doi:
779 10.1016/j.expneurol.2011.05.007 [published Online First: 2011/05/31]
- 780 47. Neymotin A, Calingasan NY, Wille E, et al. Neuroprotective effect of Nrf2/ARE activators,
781 CDDO ethylamide and CDDO trifluoroethylamide, in a mouse model of amyotrophic
782 lateral sclerosis. *Free Radical Biology and Medicine* 2011;51(1):88-96. doi:
783 10.1016/j.freeradbiomed.2011.03.027
- 784 48. Del Signore SJ, Amante DJ, Kim J, et al. Combined riluzole and sodium phenylbutyrate
785 therapy in transgenic amyotrophic lateral sclerosis mice. *Amyotrophic lateral sclerosis :*
786 *official publication of the World Federation of Neurology Research Group on Motor*
787 *Neuron Diseases* 2009;10(2):85-94. doi: 10.1080/17482960802226148 [published
788 Online First: 2008/07/12]
- 789 49. Tada S, Okuno T, Yasui T, et al. Deleterious effects of lymphocytes at the early stage of
790 neurodegeneration in an animal model of amyotrophic lateral sclerosis. *Journal of*
791 *neuroinflammation* 2011;8(1):19. doi: 10.1186/1742-2094-8-19 [published Online First:
792 2011/02/25]
- 793 50. Pregibon D. Goodness of link tests for generalized linear models. *Applied statistics* 1980:15-
794 14.
- 795 51. Breusch TS, Pagan AR. A simple test for heteroscedasticity and random coefficient
796 variation. *Econometrica: Journal of the Econometric Society* 1979:1287-94.
- 797 52. Ramsey JB. Tests for specification errors in classical linear least-squares regression analysis.
798 *Journal of the Royal Statistical Society Series B (Methodological)* 1969:350-71.
- 799 53. Franco NH, Olsson IAS. Is the ethical appraisal of protocols enough to ensure best practice
800 in animal research? *Alternatives to laboratory animals: ATLA* 2013;41(1):P5-7.
- 801 54. Olsson IAS, Hansen AK, Sandoe P. Animal welfare and the refinement of neuroscience
802 research methods - a case study of Huntington's disease models. *Lab Anim*
803 2008;42(3):277-83. doi: 10.1258/la.2008.007147
- 804 55. Lilley E, Hawkins P, Jennings M. A 'road map' toward ending severe suffering of animals
805 used in research and testing. *ATLA - Alternatives to Laboratory Animals*
806 2014;42(4):267-72. [published Online First: 2014/10/08]
- 807 56. Boill e S, Yamanaka K, Lobsiger CS, et al. Onset and progression in inherited ALS
808 determined by motor neurons and microglia. *Science* 2006;312(5778):1389-92. doi:
809 10.1126/science.1123511
- 810 57. Kimmelman J, Mogil JS, Dirnagl U. Distinguishing between Exploratory and Confirmatory
811 Preclinical Research Will Improve Translation. *PLoS biology* 2014;12(5):e1001863. doi:
812 10.1371/journal.pbio.1001863

- 813 58. Benatar M. Lost in translation: Treatment trials in the SOD1 mouse and in human ALS.
814 *Neurobiology of Disease* 2007;26(1):1-13. doi:
815 <http://dx.doi.org/10.1016/j.nbd.2006.12.015>
- 816 59. Reilly J, Blackshaw AW. Euthanasia of animals used for scientific purposes: ANZCCART
817 2001.
- 818 60. Artwohl J, Brown P, Corning B, et al. Report of the ACLAM Task Force on Rodent
819 Euthanasia. *Journal of the American Association for Laboratory Animal Science*
820 2006;45(1):98-105.
- 821 61. Bame M, Pentiak PA, Needleman R, et al. Effect of Sex on Lifespan, Disease Progression,
822 and the Response to Methionine Sulfoximine in the SOD1 G93A Mouse Model for ALS.
