PEER REVIEW HISTORY

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	***Malcolm Macleod*** ***University of Edinburgh***
REVIEW RETURNED	*16/09/2018*

GENERAL COMMENTS	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 datas 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

	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?

REVIEWER 1	Shai D Silberberg ***NIH***
REVIEW RETURNED	*09/08/2018*

GENERAL COMMENTS	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. 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 liner 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 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. 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

	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".

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

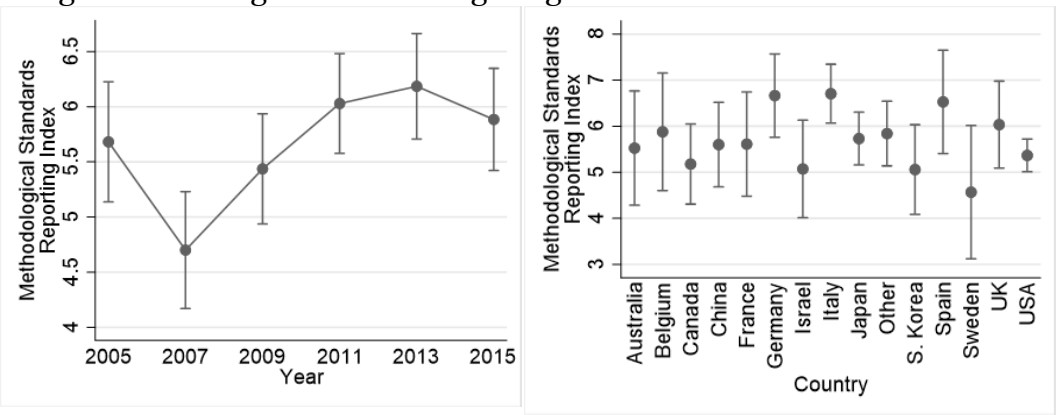
In response to the questions about the use of weighting.
We chose to create two aggregate measures (the MSR and RCR measures) in order to be able to use statistics that are more robust. We included in these aggregate measures a number of items which we consider relevant for determining the methodological 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.

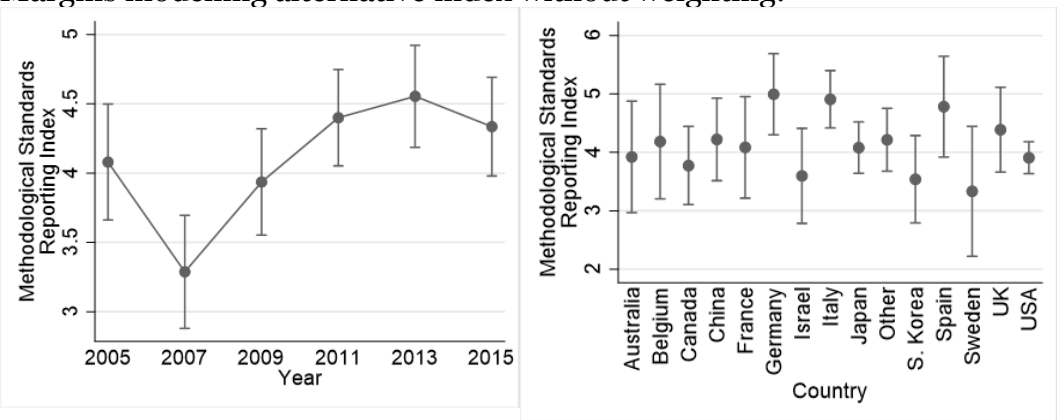
We have instead followed the recommendation by Reviewer 2, to test the consequences of weighting the MSR items, instead of using an unweighted sum, by modelling an alternative index formed without weighting. The plots below show that

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 datas 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 liner 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-hat=0.23 and p-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-hat=0.000 and p-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".

