PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Did a change in Nature journals' editorial policy for life sciences research improve reporting?
AUTHORS	**Professor Malcolm Macleod (Corresponding Author)** **The NPQIP Collaborative group**

VERSION 1 - REVIEW

REVIEWER 1	***Marcus Munafo*** ***University of Bristol***
REVIEW RETURNED	26-01-18

GENERAL COMMENTS	The authors report evidence that the introduction of checklists at NPG journals improved the reporting of various measures that protect against bias. This project is clearly a substantial undertaking, and the results potentially important, but I have some suggestions which may help improve the manuscript. First, it may be too late to change this, but I was surprised that the comparator comprised individual articles from (presumably) a wide range of other journal. Would it not have been more appropriate to choose comparator journals, since the primary unit of analysis is the journal (or even the publishing group)? This was the approach taken to evaluate the potential impact of the introduction of badges for data sharing at Psychological Science. The authors may wish to justify this choice in more detail, and perhaps discuss this as a potential limitation (e.g., more variance in the comparator group potentially masking similar change in some other journals). Second, the authors describe the results as reflecting major improvements, but that is a rather subjective interpretation. In my view some of the improvements are rather modest, and in many cases seems to be driven in large part by an increase in box checking rather than genuine engagement with the underlying issues (for example, there is a greater increase in the proportion of articles mentioning sample size with no formal sample size calculation than in those including a formal power calculation, which may actually make matters worse by encouraging phrases such as "sample size typical for the field"). Third, was there any evidence of differences across the different journals included? It's possible that unintended consequences (e.g., box ticking rather than genuine engagement) may have been more common in some disciplines than others. I appreciate this would be an exploratory analysis but it might be informative if

	it identifies areas where unintended consequences of the introduction of checklist are perhaps more widespread). I sign my reviews. Marcus Munafo

REVIEWER 2	***Anne Scheel*** ***Technische Universiteit Eindhoven***
REVIEW RETURNED	28-02-18

GENERAL COMMENTS	The authors report a study on the effects of an editorial intervention by Nature Journals regarding transparent reporting of research methods. The study has been planned with unusual rigour and transparency (as evidenced by the published preregistration and shared data) using a substantial sample size and an interesting new approach to crowd-source assessments of published articles while ensuring the quality of these assessments. In sum, I found the manuscript very interesting and relevant, but I think several aspects should be explained more plainly, put in a clearer structure, and the relation between the published preregistration of this study and the present manuscript should be made as transparent as possible. Below I list a number of major and minor comments, not necessarily in order of priority. 1. My first comment is probably the most basic and important one: After reading the manuscript and the published preregistration, I still fail to understand a) what the checklist that was implemented by NPG on 1st May 2013 looks like and b) how (if at all) the checklist used to measure the outcome of the present study differs from it. This could partly be due to the fact that I am not working in bio sciences or medicine, but since this study is of interest for meta researchers as well, I think that it should be written in a way to make it accessible to a broader audience. This confusion is not just a technical issue, but a conceptual one. Was this study designed to test simple compliance with new publication requirements (i.e., intervention and outcome are based on the same instrument) or to test if new publication requirements have an effect on other outcomes which are not part of these same guidelines (i.e., intervention and outcome are based on different instruments), or both (i.e., the instrument used for the outcome measure contains the intervention checklist in addition to other items)? All of these questions are potentially valuable, but the conclusions will be different depending on what is being asked. For example, if the checklist used by the authors is more or less identical to the NPG checklist, I would be inclined to judge post-intervention compliance rates of less than 20% as outrageously low, and I would say that the authors' description of "major improvements" in their Discussion section seems a little misplaced. If, however, what they measured went beyond the scope of the NPG checklist, I would agree that smaller changes should be considered a success. In the former case my conclusion would be "NPG publications do not adhere to NPG rules" (which should be very

	worrying for NPG), in the latter case it would be "NPG publication requirements have a positive, although not overwhelming, effect on transparent reporting". I think it would be important to make this issue as clear as possible very early on in the paper, and clearly reference which checklist items are used by NPG and which ones were used and added by the authors. As a slightly more minor point I found it difficult to understand how the authors categorised checklist items with regard to their research question. It seems clear that the "Landis 4" items were of primary interest, but the additional items and the way compliance with them was analysed (e.g. in conjunction vs. individually) and why should be described more clearly. Again, my struggles here may be due to the fact that my background is in a different research field, but I find it important to make the manuscript accessible to researchers outside of biomedical areas. 2. Why were NPG publications pre- and post-intervention matched for country? Of course it is possible that this variable has an influence on the quality of reporting, but the same is true for many other variables which were not taken into account (e.g., number of authors, COI statements, research area...). The problem I have with this decision is that publications for which no match on the country variable could be established for the pre- and post-intervention group were excluded. This introduces a bias in favour of publications from very prolific countries (for less prolific countries, finding matching pre- and post-intervention publications will be less likely). Whether or not this has an impact on the results of this study cannot be determined without looking at the full dataset without such exclusions. A second problem with this decision is that the same restriction was not made for the non-NPG sample (as a minor note, I found it somewhat confusing that the preregistered inclusion criteria for non-NPG publications do not mention the country variable, but the final manuscript implies that matching publications for country was attempted but found to not be possible). I suggest that the authors discuss this potential pitfall and the risk of collider bias and include a table or figure showing/comparing country distributions in all subgroups. 3. Both adherence to and deviation from the published preregistration should be made transparent. Instances of discrepancies which struck me (I may not have noticed all of them): 3.1 Coder recruitment: The preregistration states: "We will recruit individuals experienced in the critical appraisal of published materials (through for instance involvement with previous systematic reviews)," whereas the present manuscript says "We had no prior requirements for the skills required of these individuals". However, later the authors go on to say that coders were actually recruited from two different populations: "Each manuscript was scored by 2 individuals, one with experience in systematic review and risks of bias annotation and one recruited from outside this community." The recruitment procedure should be made clear and unambiguous, and