823 *Gender Medicine* 2012;9(6):524-35. doi:
824 <http://dx.doi.org/10.1016/j.genm.2012.10.014>
- 825 62. McCombe PA, Henderson RD. Effects of gender in amyotrophic lateral sclerosis. *Gender*
826 *Medicine* 2010;7(6):557-70. doi: <http://dx.doi.org/10.1016/j.genm.2010.11.010>
- 827 63. Alves CJ, de Santana LP, Santos AJDd, et al. Early motor and electrophysiological changes in
828 transgenic mouse model of amyotrophic lateral sclerosis and gender differences on
829 clinical outcome. *Brain Research* 2011;1394(0):90-104. doi:
830 <http://dx.doi.org/10.1016/j.brainres.2011.02.060>
- 831 64. Hooijmans CR, Leenaars M, Ritskes-Hoitinga M. A gold standard publication checklist to
832 improve the quality of animal studies, to fully integrate the Three Rs, and to make
833 systematic reviews more feasible. *Alternatives to laboratory animals : ATLA*
834 2010;38(2):167-82. [published Online First: 2010/05/29]
- 835 65. Banwell V, Sena ES, Macleod MR. Systematic review and stratified meta-analysis of the
836 efficacy of interleukin-1 receptor antagonist in animal models of stroke. *Journal of*
837 *stroke and cerebrovascular diseases : the official journal of National Stroke Association*
838 2009;18(4):269-76. doi: 10.1016/j.jstrokecerebrovasdis.2008.11.009 [published Online
839 First: 2009/06/30]
- 840 66. Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research
841 evidence. *Lancet* 2009;374(9683):86-9. doi: 10.1016/s0140-6736(09)60329-9
842 [published Online First: 2009/06/16]
- 843 67. Festing MF, Altman DG. Guidelines for the design and statistical analysis of experiments
844 using laboratory animals. *ILAR journal / National Research Council, Institute of*
845 *Laboratory Animal Resources* 2002;43(4):244-58. [published Online First: 2002/10/23]
- 846 68. Festing MFW. Warning: the use of heterogeneous mice may seriously damage your
847 research. *Neurobiology of Aging* 1999;20(2):237-44. doi:
848 [http://dx.doi.org/10.1016/S0197-4580\(99\)00040-8](http://dx.doi.org/10.1016/S0197-4580(99)00040-8)
- 849 69. Mead RJ, Bennett EJ, Kennerley AJ, et al. Optimised and Rapid Pre-clinical Screening in the
850 SOD1^{G93A} Transgenic Mouse Model of Amyotrophic Lateral Sclerosis
851 (ALS). *PLoS One* 2011;6(8):e23244. doi: 10.1371/journal.pone.0023244
- 852 70. Perrin S. Preclinical research: Make mouse studies work. *Nature* 2014;507:423-25. doi:
853 10.1038/507423a
- 854 71. Landis SC, Amara SG, Asadullah K, et al. A call for transparent reporting to optimize the
855 predictive value of preclinical research. *Nature* 2012;490(7419):187-91. doi:
856 10.1038/nature11556
- 857 72. Bara M, Joffe AR. The methodological quality of animal research in critical care: the public
858 face of science. *Annals of intensive care* 2014;4(1):1-9. doi: 10.1186/s13613-014-0026-
859 8
- 860 73. Howells DW, Macleod MR. Evidence-based Translational Medicine. *Stroke; a journal of*
861 *cerebral circulation* 2013;44(5):1466-71. doi: 10.1161/strokeaha.113.000469
- 862 74. Bebarta V, Luyten D, Heard K. Emergency medicine animal research: does use of
863 randomization and blinding affect the results? *Academic Emergency Medicine*
864 2003;10(6):684-87.

- 865 75. Macleod MR, van der Worp HB, Sena ES, et al. Evidence for the efficacy of NXY-059 in
866 experimental focal cerebral ischaemia is confounded by study quality. *Stroke; a journal*
867 *of cerebral circulation* 2008;39(10):2824-29.
- 868 76. Perrin S. Preclinical research: Make mouse studies work. *Nature* 2014;507(7493):423-25.