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

Joana G Fernandes[1,2¶], Nuno H Franco[1,2¶], Andrew J Grierson[3], Jan Hultgren[4], Andrew JW Furley[5&], I Anna S Olsson[1,2&*]

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

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

[3] Sheffield Institute for Translational Neuroscience, Department of Neuroscience, University of Sheffield, Sheffield, United Kingdom

[4] Department of Animal Environment and Health, Swedish University of Agricultural Sciences, Skara, Sweden

[5] Department of Biomedical Science, University of Sheffield, Western Bank, Sheffield, United Kingdom

*Corresponding author

E-mail: olsson@ibmc.up.pt

¶ These authors share the first authorship

[&] These authors share the last authorship

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

Word count: 5248

Abstract

Objectives

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

Methods

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

Results

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

Conclusions

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

Strengths and limitations:

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

1. Introduction

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

While the use of animal models is relevant for advancing knowledge and considered essential for testing putative treatments, it also presents ethical challenges and this may be a reason for public concern. As a result, a common legal requirement in many countries is that animal research projects undergo an evaluation process intended to ensure that protocols are designed and carried out in compliance with the 3Rs principle: *replacement* of animal use by non-animal methods, *reduction* of animal numbers needed to achieve the scientific objectives, and *refinement* of procedures to reduce or prevent harm to animals and improve their wellbeing. 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, but that significant progress could still be made to minimise and prevent avoidable suffering of laboratory animals. One key measure for accomplishing this is the termination of experiments during less severe stages of disease development where it is scientifically valid to do so. Endpoints based on early obtainable and scientifically sound indicators of phenotype progression can not only improve the ethical acceptability of animal studies, but also prevent the confounding influence of secondary factors; in the case of animal models of neurodegenerative diseases, starvation and dehydration arising from difficulties in eating and drinking due to progressive motor impairment can affect the phenotype and the readout of survival studies [8-10]. Simple refinements – such as adding mash food and longer bottle spouts – can however help reduce the influence of such factors [11-13].

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

limited translational value [14-19]. This also holds true for neuroscience [20-23], with concerns over the overall quality and reproducibility of published results being raised for several neuroscience sub-fields, including multiple sclerosis [24], stroke [25], spinal cord injury [26] Alzheimer's [27], Parkinson's [28], Huntington's [13] and ALS [29] research. This has led major science funders, including the National Institutes of Health [30] and Research Councils UK [31] to demand that future grant proposals attest to the likelihood of providing reliable results, by including details of experimental design and adequate justification of sample sizes. Reproducibility is further hindered by insufficient provision of information on methodology in published research [32] – including failure to account for key variables such as sex, genotype, age, and weight of animals, anaesthetics used or methods of euthanasia. Omitting information also makes it impossible to evaluate the study quality and there is evidence that papers that do not report randomization or blinding report exaggerate biological effects [33-35].

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

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

2. Methods

2.1 Database search

An advanced search was conducted on the *ISI Web of Science*® database with the query *TS = ((mice OR mouse) SAME (ALS OR "amyotrophic lateral sclerosis")).* Results were refined to include only original research articles written in English and published in 2005, 2007, 2009, 2011, 2013 and 2015. Years of publication were selected to include papers reporting research planned and carried out prior to and after the publication of guidelines for ALS research in 2007 [39] and 2010 [2], resulting from two international meetings held in 2006 and 2009, respectively (Figure 1).

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

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

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

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

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

2.2 Data collection

Each published study was categorised as either a 'preclinical' (i.e., carried out "to evaluate a drug for use in humans") or 'proof-of-concept' (i.e., aiming "to elucidate the mechanism of the disease"), according to the suggested classification for animal studies on ALS [2 39]. Table 1 describes the information retrieved regarding regulatory compliance, animal models, experimental design and animal welfare. This information was retrieved through careful reading of the full papers.

Table 1. Data retrieved. A description of the information collected from revised papers is presented for 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 *both*, 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".

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

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

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

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.

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

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

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

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

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

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

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

	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	*comply*	1	98	21.3	28	25.9
Project approval reported	*protocol*	1	315	68.3	66	61.1
Refinement measures (e.g. to aid feed and hydrate)	*refine*	1	29	6.3	14	13

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

Predictive marginal means were calculated for all significant model effects. Both models were checked using the Pregibon link test [53], and by examining standardised residuals. The MSR model was also checked with the Breusch-Pagan/Cook-Weisberg test for heteroscedasticity [54], the Ramsey regression specification-error test for omitted variables [55], and the RCR b model by examining delta-betas to identify influential observations. The proportion of the total variation in MSR and RCR b that could be explained by differences between countries or journals was determined by running empty mixed models with country and journal, respectively, as a random effect, and calculating the intra-class correlation coefficients.

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

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

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

3. Results

3.1. Quality of research and reporting

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

3.1.1 Methodological Standards Reporting index

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

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

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

3.1.2. Reporting of relevant research parameters

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

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

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

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

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

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

CO_2 asphyxiation (7%; 26/367) or others such as decapitation or cervical dislocation (7%; 24/367) were seldom used. Very few studies (15 out of 569) were identified as not performing euthanasia of any kind. The remaining 21 articles were deemed "inconclusive", for neither reporting euthanizing animals at any point nor reporting deaths.