	adherence to/deviation from the preregistration must be made explicit. 3.2 Regarding the coding process, the preregistration states "Monitoring of outcome assessment after 10 % of manuscripts have been scored and adjudicated; we will review performance and if there are questions that are highly represented in those resulting in disagreements we will review the training materials and amend them as appropriate." I could not find any statement in the final manuscript about whether this actually took place, and if so, what the result was. 3.3 Primary and secondary outcomes: I found this part of the preregistration a little hard to comprehend, but if I understood correctly, it seems as if primary and secondary outcome were switched. Primary outcome according to the preregistration: "The proportion of publications in the intervention group describing in vivo research that meet the Landis criteria (item numbers #2, #3, #4, #5 of ‘‘Appendix 2’’)." Secondary outcome 1 (in vivo) according to the preregistration: "The change in prevalence of reporting of all of the Landis criteria (#2,3,4 and 5 together)." Primary outcome according to the present manuscript: "Our primary outcome was the change in the proportion of publications describing in vivo experiments published by NPG before and after May 2013 that meet all of the relevant Landis 4 criteria." Unfortunately I struggle to map the secondary outcomes reported in the final manuscript onto the secondary outcomes in the preregistration. This could well be due to my lack of familiarity with the research topic, but it would be good to have a clearer overview of this relation and any changes. I would recommend listing the outcomes/analysis in the same way as has been done in the preregistration. 3.4 The preregistration states "We will conduct sub-group analyses in groups defined by country of origin; categorisation of research; and whether the study is predominantly in silico; in vitro; in vivo; or involves human subjects," but I could not find these analyses in the final manuscript. 4. Coding and reviewer recruitment: 4.1 Was the "Gold standard" pool of 10 papers to train coders part of the studied sample? If it was, which of the many reviews for these papers were used for the study, and how was this determined? 4.2 Did any coders fail to reach sufficient concordance for three consecutive papers in the 10-paper pool? 4.3 How many papers had to be reconciled? 4.4 What is meant by "The agreement between the initial pair of outcome assessors ranged from..."? Were there other assessors beside the "initial pair" and the reconciler?

	5. Exploratory studies where not included in some or all of the analyses. It did not become entirely clear to me if this only referred to the Landis 4 items or to all outcomes. How many publications were excluded due to this criterion? Does this explain the varying group sizes for the individual outcome reports? Or were exploratory studies counted as complying with the checklist? 6. Power calculations: The power analyses reported in the preregistration are exceptionally detailed and well thought-out, which I see as a great strength of this study. However, the final manuscript only mentions achieved power and does not explicitly mention which sample size the calculation refers to, which is particularly relevant given that sample sizes vary between the different groups that are being studied. Here again I would appreciate a tighter and more transparent link between the preregistration and the final manuscript. 7. Another great strength of the preregistration is that the authors define the size of an "editorially significant change or prevalence" (Table 1). Explicitly setting a "smallest effect size of interest" (SESOI) like this is crucial to ensure hypotheses are falsifiable, and they greatly increase the practical value of investigations like this one (see e.g. Lakens, Scheel, & Isager, 2018). In this particular case I think that considering "editorial significance" is an excellent and helpful idea and it would be nice to spark a discussion about what should be considered an editorially significant change. However, as far as I could see, not all of the SESOIs the authors set for themselves were followed up with the appropriate tests in the final manuscript. That would mean to test both a) if changes/differences are greater than zero and b) if they are significantly smaller than the SESOI. Unless I missed something, it seems that this was only done for the very first analysis (change of proportion of NPG in vivo studies reaching full compliance with the Landis 4 criteria before vs. after the intervention). In my view it would be important to add these tests against the SESOI for each confirmatory analysis (i.e., each comparison that is described in the preregistration). NB: The authors do explain the level of power they had to detect changes of 15% or more (15% is defined as the SESOI for many of the analyses in the preregistration), but this is not sufficient to conclude the absence of differences as large as 15% when no significant result is obtained in a null-hypothesis test (see e.g. Lakens et al., 2018). On a minor note, as happy as I am about Table 1 in the preregistration, I was wondering why the values set in there do not seem to be discussed in the text. I think it would be nice to add this to the final manuscript (i.e. mention and discuss "editorially significant changes" and explain the reasoning behind the SESOIs of 80% and 15% which are mentioned in the preregistration). 8. The authors report picking papers for the non-NPG control group by matching non-NPG papers to the NPG groups based on

	publication date and whether they reported in vitro or in vivo research. However, the final result are uneven group sizes before vs. after 1st May 2013. The authors write: "The difference in numbers for NPG and non-NPG before and after 1st May 2013 is because some of the NPG “before” papers matched best with publications in other journals published in the few months following May 2013." I find this problematic, because the crucial comparison for all groups is before vs. after intervention. Prioritising a match on the in vivo/in vitro status criterion over the before vs. after criterion therefore makes little sense to me - it means that a "match" is being established between two publications that do not need to be matched, and a match between two studies that ought to match is sacrificed. Given that none of the resulting subgroups is particularly small, one may argue that any resulting problems are negligible - but note that e.g. the difference in group size between NPG in vivo before and non-NPG in vivo before is 20%, which is substantial in my view and lowers statistical power to detect changes in the non-NPG group. 9. Reporting of results: 9.1 First, all descriptive and test statistics are reported in the text, which I found very hard to read. Almost all of the reported numbers are also presented in tables 2-6, which in my view is a much better format for this. I would recommend to reduce redundancy between text and tables as much as possible by replacing most of the numbers in the text with references to the respective table, adding test statistics which are currently only reported in the text to the tables (Chi-square and df), and structuring the tables such that it becomes clear which of the preregistered analyses each result refers to. This would make the Results section much easier to comprehend, but it would also reduce additional sources of error. For example, the number of NPG in vitro studies after 1st May 2013 seems to be misreported as 176 in the text, although the (I assume) correct number 182 is apparent in Figure 1 and Table 3. On a related note, the abstract seems to contain a similar error: "The number of NPG publications meeting all relevant Landis 4 criteria increased from 0/203 prior to May 2013 to 31/181 (16.4%)" It seems to me that these should be 0/204, 31/190, and 16.3%, respectively. 9.2 The text in the Results section should be structured and labelled in the same way as the analyses are described in the Method section (e.g. by numbering them). 9.3 The size of the basic set of studies that compliant studies are compared against for the different checklist items varies wildly, and I found it hard to understand the reason for this. For example, for NPG in vivo studies there were 204 "before" and 190 "after" publications, but studies mentioning randomisation are reported as 14/169 before and 97/151 after. I assume that it is due to the fact that most items did not apply to all studies, and such studies were then excluded for the individual comparisons. In any case I believe it would be good to make this more clear to avoid confusion.