- 869 77. Avey MT, Moher D, Sullivan KJ, et al. The Devil Is in the Details: Incomplete Reporting in
870 Preclinical Animal Research. *PLoS One* 2016;11(11):e0166733. doi:
871 10.1371/journal.pone.0166733
- 872 78. Vogt L, Reichlin TS, Nathues C, et al. Authorization of Animal Experiments Is Based on
873 Confidence Rather than Evidence of Scientific Rigor. *PLoS biology*
874 2016;14(12):e2000598. doi: 10.1371/journal.pbio.2000598
- 875 79. Baker D, Lidster K, Sottomayor A, et al. Two years later: journals are not yet enforcing the
876 ARRIVE guidelines on reporting standards for pre-clinical animal studies. *PLoS biology*
877 2014;12(1):e1001756. doi: 10.1371/journal.pbio.1001756
- 878 80. Gulin JEN, Rocco DM, García-Bournissen F. Quality of Reporting and Adherence to ARRIVE
879 Guidelines in Animal Studies for Chagas Disease Preclinical Drug Research: A
880 Systematic Review. *PLOS Neglected Tropical Diseases* 2015;9(11):e0004194. doi:
881 10.1371/journal.pntd.0004194
- 882 81. Macleod MR, Lawson McLean A, Kyriakopoulou A, et al. Risk of Bias in Reports of In Vivo
883 Research: A Focus for Improvement. *PLoS biology* 2015;13(10):e1002273. doi:
884 10.1371/journal.pbio.1002273
- 885 82. Ting KH, Hill CL, Whittle SL. Quality of reporting of interventional animal studies in
886 rheumatology: a systematic review using the ARRIVE guidelines. *International Journal*
887 *of Rheumatic Diseases* 2015;18(5):488-94. doi: 10.1111/1756-185x.12699
- 888 83. Muhlhausler BS, Bloomfield FH, Gillman MW. Whole animal experiments should be more
889 like human randomized controlled trials. *PLoS biology* 2013;11(2):e1001481.
- 890 84. McGonigle P, Ruggeri B. Animal models of human disease: Challenges in enabling
891 translation. *Biochemical Pharmacology* 2014;87(1):162-71. doi:
892 <http://dx.doi.org/10.1016/j.bcp.2013.08.006>
- 893 85. Hackam DG. Translating animal research into clinical benefit. *BMJ: British Medical Journal*
894 2007;334(7586):163. doi: 10.1136/bmj.39104.362951.80
- 895 86. de Vries RBM, Wever KE, Avey MT, et al. The Usefulness of Systematic Reviews of Animal
896 Experiments for the Design of Preclinical and Clinical Studies. *ILAR Journal*
897 2014;55(3):427-37. doi: 10.1093/ilar/ilu043
- 898 87. Jansen of Lorkeers SJ, Doevendans PA, Chamuleau SAJ. All preclinical trials should be
899 registered in advance in an online registry. *European Journal of Clinical Investigation*
900 2014;44(9):891-92. doi: 10.1111/eci.12299
- 901 88. Dal-Ré R, Ioannidis JP, Bracken MB, et al. Making prospective registration of observational
902 research a reality. *Science translational medicine* 2014;6(224):224cm1-24cm1. doi:
903 10.1126/scitranslmed.3007513
- 904 89. Rollin BE. Animal Research, Animal Welfare, and the Three R's *The Journal of Philosophy,*
905 *Science & Law* 2010;10
- 906 90. Osborne NJ, Phillips BJ, Westwood K. Journal editorial policies as a driver for change -
907 animal welfare and the 3R. *New Paradigms In Laboratory Animal Science - Proceedings*
908 *of the Eleventh FELASA symposium and the 40th Scand-LAS Symposium. Helsinki,*
909 *2010:18-23.*
- 910 91. Martins AR, Franco NH. A Critical Look at Biomedical Journals' Policies on Animal Research
911 by Use of a Novel Tool: The EXEMPLAR Scale. *Animals* 2015;5(2):315-31.
- 912 92. Editorial. Checklists work to improve science. *Nature* 2018;556:273-74. doi:
913 10.1038/d41586-018-04590-7

914