3.2 Regulatory compliance and estimated severity

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

3.2.1. Regulatory compliance reporting index (RCR)

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

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

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

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

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

3.2.2 Severity and refinement measures

We have found in previous systematic reviews [6 56] that self-reported compliance with regulations may not necessarily affect the severity of the experiments being conducted. To test whether actual experimental practice has changed over the study period, we classified the severity of each study according to the criteria in Table 2. The majority of publications (60.7%) (346/569) included experiments at level-4 severity (Figure 6-A). Of the 64 studies classified as Level 5 (allowing animals to die of disease progression or to reach complete paralysis), 89% reported regulatory compliance (70% ethical approval from a national authority or institutional ethics committee and 19% compliance with relevant legislation or animal use guidelines). However, between those studies that reported regulatory compliance and those that did not, there was no difference in the proportion that were Level 5 ($x^2(5) = 2.855$, p = 0.722) (Figure 6-B).
On the other hand, we did observe a difference between preclinical and proof-of-concept studies: preclinical studies included a higher proportion of studies within the highest severity categories (77.9% (81/104) classified as level 4 and 19.2% (20/104) as level 5) than did proof-of-concept studies (68.7% (265/386) classified as level 4 and 11.4% (44/386) as level 5). Moreover, no preclinical studies were given a level 1 or level 2 severity ($x^2(5) = 19.593$, p = 0.001) (Figure 6-C).

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

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

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

4. Discussion

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

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

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

paralysis), which is much lower than that found in research using mouse models of Huntington's Disease (38%) [51] and Tuberculosis (66%) [56]). Moreover, most endpoints applied in ALS studies adhered to the same basic criterion for euthanizing animals, namely the point at which animals are unable to resume their position if laid recumbent within 10-30 seconds. This is the primary endpoint proposed in existing guidelines for preclinical ALS [2 39] and the ALS Treatment Development Institute's recommendations [29] (level-4 severity on our scale), suggesting widespread compliance to published guidance in this respect. However, this endpoint was already broadly used before the publication of the guidelines suggesting that these reflect common practice at the time of publication.

Applying predefined endpoints is important to prevent the loss of biological samples from animals found dead and for which time of death therefore cannot be defined [6], hence maintaining numbers of animals and avoiding loss of statistical power and subsequent inconclusive results. However, from an animal welfare perspective, the current standard endpoint for ALS studies corresponds to an end-stage where euthanasia may prevent deaths from respiratory failure, but since they seldom anticipate death by more than a day, or even just a few hours, late stage endpoints only curtail a small part of animal suffering [8]. Very late endpoints increase the likelihood that at least some animals will die unsupervised (e.g. overnight), while the confounding effect of starvation and dehydration in survival data increases as animals become progressively less able to reach the bottle spout or the food hopper [5]. At advanced clinical stages, refinements such as providing mashed food on the cage floor, long-spouted water bottles or fluid administration are therefore crucial to avoid unnecessary animal suffering and to improve validity by bringing the model closer to the clinical setting, where late-stage human patients are provided palliative care [58]. Defining endpoints also needs to take the research purpose into account. In ALS, the mechanisms operating at different stages of the disease are known to be different, principally affecting distal axons at the onset of symptoms, but developing an immune/inflammatory phenotype during the end stages [59]. Therefore, endpoints relevant to the treatment strategies must be used, particularly when targeting neuroinflammation.

Methodological standards reporting improved over the time period under study. Studies classified as 'preclinical' reported methodology in more detail than those deemed 'proof-of-concept', consistent with the view that a more rigorous design and execution should be demanded for preclinical studies [60]. Nevertheless, the checklist provided in the 2010 edition of the guidelines for ALS research sets high methodological standards for both types of studies [2].

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.