	9.4 Significant changes in "statistical reporting" are claimed but not backed up with any test statistics (admittedly this feels somewhat ironic); these should be added, preferably also in form of a table rather than in the text. 10. Discussion: Tying back in with my first point, I think in the Discussion section the authors could make it a bit more clear what the results mean in relation to which outcomes could have reasonably been expected. I also think it would be great to pick up the notion of "editorially significant changes" from the preregistration and discuss this concept. One final minor comment is that regarding the authors' suggestion that machine learning approaches could potentially decrease bias in assessments like the one undertaken here and/or make the process more efficient: To this I would add that any coding unreliability the authors observed in their study hints at problems regarding how researchers report their work - in other words, if two trained reviewers cannot agree whether or not I report if my study was blinded, maybe the problem lies with me rather than the reviewers. So working on the end of reporting standards might be as important or even more important than finding new ways of coding research reports. 11. A list of additional (very) minor comments: - Please provide references for the first sentence on page 4. - Similarly, the first sentence of the following sentence could be read to mean that in vivo research has been shown to be poorly replicable. If this is the case, please provide a reference, if not, if might be good to make the sentence sound less ambiguous ("Poor replication of in vitro molecular and cellular biology studies has also been reported" - "also" could be read to mean different things). - First line of the following paragraph: It would be good to add a comma at the beginning of the sentence to avoid confusion ("In May 2013, Nature Journals", otherwise it could be read to mean that 2013 journals did something) - p. 4, last paragraph, 3rd sentence: Something went wrong here, I assume "and before November 1st 2014" belongs to the following sentence - Why were papers reporting only research on humans not included? (Not a criticism, as a non-medical/biological researcher this is not obvious to me) - p. 5, 2nd paragraph, 3rd sentence: "the" should probably be "they" - same paragraph, last sentence: "paid" should probably be "played" - p. 7, 2nd paragraph, 2nd sentence: Please explain/define the word "reconciler" (I was initially struggling to understand this)

	- p. 7, last paragraph: "One NPG publication and one non-NPG publication were adjudged at the time of outcome assessment to report neither in vivo nor in vitro research and so were excluded." This sounds a bit cryptic. Does that simply mean the paper was erroneously judged to contain in vivo or in vitro research when it was included? - p. 9, 4th paragraph, last sentence: "The prevalence of reporting the different items before and after is shown in table 2" I think this should be Table 4, not 2. - p. 10, first line: According to my calculation, the 62% figure is 62.5%, which I assume should be rounded to 63%. But I might have made a mistake in my recalculation. - Figure 1: This figure is exceptionally helpful (no criticism). - Figure 2: Using red-green as the primary colour contrast in figures should be avoided since red-green blindness is the most common form of colour blindness. It would be nice if a more colour-blind-friendly palette could be used. - Figure 3: As mentioned in the paper, power for the interrupted time series analysis was substantially reduced due to the temporal distribution of the data, which in my view means that the results should be treated with caution. Therefore I would suggest to change Figure 3 to put more visual emphasis on the data points and less emphasis on the fitted lines. - Figure 4: Font size should be increased. - Tables: In the text, the comparison group is consistently labelled as "non-NPG" but in the tables it is referred to as "matched". I think text and tables should use the same label, in my view "non-NPG" is the better choice.

REVIEWER 3	***Michele Nuijten*** ***Tilburg University***
REVIEW RETURNED	19-02-18

GENERAL COMMENTS	This paper investigated whether a change in Nature's editorial policy improved reporting to decrease risk of bias. I think this is an interesting and timely research question, and I applaud the authors for taking an empirical approach to answer it. Very often policies seem to be implemented without any form of scientific evaluation. This is a great attempt to see whether implementing journal policies (a relatively low-cost intervention) can substantively improve scientific reporting. Even though I think this project has great potential, I see several important methodological problems that need to be addressed. First, even though the authors included a "control group" of comparable articles, they did not investigate whether these articles

	were also published in journals with checklists, nor did they include this control group in their main analysis as a statistical control. Second, several aspects of the manuscript need more thorough explanation. For instance, I missed a clear explanation of the checklist (preferably with examples) in the beginning of the paper. Furthermore, the description of the selection of the articles was not clear to me. I noticed that the descriptions in the protocol are much clearer (e.g., the bullets that describe the selection criteria), so perhaps parts of that paper can by copied to this manuscript. If that is not possible, then perhaps explicit references to specific parts of the protocol might help clarify things. I also think the manuscript needs some thorough editing to rephrase long sentences that are hard to follow, to correct mistakes in punctuation and spelling, and to improve the figures. Below, I will first list some strong points of this manuscript, then list my main concerns in more detail, and finally end with some minor remarks. Please note: page numbers correspond to the pages of the PDF file, not the page numbers of the manuscript. Signed, Michèle Nuijten Strong points • The study protocol and data analysis plan were preregistered and highly detailed, and the raw data are available. • Efforts were made to blind coders to the journal and period the articles were published. • To be able to code a large number of articles, the authors chose to invite people from all over the world to participate in this study. I think collaborations such as this one are the future of science, and should be encouraged. • The authors are careful in wording their conclusion and addressing limitations of the study, and I really like the suggestion for further research to investigate ways of automatically scanning articles to see if they comply with quality enhancing checklists. Remarks • In the abstract, the primary outcome is identified as a change in reporting standards over time in Nature. This is repeated later in the manuscript, but I'd say it is the relative change, compared to other journals. This also means that in the main analysis, it is not sufficient to only look at a change in reporting standards in Nature, and later look at the comparison articles. This should be done in a single analysis with an interaction term (see https://www.nature.com/articles/nn.2886;

	http://www.tandfonline.com/doi/abs/10.1198/000313006X152649) • It seems that the editorial policies of the journals in which the comparison articles were published were not examined. This seriously diminishes the value of this sample as a control group: it is imaginable that due to a changing scientific culture, journals included in the control group also changed their editorial policy, or already had such a policy in place. Moreover, it is not clear to me which journals are included in the control sample, but this seems important information. • I think it is shocking that before the improved editorial policies, the Landis 4 items were not met in a single case! And after the new policy, I'd still say that 16% is incredibly low. Especially since Nature "requires" completion of these checklists, I'm confused about this low compliance rate, and I think this deserves more attention in the manuscript, even if it's only the interpretation of the authors of how serious this problem is. • Quite some important details about the data collection are unclear from the manuscript. Some of them are reported in the protocol, but in the manuscript, the authors often tend to only quickly mention part of the used criteria, methodology, or other choices. Even though these things can be found in the protocol, I think the manuscript would benefit from a clearer explanation of the methodological choices & strategies. Some more explicit examples: o The authors report a power analysis, but judging from the description in the manuscript, the analysis seems to have been performed when the data were already collected (since it did not seem to have served as a justification for the sample size). This is not a problem per se, but it is misleading to then state it was run during the planning phase of the study. Furthermore, are the effect sizes the study can detect with 80% reasonable? I would like an explanation/interpretation of these numbers. o P.4: I would like to see some examples of items on this checklist here, or somewhere else in the beginning of the manuscript. It is still not entirely clear to me what this checklist entails. Also, why would there be a need to design a new series of questions to see if the criteria on the checklist are met? Why not just use the checklist itself? o P.5: why was the initial sample size set to 40 articles? Also, this sample size of 40 does not match sample sizes reported later in the paper (and in Figure 1). What does this initial subsample (?) mean? o P.5: country of origin of whom? o P.5: why are articles with human subjects not included? o P.5: what went wrong in including the articles? It seems

	that the inclusion criteria are very clear, so what type of problems occurred here? o P.6-7: some of the comparison articles did not fall into the "before" time period when they should have. This makes it unclear in which periods the comparison articles were published. This needs to be made explicit. o P.8: "The proportion of NPG in vivo studies reaching full compliance […] remained significantly lower than the target of 80%." In the manuscript, it is unclear whose target this is and where it comes from. It is also unclear if this is the number the power analysis was based on. • P.6: I can see why the Landis 4 items are less relevant for exploratory research than for confirmatory research (although this deserves a bit more attention in the manuscript, and preferably some relevant references), but I then do not understand why exploratory studies are included in the sample in the first place. If the goal is to see if editorial policies can increase reporting standards, but the reporting standards are not applicable to exploratory research, it seems that these studies should not have been examined here. Minor remarks • The title is quite long, maybe consider shortening it • I may have missed it, but I would like to see a codebook for the data published along side them. The variable names in the data set look pretty straightforward, but to ensure reproducibility, the details about the data need to be stored in the same place. If this code book is available, please refer to it explicitly when referring to the location of the raw data. • Past and present tenses are used inconsistently • In the abstract, the abbreviation "NPG" is not explained • P.4: Something went wrong with the punctuation in the following sentence: "The study populations comprised (1) Published articles accepted for publication in Nature journals, which described research in the life sciences and which were submitted after May 1st 2013, at which time the mandatory completion of a checklist at the stage of manuscript revision, was introduced. This checklist required authors to indicate where details relating to study design could be found in the manuscript at the point of manuscript revision. and before November 1st 2014; […]" • P.5: the paragraph about selecting relevant manuscripts contains a lot of long sentences with conditional inclusion

	decisions, and is hard to follow. At the end of this paragraph I also expected the final sample size. The final sample sizes are discussed in the Results section, which is justifiable, but I think it might fit better to immediately report these after the explanation of the inclusion criteria. • P.5: missing space in “(less than10%). There are quite a number of other punctuation issues and typos throughout the manuscript (missing spaces, double periods, “Boneferroni”, etc.), so I would encourage another round of careful revision to solve these. • P.5: “the same individual”, this is unclear. Several individuals have been discussed in the meantime. • P.6: the reference to Landis et al. misses a year • Figure 1: please use spaces or hyphens in the dates, without them they’re hard to read • P.8: the t-test misses degrees of freedom. • Figure 2: the red-green color scheme is hard to read for people who are color blind. Maybe change this to different patterns in black & white. Furthermore, this figure would be more informative if the compliance of the comparison articles were also depicted. • P. 8: Figure 2 is referred to as Figure 1 (b). Furthermore, the panels in the figure are not labelled a, b, c, etc. • The axes labels, ticks, and data points in Figure 3 is too small to read • Figure 4 (?; it had no label, the one with the spiderwebs) should also not make use of a red-green color scheme

VERSION 1 – AUTHOR RESPONSE

Re: **Did a change in Nature journals' editorial policy for life sciences research improve reporting?**

(previously “Findings of a retrospective, controlled cohort study of the impact of a change in Nature journals' editorial policy for life sciences research on the completeness of reporting study design and execution“)

The NPQIP Collaborative group

We are grateful to the reviewers for their comments, which we have addressed in a revised manuscript and below. In addition we have sought to improve our use of language. For instance, we had variously described “manuscripts”, “publications” and “articles” all to mean the same thing, and now we consistently use the term “articles”

Further, we are now clearer in our description of the primary outcome measure. When writing our data analysis plan, and prior to any data inspection or analysis, we swapped our “first” secondary outcome (the change in proportion of articles describing in vivo research meeting the 4 Landis criteria) for our original (in the published study protocol) primary outcome (whether compliance in the post intervention group of articles reached 80%). This was because our primary intention had been to observe any effect of a change in publication policy, and, with the benefit of hindsight, this was not captured in our original primary outcome. We have also discussed this as a limitation.

We apologise that, due to an oversight on our part, the Data Analysis Plan and Code deposited on the Open Science Framework were not made public at the time of submission. The link was given, but I suspect your reviewers were not able to see it. It has now been made public.

In the accompanying dataset 4 articles had had dates of publication as 01/01/2000. The date of publication used to identify comparator articles was not affected.

Finally, in checking the integrity of our data we found that two articles had been considered by both reviewers to include results from both in vitro and in vivo experiments; but that both had considered that every question relating in in vivo research was not relevant. The articles in question does not, on inspection, include reports of in vivo experiments. In the original analysis these article contributed, in error, to the denominator for the primary outcome measure in the NPG pre and post intervention groups. This has been corrected in this revision to 203 instead of 204 in the pre-intervention group and to 189 rather than 190 in the post intervention group, with a consequent (trivial) change in the % compliance (from 16.3% to 16.4%) and the 95% confidence intervals. One article in the pre-intervention comparator group had, similarly, been categorised at “both” when in fact it only reported in vivo research. These changes are explained in the manuscript. None of the assessments for individual items were changed, because the “not applicable” response had removed those articles from the denominator.

We have uploaded a revised version of the study dataset to Figshare, along with the data dictionary and the previous version of the dataset, along with an explanatory note setting out the revisions. Therefore it is possible for others to run the analysis both on the original dataset or on the revised dataset, if they so wish.

Reviewer 1:

The authors report evidence that the introduction of checklists at NPG journals improved the reporting of various measures that protect against bias. This project is clearly a substantial undertaking, and the results potentially important, but I have some suggestions which may help improve the manuscript. First, it may be too late to change this, but I was surprised that the comparator comprised individual articles from (presumably) a wide range of other journal. Would it not have been more appropriate to choose comparator journals, since the primary unit of analysis is the journal (or even the publishing group)? This was the approach taken to evaluate the potential impact of the introduction of badges for data sharing at Psychological Science. The authors may wish to justify this choice in more detail, and perhaps discuss this as a potential limitation (e.g., more variance in the comparator group potentially masking similar change in some other journals).