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

Sample size was generally well reported, but of those reporting this parameter, only a small minority used the 24 per group recommended in the 2010 guidelines [2]. Furthermore, only three studies clearly justified group size, in agreement with previous reports that this is frequently overlooked e.g. [32 68]. Adequate sample size is paramount to ensure that animals, time and resources are not wasted as a result of underpowering experiments by using too few animals [69 70]. Noise reduction by genetic standardisation could also help reduce the number of animals needed per study, as the reduced inter-individual variability of isogenic strains allows increasing power without requiring more animals [71] and is indeed mentioned in the 2007 guidelines as a way of reducing variability in drug testing [39]. Mead and colleagues [72], for instance, have shown great consistency of results by using SOD1G93A transgenic mice on an inbred C57BL/6 genetic background, with the added advantage of presenting early indicators of disease progress, allowing for faster and more humane drug screening. Only 11% of the preclinical studies reviewed, however, used a fully inbred background. The use of a single well characterised model

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

Most articles did not report random assignment of animals to groups or blinded outcome assessment. This reflects similar data from reviews on the methodological quality of preclinical research on ALS [29 61 73] and other fields [7 32 34 74 75]. This lack of attention to measures to avoid noise and biases in animal experiments is cause for concern, given their role in improving the reliability of results, as well as the translational value of preclinical research [17 25 34 70 74]. While it cannot be excluded that in some cases blinding and randomisation were applied but not reported, one might expect that researchers carrying out well thought-out and planned experiments would state such measures, since this strengthens their results and conclusions. There is ample evidence for ALS [29 76] and other areas [34 75 77-79] 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 is particularly relevant in the light of the ongoing discussion of why promising pre-clinical results of candidate drugs for ALS have not translated into the clinic. Although the disappointing outcomes of clinical trials apparently contradict the promising preclinical results that elicited them, they may actually mirror the results obtained from adequately designed animal studies carried out to high methodological standards [29 76].

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

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

al. guidelines[2 39]. Given the slow adoption of the ARRIVE guidelines [82], it seems likely it may also take some time for the ALS guidelines to have a detectable effect.

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

One way of addressing the problems with study quality could be for preclinical researchers to adopt the standards of randomised controlled trials in humans [86-89], including trial pre-registration [90 91]. Compliance with existing guidelines would seem a more readily achievable goal, however other self-regulatory mechanisms may be warranted to improve compliance, such as changes to the publishing requirements of biomedical journals [92-94] or more demanding requirements by science funders, both of which are clearly on the horizon [31 95].

5. Conclusion

The ALS research community pioneered the development of field-specific guidelines, setting science community-based standards for animal research methodology and reporting [2 39]. Whereas we found significant improvement over time, it is less clear to what extent this is linked to the guidelines, which are rarely referred to. Animal research in the field of ALS does however differ from comparable research in other reviewed fields in one aspect: the implementation of predefined endpoints in studies of advanced disease stages. This practice is important both for research quality and animal welfare and is indeed coherent with the field-specific guidelines. We propose that future guidelines should address measures to raise standards in the design, conduct and reporting of experiments as well as to reduce the impact on animal welfare, as part of a concerted effort to make biomedical research using animals more ethically and socially acceptable and effective.

Acknowledgements

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

Funding

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

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

Competing Interests

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

Role of Authors and Contributors

Original idea for this study: NF, AO

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

Data collection: JF, NF

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

Drafting the article: NF, JF

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

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

Data access statement

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

References

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

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

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

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

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

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

Joana G Fernandes[1,2¶], Nuno H Franco[1,2¶], Andrew J Grierson[3], Jan Hultgren[4], Andrew JW Furley[5&], I Anna S Olsson[1,2&*]

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

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

[3] Sheffield Institute for Translational Neuroscience, Department of Neuroscience, University of Sheffield, Sheffield, United Kingdom

[4] Department of Animal Environment and Health, Swedish University of Agricultural Sciences, Skara, Sweden

[5] Department of Biomedical Science, University of Sheffield, Western Bank, Sheffield, United Kingdom

*Corresponding author

E-mail: olsson@ibmc.up.pt

¶ These authors share the first authorship

[&] These authors share the last authorship

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

Abstract

Objectives

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

Methods

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

Results

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

Conclusions

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

Strengths and limitations:

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

1. Introduction

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

While the use of animal models is relevant for advancing knowledge and considered essential for testing putative treatments, it also presents ethical challenges and thus may be a reason for public concern. As a result, a common legal requirement in many countries is that animal research projects undergo an evaluation process intended to ensure that protocols are designed and carried out in compliance with the 3Rs principle: *replacement* of animal use by non-animal methods, *reduction* of animal numbers needed to achieve the scientific objectives, and *refinement* of procedures to reduce or prevent harm to animals and improve their wellbeing. Systematic reviews of animal use in both neuroscience[5] and infection[6] research indicate that self-reported regulatory compliance – including of ethical approval of protocols – has steadily increased over the last decade, but that significant progress could still be made to minimise and prevent avoidable suffering of laboratory animals. One key measure for accomplishing this is the termination of experiments during less severe stages of disease development where it is scientifically valid to do so. Endpoints based on early obtainable and scientifically sound indicators of phenotype progression can not only improve the ethical acceptability of animal studies, but also prevent the confounding influence of secondary factors; in the case of animal models of neurodegenerative diseases, starvation and dehydration arising from difficulties in eating and drinking due to progressive motor impairment can affect the phenotype and the readout of survival studies [7-9]. Simple refinements – such as adding mash food and longer bottle spouts – can however help reduce the influence of such factors [10-12].

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

limited translational value [13-18]. This also holds true for neuroscience [19-22], with concerns over the overall quality and reproducibility of published results being raised for several neuroscience sub-fields, including multiple sclerosis [23], stroke [24], spinal cord injury [25] Alzheimer's [26], Parkinson's [27], Huntington's [12] and ALS [28] research. This has led major science funders, including the National Institutes of Health [29] and Research Councils UK [30] to demand that future grant proposals attest to the likelihood of providing reliable results, by including details of experimental design and adequate justification of sample sizes. Reproducibility is further hindered by insufficient provision of information on methodology in published research [31] – including failure to account for key variables such as sex, genotype, age, and weight of animals, anaesthetics used or methods of euthanasia. Omitting information also makes it impossible to evaluate the study quality and there is evidence that papers that do not report randomization or blinding exaggerate biological effects [32-34].

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

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

2. Methods

2.1 Database search

An advanced search was conducted on the *ISI Web of Science*® database with the query *TS = ((mice OR mouse) SAME (ALS OR "amyotrophic lateral sclerosis")).* Results were refined to include only original research articles written in English and published in 2005, 2007, 2009, 2011, 2013 and 2015. Years of publication were selected to include papers reporting research planned and carried out prior to and after the publication of guidelines for ALS research in 2007 [38] and 2010 [2], resulting from two international meetings held in 2006 and 2009, respectively (Figure 1).

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

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

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

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

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

2.2 Data collection

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

Table 1. Data retrieved. A description of the information collected from revised papers is presented for 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 *both*, 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".

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

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

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

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.

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

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

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

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

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

Table 3. List of items integrated in the MSR and the RCR scores for preclinical (n=108) and proof-of-concept (n=461) animal studies on ALS reporting this information. The score for each variable is provided (MSR score ranging from 0 to 12.5, and RCR score ranging from 0 to 3). Greater weight (1.5 versus 1) was attributed to items which are also part of the ALS guidelines. The internal consistency reliability of each score was checked using Cronbach's alpha [50], estimating the degree to which the items measured a latent construct. For MSR, alpha was 0.58, which indicates poor internal consistency, which however does not disqualify the score. Omission of items *exclus*, *cagesize* and *blinded* increased alpha only slightly, thus the original MSR score was retained. For RCR, alpha was 0.30, but increased to 0.42 when item *severity* was omitted. For purposes of statistical modelling, RCR (only including items *comply*, *protocol* and *refine*) was later simplified to a binary variable RCRb coded as 1 for RCR values 2-3 and as 0 for RCR values 0-1.

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

	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	*comply*	1	98	21.3	28	25.9
Project approval reported	*protocol*	1	315	68.3	66	61.1
Refinement measures (e.g. to aid feed and hydrate) to aid feed and hydrate)	*refine*	1	29	6.3	14	13

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

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

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

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

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

3. Results

3.1. Quality of research and reporting

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

3.1.1 Methodological Standards Reporting Score

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

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

Figure 3. **Predictive marginal means ± 95% confidence interval of publication year (left panel) and country (right panel) based on a model of a Methodological Standards Reporting (MSR) Score in 487 ALS studies (predicted score values).** According to the linear regression model, MSR could be expected 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 0.002), respectively, and 0.98 units higher in 2005 (p=0.011). No significant interactions were found (e.g. between country and year). According to the R-square statistic the model explained 25% of the total variation in MSR.