> For our comparator group we chose similar publications with a similar date of publication identified using the PubMed “related citations” tool. The journals in which these works were published will vary in the attention which they have given to transparency in reporting, and it may be that for some journals there have been changes similar to those observed in the

Nature Publishing Group publications. While we might have restricted our comparator group to journals more similar to NPG publications (for instance by Impact Factor, or extent of editorial intervention) this would have meant lower fidelity of matching by subject area or date of publication or both, and we considered these factors to be more important. For this reason, our findings for NPG publications cannot be interpreted as showing improved reporting compared with similar articles in similar journals. The representation of such “similar” journals in the comparator group is too small to allow meaningful conclusions to be drawn.

Second, the authors describe the results as reflecting major improvements, but that is a rather subjective interpretation. In my view some of the improvements are rather modest, and in many cases seems to be driven in large part by an increase in box checking rather than genuine engagement with the underlying issues (for example, there is a greater increase in the proportion of articles mentioning sample size with no formal sample size calculation than in those including a formal power calculation, which may actually make matters worse by encouraging phrases such as "sample size typical for the field").

We inserted the text

“While we saw improvements in the transparency of reporting, the observed improvements in experimental design were much more modest. However, peer review may not ensure the quality of published work {Smith, 2006 5381 /id}, as evidenced for in vivo research by poor reporting of measures to reduce risks of bias {Macleod, 2015 5508 /id}. We believe that the ultimate responsibility for assessing research quality (and therefore the validity of the findings presented) rests with the reader, and transparency in reporting is fundamental to this assessment.

Third, was there any evidence of differences across the different journals included? It's possible that unintended consequences (e.g., box ticking rather than genuine engagement) may have been more common in some disciplines than others. I appreciate this would be an exploratory analysis but it might be informative if it identifies areas where unintended consequences of the introduction of checklist are perhaps more widespread).

The proportions complying fully ranged from 0 (but Nature Structural and Molecular Biology only had 4 papers with in vivo experiments of which 0 were compliant) to 32% (7/22) at Nature Medicine. We think, given the small numbers, it might be misleading to report these.

Reviewer 2:

This paper investigated whether a change in Nature’s editorial policy improved reporting to decrease risk of bias. I think this is an interesting and timely research question, and I applaud the authors for taking an empirical approach to answer it. Very often policies seem to be implemented without any form of scientific evaluation. This is a great attempt to see whether implementing journal policies (a relatively low-cost intervention) can substantively improve scientific reporting. Even though I think this project has great potential, I see several important methodological problems that need to be addressed.

First, even though the authors included a “control group” of comparable articles, they did not investigate whether these articles were also published in journals with checklists, nor did they include this control group in their main analysis as a statistical control.

Most journals have made some efforts to improve the quality of reporting, whether through signalling (e.g. endorsement of the ARRIVE guidelines), training or guidance for peer reviewers, or formal adoption of for instance a reporting checklist. The primary question here is whether the change at NPG was associated with an improvement at NPG. The comparator group was included to establish whether any improvement at NPG was simply a reflection of a general improvement in similar articles across biomedical publishing. We expect that within

that group there will be some journals which have made substantial progress on a par with that seen at NPG, while others will have not. Given differences in the timing and characteristics of these diverse interventions we do not think that it is something we can measure.

Second, several aspects of the manuscript need more thorough explanation. For instance, I missed a clear explanation of the checklist (preferably with examples) in the beginning of the paper.

We inserted the text

"The development of this checklist was prompted in part by a consensus statement {Landis, 2012 5292 /id} setting out key aspects of study design and conduct which were necessary to allow the reader to assess the validity of the findings presented; it identified these as randomisation; blinding; sample size estimation and data handling (the "Landis 4"). The Nature Journals' checklist also included items relating to figures and statistical representation of data; reagents used; species, strain, and sex of experimental animals, reporting of relevant ethical approvals; consent (for research involving human subjects); data deposition; and availability of any bespoke computer code. The full checklist is given in Appendix 1.

Furthermore, the description of the selection of the articles was not clear to me. I noticed that the descriptions in the protocol are much clearer (e.g., the bullets that describe the selection criteria), so perhaps parts of that paper can by copied to this manuscript. If that is not possible, then perhaps explicit references to specific parts of the protocol might help clarify things.

We changed the text:

"Non- NPG publications: The same individual was responsible for identifying matching publications in other journals. Using PubMed, they entered the Nature Publishing Group publication title to retrieve the relevant record. Then they added the "related citations for PubMed" result to the search builder. In the second line search field of the search builder they searched for "Date of publication" in the same calendar month and year, and performed the search. In the results returned they started with the first result returned and established whether it was published in a Nature Publication Group Journal (given in Appendix 2). If it was not, they applied the study inclusion criteria (in vivo or in vitro research or both, as defined above), ensuring that there is a match on the in vivo/in vitro status between the index Nature Publishing Group publication and the non-Nature Publishing Group publication. Where these criteria were met they selected the publication for the study and retrieved the pdf, through open access, online institutional subscription, interlibrary loan, or by request from the authors. If the first related citation did not fulfil these criteria, they moved to the next, until an appropriate publication was found. If an appropriate publication was not found, they repeated these steps but with the date of publication used in the search extended by 1 month earlier and 1 month later. If this process did not identify an eligible publication, they again extended the search by a month in each direction, and continued until a matching publication was found. They then recorded the difference in calendar months between the date of publication of the index NPG article and the date of publication of the matching non-NPG article. Because of a limited number of potential matching publications it was not possible to match non NPG manuscripts by country.

The same individual used Adobe Acrobat to redact information relating to author names or affiliations, dates, volumes or page numbers; and the reference list; to minimise awareness of outcome assessors to whether the manuscript was pre- or post- intervention.

The individual making this selection and redacting information from publications paid no further part in the study.

I also think the manuscript needs some thorough editing to rephrase long sentences that are hard to follow, to correct mistakes in punctuation and spelling, and to improve the figures. Below, I will first list

some strong points of this manuscript, then list my main concerns in more detail, and finally end with some minor remarks.

> Apologies, we've tried to revise accordingly

Strong points • The study protocol and data analysis plan were preregistered and highly detailed, and the raw data are available. • Efforts were made to blind coders to the journal and period the articles were published. • To be able to code a large number of articles, the authors chose to invite people from all over the world to participate in this study. I think collaborations such as this one are the future of science, and should be encouraged. • The authors are careful in wording their conclusion and addressing limitations of the study, and I really like the suggestion for further research to investigate ways of automatically scanning articles to see if they comply with quality enhancing checklists.