3.1.2. Reporting of relevant research parameters

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

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

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

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

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

Of the 569 papers reviewed, 38% (214/569) did not report the method for killing animals despite the fact that in 91% (195/214) of these, terminal procedures requiring anaesthesia for ethical and practical reasons were identified (e.g. transcardial perfusion fixation). The most commonly used euthanasia method – of the papers reporting this information – was anaesthetic overdose or the use of another method under anaesthesia (86%; 317/367) while other methods such as CO_2 asphyxiation (7%; 26/367) or others such as decapitation or cervical dislocation (7%; 24/367) were seldom used. Very few studies (15 out of 569) were identified as not performing euthanasia of any kind. The remaining 21 articles were deemed "inconclusive", for neither reporting euthanizing animals at any point nor reporting deaths.

3.2 Regulatory compliance and estimated severity

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

3.2.1. Regulatory Compliance Reporting Score (RCR)

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

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

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

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

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

3.2.2 Severity and refinement measures

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

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

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

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

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

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

4. Discussion

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

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

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

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

Applying predefined endpoints is important to prevent the loss of biological samples from animals found dead and for which time of death therefore cannot be defined[5] hence maintaining numbers of animals and avoiding loss of statistical power and subsequent inconclusive results. However, from an animal welfare perspective, the current standard endpoint for ALS studies corresponds to an end-stage where euthanasia may prevent deaths from respiratory failure, but since they seldom anticipate death by more than a day, or even just a few hours, late stage endpoints only curtail a small part of animal suffering [7]. Very late endpoints increase the likelihood that at least some animals will die unsupervised (e.g. overnight), while the confounding effect of starvation and dehydration in survival data increases as animals become progressively less able to reach the bottle spout or the food hopper [55]. At advanced clinical stages, refinements such as providing mashed food on the cage floor, long-spouted water bottles or fluid administration are therefore crucial to avoid unnecessary animal suffering and to improve validity by bringing the model closer to the clinical setting, where late-stage human patients are provided palliative care [56]. Defining endpoints also needs to take the research purpose into account. In ALS, the mechanisms operating at different stages of the disease are known to be different, principally affecting distal axons at the onset of symptoms, but developing an immune/inflammatory phenotype during the end stages [57]. Therefore, endpoints relevant to the treatment strategies must be used, particularly when targeting neuroinflammation.

Methodological standards reporting improved over the time period under study. Studies classified as 'preclinical' reported methodology in more detail than those deemed 'proof-of-concept', consistent with the view that a more rigorous design and execution should be demanded for preclinical studies [58]. Nevertheless, the checklist provided in the 2010 edition of the guidelines for ALS research sets high methodological standards for both types of studies [2]. 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 remain substantially below the recommendations in the guidelines.

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

Sample size was generally well reported, but of those reporting this parameter, only a small minority used the 24 per group recommended in the 2010 guidelines [2]. Furthermore, only three studies clearly justified group size, in agreement with previous reports that this is frequently overlooked e.g. [31 66]. Adequate sample size is paramount to ensure that animals, time and resources are not wasted as a result of underpowering experiments by using too few animals [67 68]. Noise reduction by genetic standardisation could also help reduce the number of animals needed per study, as the reduced inter-individual variability of isogenic strains allows increasing

519 power without requiring more animals [69] and is indeed mentioned in the 2007 guidelines as a
520 way of reducing variability in drug testing [38]. Mead and colleagues [70], 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 guidelines [80], 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 humans [84-87], 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 journals [90-92] 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

Acknowledgements

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

Funding

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

Competing Interests

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

Role of Authors and Contributors

Original idea for this study: NF, AO

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

Data collection: JF, NF

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

Drafting the article: NF, JF

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

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

Data access statement

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

References

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

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

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

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

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

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

Joana G Fernandes[1,2¶], Nuno H Franco[1,2¶], Andrew J Grierson[3], Jan Hultgren[4], Andrew JW Furley[5&], I Anna S Olsson[1,2&*]

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

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

[3] Sheffield Institute for Translational Neuroscience, Department of Neuroscience, University of Sheffield, Sheffield, United Kingdom

[4] Department of Animal Environment and Health, Swedish University of Agricultural Sciences, Skara, Sweden

[5] Department of Biomedical Science, University of Sheffield, Western Bank, Sheffield, United Kingdom

*Corresponding author

E-mail: olsson@ibmc.up.pt

¶ These authors share the first authorship

[&] These authors share the last authorship

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

Abstract

Objectives

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

Methods

We searched for papers reporting ALS research on SOD1 mice published between 2005 and 2015 on the ISI Web of Science® database, resulting in a sample of 569 papers to review, after triage. Two scores – one for methodological quality, one for regulatory compliance –_ were built from

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

Results

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

Conclusions

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

Strengths and limitations:

- The approach for this systematic review is unique in covering methodological quality, regulatory compliance, and severity or animal welfare.
- We built two comprehensive scores (for methodological standards and for regulatory compliance) which were subjected to multivariable regression analysis to investigate how these scores were related to publication year, type of study, country of origin and journal, simultaneously accounting for all these factors.
- Our large sample (N=569 papers) included half the total population of published papers between 2005 and 2015.
- While more models of ALS are now available, only studies using the SOD-1 mouse were included.
- The protocol was defined prior to data collection but was not registered prior to the study.
- Information retrieval and assessment was not blinded.

1. Introduction

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

While the use of animal models is relevant for advancing knowledge and considered essential for testing putative treatments, it also presents ethical challenges and thus may be a reason for public concern. As a result, a common legal requirement in many countries is that animal research projects undergo an evaluation process intended to ensure that protocols are designed and carried out in compliance with the 3Rs principle: *replacement* of animal use by non-animal methods, *reduction* of animal numbers needed to achieve the scientific objectives, and *refinement* of procedures to reduce or prevent harm to animals and improve their wellbeing. Systematic reviews of animal use in both neuroscience[5] and infection[6] research indicate that self-reported regulatory compliance – including of ethical approval of protocols – has steadily increased over the last decade, but that significant progress could still be made to minimise and prevent avoidable suffering of laboratory animals. One key measure for accomplishing this is the termination of experiments during less severe stages of disease development where it is scientifically valid to do so. Endpoints based on early obtainable and scientifically sound indicators of phenotype progression can not only improve the ethical acceptability of animal studies, but also prevent the confounding influence of secondary factors; in the case of animal models of neurodegenerative diseases, starvation and dehydration arising from difficulties in eating and drinking due to progressive motor impairment can affect the phenotype and the readout of survival studies [7-9]. Simple refinements – such as adding mash food and longer bottle spouts – can however help reduce the influence of such factors [10-12].

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

limited translational value [13-18]. This also holds true for neuroscience [19-22], with concerns over the overall quality and reproducibility of published results being raised for several neuroscience sub-fields, including multiple sclerosis [23], stroke [24], spinal cord injury [25] Alzheimer's [26], Parkinson's [27], Huntington's [12] and ALS [28] research. This has led major science funders, including the National Institutes of Health [29] and Research Councils UK [30] to demand that future grant proposals attest to the likelihood of providing reliable results, by including details of experimental design and adequate justification of sample sizes. Reproducibility is further hindered by insufficient provision of information on methodology in published research [31] – including failure to account for key variables such as sex, genotype, age, and weight of animals, anaesthetics used or methods of euthanasia. Omitting information also makes it impossible to evaluate the study quality and there is evidence that papers that do not report randomization or blinding exaggerate biological effects [32-34].

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

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

2. Methods

2.1 Database search

An advanced search was conducted on the *ISI Web of Science*® database with the query *TS = ((mice OR mouse) SAME (ALS OR "amyotrophic lateral sclerosis")).* The database choice followed the protocol established for our previous reviews [5 6], based on considerations of access, search function and wide coverage of life sciences research. Results were refined to include only original research articles written in English and published in 2005, 2007, 2009, 2011, 2013 and 2015. Years of publication were selected to include papers reporting research planned and carried out prior to and after the publication of guidelines for ALS research in 2007 [38] and 2010 [2], resulting from two international meetings held in 2006 and 2009, respectively (Figure 1).

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

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

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

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

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

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

2.2 Data collection

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

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

Table 1. Data retrieved. A description of the information collected from revised papers is presented for 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 *both*, 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".

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

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

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

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

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.

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

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

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

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

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

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

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

	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	*comply*	1	98	21.3	28	25.9
Project approval reported	*protocol*	1	315	68.3	66	61.1
Refinement measures (e.g. to aid feed and hydrate) to aid feed and hydrate)	*refine*	1	29	6.3	14	13

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

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

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

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

3. Results

3.1. Quality of research and reporting

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

3.1.1 Methodological Standards Reporting Score

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

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

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

3.1.2. Reporting of relevant research parameters

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

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

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

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

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

Of the 569 papers reviewed, 38% (214/569) did not report the method for killing animals despite the fact that in 91% (195/214) of these, terminal procedures requiring anaesthesia for ethical and practical reasons were identified (e.g. transcardial perfusion fixation). The most commonly used euthanasia method – of the papers reporting this information – was anaesthetic overdose or the use of another method under anaesthesia (86%; 317/367) while other methods such as CO_2 asphyxiation (7%; 26/367) or others such as decapitation or cervical dislocation (7%; 24/367) were seldom used. Very few studies (15 out of 569) were identified as not performing euthanasia of any kind. The remaining 21 articles were deemed "inconclusive", for neither reporting euthanizing animals at any point nor reporting deaths.