Remarks • In the abstract, the primary outcome is identified as a change in reporting standards over time in Nature. This is repeated later in the manuscript, but I'd say it is the relative change, compared to other journals. This also means that in the main analysis, it is not sufficient to only look at a change in reporting standards in Nature, and later look at the comparison articles. This should be done in a single analysis with an interaction term (see https://www.nature.com/articles/nn.2886; http://www.tandfonline.com/doi/abs/10.1198/000313006X152649)

> Our intention was to determine whether there was any change in reporting standards over time in NPG papers. It was not our intention to seek evidence of relative change, and these data were only collected to provide context in which to interpret the findings for the NPG papers. Given prior publication of our study protocol and statistical analysis plan, and the concerns raised by reviewer 1 about the appropriateness of our selection of comparator articles, we are reluctant to embark on such a post hoc analysis, particularly as it is highly likely to deliver a "significant" result. However, since the data are in the public domain, it would be relatively straightforward for an interested used to conduct such a test.

It seems that the editorial policies of the journals in which the comparison articles were published were not examined. This seriously diminishes the value of this sample as a control group: it is imaginable that due to a changing scientific culture, journals included in the control group also changed their editorial policy, or already had such a policy in place.

> See above regarding our views of the control group (which we consider as NPG, pre-intervention) and the comparator group (articles in non-NPG publications).

Moreover, it is not clear to me which journals are included in the control sample, but this seems important information.

> The journals in which the NPG publications were published is now given in full.

I think it is shocking that before the improved editorial policies, the Landis 4 items were not met in a single case! And after the new policy, I'd still say that 16% is incredibly low. Especially since Nature "requires" completion of these checklists, I'm confused about this low compliance rate, and I think this deserves more attention in the manuscript, even if it's only the interpretation of the authors of how serious this problem is. •

> It is notable that even with considerable investment in designing and implementing a checklist, and working with authors to encourage its completion, that compliance remains so low. This stands rather in contrast to the belief that "all" that is required to ensure transparency in reporting is that journals "insist" that authors do the right thing. Securing transparency in research reports is a complex challenge, and experience in other fields (MM is also clinical lead for a clinical Neurology service) suggests such challenges require a range of complementary approaches with commitment from all stakeholders, might best be

achieved through formal improvement activity, and often take multiple attempts to achieve and sustain change.

Quite some important details about the data collection are unclear from the manuscript. Some of them are reported in the protocol, but in the manuscript, the authors often tend to only quickly mention part of the used criteria, methodology, or other choices. Even though these things can be found in the protocol, I think the manuscript would benefit from a clearer explanation of the methodological choices & strategies. Some more explicit examples:

Thank you, we've tried to provide greater detail and transparency

The authors report a power analysis, but judging from the description in the manuscript, the analysis seems to have been performed when the data were already collected (since it did not seem to have served as a justification for the sample size). This is not a problem per se, but it is misleading to then state it was run during the planning phase of the study.

We included the text:

"Power calculations were performed in STATA prior to commencement of the study. For the primary outcome measure we approximated required sample sizes using power calculations for a one sided two sample Chi squared test in STATA seeking a significance level of $p<0.01$ and with varying estimates of compliance with the Landis 4 criteria in the pre-intervention group. With 200 manuscripts in each group we had 80% power to detect an increase from 10% to 21%, or from 20% to 34%, or from 30% to 45%, or from 40% to 56%, or from 50% to 66%. We wanted to detect an absolute difference of 10% or more, and thought that compliance with the Landis 4 criteria in the pre-intervention group would be around 10%, so thought that having 200 studies in each group would be sufficient. For the primary outcome measure proposed in the original study protocol (that compliance with the Landis 4 criteria in the post-intervention group reached 80%), 200 studies in each group would be sufficient to reject the alternative hypothesis if the observed compliance was 72% or lower and again, we considered this to be sufficient. After correcting for multiple comparisons, and where the level of reporting in the pre intervention group was between 15% and 85%, with 200 studies per group we would have 80% power to detect an absolute increase of 15% in the reporting of each item. The power calculations are described in greater detail in the study protocol {Cramond, 2016 43 /id}.

Furthermore, are the effect sizes the study can detect with 80% reasonable? I would like an explanation/interpretation of these numbers.

At the stage of study design we discussed with the NPG editorial team what a meaningful improvement might be. We agreed that the improvements in the completeness of reporting of an item of less than 15% would not represent a very important change, but differences of more than this could be considered to represent progress towards better reporting.

P.4: I would like to see some examples of items on this checklist here, or somewhere else in the beginning of the manuscript. It is still not entirely clear to me what this checklist entails. Also, why would there be a need to design a new series of questions to see if the criteria on the checklist are met? Why not just use the checklist itself?

We now include more details of the checklist in the introduction; in the methods section on outcome assessment we have added a sentence explaining why we needed to "operationalise" the checklist items – essentially this is because a single checklist item could encompass more than one aspect of reporting, and we wanted to make it easier for outcome assessors.

P.5: why was the initial sample size set to 40 articles? Also, this sample size of 40 does not match sample sizes reported later in the paper (and in Figure 1). What does this initial subsample (?) mean?

We now include the text

"collecting papers with the intention of identifying 40 Nature papers and 20 each from other titles (i.e. 200 papers in total) ("Post intervention" group)."

P.5: country of origin of whom?

This is based on the address of the corresponding author**,** and is now described in the text

P.5: why are articles with human subjects not included?

Articles which only included human subjects, and did not include other in vivo or in vitro research, were excluded. Articles which included human subjects, and also included other in vivo or in vitro research, were included.

P.5: what went wrong in including the articles? It seems that the inclusion criteria are very clear, so what type of problems occurred here?

Interestingly, there was a small proportion of publications (4/896) which were initially included, but where it became apparent either during processing or outcome assessment that they did not meet the inclusion criteria. Usually this was because the experimental population was in fact neither in vitro nor in vivo. For instance, some studies were simulation studies based on previous in vivo observations, while others were further analyses of previously reported data. A further 13 studies had been included twice in error.

P.6-7: some of the comparison articles did not fall into the "before" time period when they should have. This makes it unclear in which periods the comparison articles were published. This needs to be made explicit.

This is now described in the text and the discussion in greater detail, along with the mismatch on in vitro: in vivo status

"The difference in numbers for NPG and non-NPG before and after 1st May 2013 is because some of the NPG "before" articles matched best with articles in other journals published in the few months following May 2013. Specifically, 26 NPG pre-intervention articles were matched with other papers published an average of 3.2 months after May 2013 (max 8 months), and 6 NPG post-intervention articles were matched with other papers published 1,2,9,11,12 and 215 months before May 2013. Overall, 43% of matched pairs had dates of publication within 1 month, 54% within 2 months, 64% within 3 months and 81% within 6 months of each other (range -11 to +22 months). 239 articles described only in vivo research, 133 described only in vitro research, and 507 described both. 494 papers were completely matched for in vivo and in vitro status, 276 were partially matched (one member of matched pair reporting in vivo and in vitro research, the other reporting only in vitro or only in vivo research), and 36 were mismatched (one reporting only in vivo research, the other reporting only in vitro research).