3.2 Regulatory compliance and estimated severity

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

3.2.1. Regulatory Compliance Reporting Score (RCR)

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

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

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

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

3.2.2 Severity and refinement measures

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

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

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

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

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

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

4. Discussion

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

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

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

Applying predefined endpoints is important to prevent the loss of biological samples from animals found dead and for which time of death therefore cannot be defined[5] hence maintaining

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

Methodological standards reporting improved over the time period under study. Studies classified as 'preclinical' reported methodology in more detail than those deemed 'proof-of-concept', consistent with the view that a more rigorous design and execution should be demanded for preclinical studies [57]. Nevertheless, the checklist provided in the 2010 edition of the guidelines for ALS research sets high methodological standards for both types of studies [2]. 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 remain substantially below the recommendations in the guidelines.

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

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

Sample size was generally well reported, but of those reporting this parameter, only a small minority used the 24 per group recommended in the 2010 guidelines [2]. Furthermore, only three studies clearly justified group size, in agreement with previous reports that this is frequently overlooked e.g. [31 65]. Adequate sample size is paramount to ensure that animals, time and resources are not wasted as a result of underpowering experiments by using too few animals [66 67]. Noise reduction by genetic standardisation could also help reduce the number of animals needed per study, as the reduced inter-individual variability of isogenic strains allows increasing power without requiring more animals [68] and is indeed mentioned in the 2007 guidelines as a way of reducing variability in drug testing [38]. Mead and colleagues [69], for instance, have shown great consistency of results by using SOD1G93A transgenic mice on an inbred C57BL/6 genetic background, with the added advantage of presenting early indicators of disease progress, allowing for faster and more humane drug screening. Only 11% of the preclinical studies reviewed, however, used a fully inbred background. The use of a single well characterised model for initial studies can be supported further by independent replication studies in a different disease model.

Most articles did not report random assignment of animals to groups or blinded outcome assessment. This reflects similar data from reviews on the methodological quality of preclinical research on ALS [28 58 70] and other fields [31 33 71-73]. This lack of attention to measures to avoid noise and biases in animal experiments is cause for concern, given their role in improving the reliability of results, as well as the translational value of preclinical research [16 24 33 67 71]. While it cannot be excluded that in some cases blinding and randomisation were applied but not reported, one might expect that researchers carrying out well thought-out and planned experiments would state such measures, since this strengthens their results and conclusions. There is ample evidence for many areas [32 33 73-75] 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 is particularly relevant in the light of the ongoing discussion of why promising pre-clinical results of candidate drugs for ALS have not translated into the clinic. Although the disappointing outcomes of clinical trials apparently contradict the promising preclinical results that elicited them, they may actually mirror the results obtained from adequately designed animal studies carried out to high methodological standards [28 76].

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

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 score. Our study, of course, is limited to the period and model under study and some improvements may have occurred as a result of the informal discussion leading up to the formal workshops and guidelines (and more recently the appearance of other transgenic models means that the study does not cover the entire field of ALS research for later years). Also, a surprisingly low number of papers (1/84 in 2009, 10/106 in 2011, 10/115 in 2013 and 14/106 in 2015) referred to the Ludolph *et al*. guidelines[2 38]. Given the slow adoption of the ARRIVE guidelines [79], it seems likely it may also take some time for the ALS guidelines to have a detectable effect.

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

One way of addressing the problems with study quality could be for preclinical researchers to adopt the standards of randomised controlled trials in humans [83-86], including trial pre-registration [87 88]. Compliance with existing guidelines would seem a more readily achievable goal, however other self-regulatory mechanisms may be warranted to improve compliance, such as

changes to the publishing requirements of biomedical journals [89-91] or more demanding requirements by science funders, both of which are clearly on the horizon [30 92].

5. Conclusion

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

Acknowledgements

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

Funding

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

Competing Interests

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

Role of Authors and Contributors

Original idea for this study: NF, AO

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

Data collection: JF, NF

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

Drafting the article: NF, JF

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

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

Data access statement

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

References

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

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

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

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

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

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