And in the discussion

"Our matching on whether studies reported in vitro or in vivo research, or both, was also reasonable in the majority of cases. Differences will have emerged where, as described above, articles were initially categorised with one set of characteristics (in vitro, in vivo or both) and matched accordingly, but later judged to have different characteristics. Our matching for date of publication worked reasonably well, with the exception of the inclusion of one comparator article published in 1995, 215 months before its "matching" NPG article. We had not anticipated that matching articles be so difficult to identify, so our matching rules did not have an upper limit of difference in date of publication. Since the comparator group do not

contribute to our primary outcome, and the matching is generally good, we do not think that these mismatches devalue our findings to any appreciable extent.

P.8: “The proportion of NPG in vivo studies reaching full compliance […] remained significantly lower than the target of 80%.” In the manuscript, it is unclear whose target this is and where it comes from. It is also unclear if this is the number the power analysis was based on.

We have clarified the discussion around the power calculations, and the reason for choosing 80% as a target

P.6: I can see why the Landis 4 items are less relevant for exploratory research than for confirmatory research (although this deserves a bit more attention in the manuscript, and preferably some relevant references), but I then do not understand why exploratory studies are included in the sample in the first place. If the goal is to see if editorial policies can increase reporting standards, but the reporting standards are not applicable to exploratory research, it seems that these studies should not have been examined here.

Indeed, and see Mogil and Macleod (PMID 28230138). However, it is often difficult to understand the intention of the investigator vis-à-vis exploratory versus hypothesis testing research, and given the preponderance of the application of hypothesis testing statistical approaches in studies where the prior articulation of an hypothesis has not been demonstrated, we do not believe that such a distinction would be possible in the context of this study. We operationalised this judgement, as we have described, according to whether authors presented statistical tests of hypotheses.

Minor remarks

• The title is quite long, maybe consider shortening it

We propose a change in title to

Did a change in Nature journals' editorial policy for life sciences research improve reporting?

• I may have missed it, but I would like to see a codebook for the data published along side them. The variable names in the data set look pretty straightforward, but to ensure reproducibility, the details about the data need to be stored in the same place. If this code book is available, please refer to it explicitly when referring to the location of the raw data.

The codebook is given on page 18 et seq of the Data Analysis Plan on the Open Science Framework, and we have now also placed it alongside the raw data on Figshare.

• Past and present tenses are used inconsistently

Thank you, we have tried to fix this

In the abstract, the abbreviation “NPG” is not explained

Thank you, fixed

• P.4: Something went wrong with the punctuation in the following sentence: “The study populations comprised (1) Published articles accepted for publication in Nature journals, which described research in the life sciences and which were submitted after May 1st 2013, at which time the mandatory completion of a checklist at the stage of manuscript revision, was introduced. This checklist required

authors to indicate where details relating to study design could be found in the manuscript at the point of manuscript revision. and before November 1st 2014; […]"

Thank you, fixed

• P.5: the paragraph about selecting relevant manuscripts contains a lot of long sentences with conditional inclusion decisions, and is hard to follow. At the end of this paragraph I also expected the final sample size. The final sample sizes are discussed in the Results section, which is justifiable, but I think it might fit better to immediately report these after the explanation of the inclusion criteria.

We have tried to improve the sentence, and have added

"In total 896 articles were selected for analysis."

• P.5: missing space in "(less than10%). There are quite a number of other punctuation issues and typos throughout the manuscript (missing spaces, double periods, "Boneferroni", etc.), so I would encourage another round of careful revision to solve these.

Thank you, we have tried to fix this

• P.5: "the same individual", this is unclear. Several individuals have been discussed in the meantime.

Thank you, fixed

• P.6: the reference to Landis et al. misses a year

Fixed, thank you

• Figure 1: please use spaces or hyphens in the dates, without them they're hard to read

Fixed, thank you

• P.8: the t-test misses degrees of freedom.

Apologies – the test actually reports a z-statistic not a t-statistic – this has now been fixed.

• Figure 2: the red-green color scheme is hard to read for people who are color blind. Maybe change this to different patterns in black & white. Furthermore, this figure would be more informative if the compliance of the comparison articles were also depicted.

Fixed, thank you. The lower pair of each set of panels is in fact the companion article performance, and we are sorry that this was not clearer; we have tried to fix this.

• P. 8: Figure 2 is referred to as Figure 1 (b). Furthermore, the panels in the figure are not labelled a, b, c, etc.

Fixed, thank you

• The axes labels, ticks, and data points in Figure 3 is too small to read

Fixed, thank you

• Figure 4 (?; it had no label, the one with the spiderwebs) should also not make use of a red-green color scheme

We have replaced with Blue/Red, which we understand is OK, and made the text bigger

VERSION 2 - Review

REVIEWER 2	***Anne Scheel*** ***Technische Universiteit Eindhoven***
REVIEW RETURNED	04-06-18

GENERAL COMMENTS	First, I should say that I think there was some kind of hiccup in the system, due to which the authors did not receive my review comments in the first round. It is possible that this had to do with me handing in quite late, for which I sincerely apologise. Unfortunately this means that many of my initial concerns are still valid after reading the revised manuscript, although it contains many substantial improvements and clarifications, for which I thank the authors. I am very sorry for putting the authors in the position of having to deal with such a large number of comments in the second round of reviews. My goal is to provide constructive feedback, and I therefore hope that my remarks can still be helpful. Substantial parts of the following text are identical with what I submitted in the first round of reviews; I apologise if this leads to any redundancy in the information the authors or the editor have/has received. In sum, I found the manuscript very interesting and relevant, and the revision offers many improvements. However, I think several aspects could still be explained more plainly, put in a clearer structure, and the relation between the published preregistration of this study and the present manuscript should be made as transparent as possible. Below I list a number of major and minor comments, not necessarily in order of priority. 1. Intervention vs. outcome After reading the initially submitted manuscript, the revised manuscript, and the published preregistration, I still fail to fully understand a) what the checklist that was implemented by NPG on 1st May 2013 looks like and b) how (if at all) the checklist used to measure the outcome of the present study differs from it. This

	could partly be due to the fact that I am not working in bio sciences or medicine, but since this study is of interest for meta researchers as well, I think that it should be written in a way to make it accessible to a broader audience. The revised manuscript refers to "Appendix 1" both for the NPG checklist and for the questions used for this study. Unfortunately I could not find an appendix in the current submission. The one in the previous submission contains one list which I presume represents the questions used in the study (i.e., not the NPG checklist), but I was not perfectly sure about it. This should be made much more transparent. Ideally, the checklist and the series of questions would be presented side-by-side. This confusion is not just a technical issue, but a conceptual one. Was this study designed to test simple compliance with new publication requirements (i.e., intervention and outcome are based on the same instrument) or to test if new publication requirements have an effect on other outcomes which are not part of these same guidelines (i.e., intervention and outcome are based on different instruments), or both (i.e., the instrument used for the outcome measure contains the intervention checklist in addition to other items)? All of these questions are potentially valuable, but the conclusions will be different depending on what is being asked. For example, if the checklist used by the authors is more or less identical to the NPG checklist, I would be inclined to judge post-intervention compliance rates of less than 20% as outrageously low. If, however, what they measured went beyond the scope of the NPG checklist, I would agree that smaller changes could be considered a success. In the former case my conclusion would be "NPG publications do not adhere to NPG rules" (which should be very worrying for NPG), in the latter case it would be "NPG publication requirements have a positive, although not overwhelming, effect on transparent reporting". As far as I understand, the checklist merely asks authors whether or not they report certain aspects of their study, and where they report it in their manuscript. The authors of the present study then coded if these aspects were reported in a given manuscript (but not the responses to the checklist during the submission/publication process). As an example, the following hypothetical scenarios are possible: a) The checklist asks "Do you report the randomisation procedure for assigning animals to experimental vs control group?" - Authors answer: "no"; coders check the manuscript and find: no, indeed

	they do not report this. b) Authors answer: "no"; coders find: contrary to their indication, the authors do report this in their manuscript. c) Authors answer: "yes"; coders find: yes, indeed the authors report this in their manuscript. d) Authors answer: "yes"; coders find: no, the authors actually do not report this in their manuscript. Or does the checklist require authors to state that they *have* reported all of these things? Whichever scenario is correct, I would encourage the authors to provide an example similar to the one I have attempted here in the beginning of the paper. I think it would make the purpose and scope of the study tremendously easier to understand (it took me several hours to get to this point). From what I understand, the answers to the NPG checklist are not compared to the coding outcomes - presumably the checklist answers were not available? This - i.e., the accuracy/honesty with which authors complete the checklist - might be another interesting question (I understand that it may be beyond the scope of the study, but pointing out the difference between that question and the actual focus of this study might make the manuscript much easier to comprehend). As a slightly more minor point I found it difficult to understand how the authors categorised checklist items with regard to their research question. It seems clear that the "Landis 4" items were of primary interest, but the additional items, the way compliance with them was analysed (e.g. in conjunction vs. individually), and why, should be described more clearly. Again, my struggles here may be due to the fact that my background is in a different research field, but I find it important to make the manuscript accessible to researchers outside of biomedical areas. Similarly, the authors explain that single checklist items were broken down into several questions for the coding procedure. I could not find a statement about how compliance with an item was established: would the article in question have to be coded as compliant with each question for an item or just a subset?

	2. Matching publications by country Why were pre- and post-intervention NPG publications matched for country? Of course it is possible that this variable has an influence on the quality of reporting, but the same is true for many other variables which were not taken into account (e.g., number of authors, COI statements, research area...). The problem I have with this decision is that publications for which no match on the country variable could be established for the pre- and post-intervention group were excluded. This introduces a bias in favour of publications from very prolific countries (for less prolific countries, finding matching pre- and post-intervention publications will be less likely). Whether or not this has an impact on the results of this study cannot be determined without looking at the full dataset without such exclusions. A second problem with this decision is that the same restriction was not made for the non-NPG sample (as a minor note, I found it somewhat confusing that the preregistered inclusion criteria for non-NPG publications do not mention the country variable, but the final manuscript implies that matching publications for country was attempted but found to not be possible). I suggest that the authors discuss this potential pitfall and the risk of collider bias and include a table or figure showing/comparing country distributions in all subgroups. 3. Preregistration Both adherence to and deviation from the published preregistration should be made transparent. Instances of discrepancies which struck me (I may not have noticed all of them):

	3.1 Coder recruitment: The preregistration states: "We will recruit individuals experienced in the critical appraisal of published materials (through for instance involvement with previous systematic reviews)," whereas the present manuscript says "We had no prior requirements for the skills required of these individuals". However, later the authors go on to say that coders were actually recruited from two different populations: "Each manuscript was scored by 2 individuals, one with experience in systematic review and risks of bias annotation and one recruited from outside this community." The recruitment procedure should be made clear and unambiguous, and adherence to/deviation from the preregistration must be made explicit. 3.2 Coding process: The preregistration states "Monitoring of outcome assessment after 10 % of manuscripts have been scored and adjudicated; we will review performance and if there are questions that are highly represented in those resulting in disagreements we will review the training materials and amend them as appropriate." I could not find any statement in the final manuscript about whether this actually took place, and if so, what the result was. 3.3 Primary and secondary outcomes: I thank the authors for clearing up the confusion about the switch of primary and secondary outcomes between the preregistration and the final paper. However, it is still tricky to comprehend the exact differences between preregistration and final analyses, especially for readers who have not read the preregistration. I strongly recommend listing the outcomes/analyses in the same way as has been done in the preregistration, ideally next to the preregistered ones (e.g. in a table). 3.4 The preregistration states "We will conduct sub-group analyses in groups defined by country of origin; categorisation of research;

	and whether the study is predominantly in silico; in vitro; in vivo; or involves human subjects," but I could not find these analyses in the final manuscript. 4. Coding and reviewer recruitment 4.1 Were the 10 papers in the "Gold standard" pool to train coders part of the studied sample? If it was, which of the many reviews for these papers were used for the study, and how was this determined? 4.2 Did any coder fail to reach sufficient concordance for three consecutive papers in the 10-paper pool? 4.3 How many papers had to be reconciled? 4.4 What is meant by "The agreement between the initial pair of outcome assessors ranged from..."? Were there other assessors beside the "initial pair" and the reconciler? 5. Exploratory studies Exploratory studies were not included in some or all of the analyses. It did not become entirely clear to me if this only referred to the Landis 4 items or to all outcomes. How many publications were excluded due to this criterion? Does this explain the varying group sizes for the individual outcome reports? Or were exploratory studies counted as complying with the checklist?

	6. Power calculations The power analyses reported in the preregistration are exceptionally detailed and well thought-out, which I see as a great strength of this study, and have been further improved in this revision by drawing a tighter link to the preregistration. 7. Editorially significant changes Another great strength of the preregistration is that the authors define the size of an "editorially significant change or prevalence" (Table 1). Explicitly setting a "smallest effect size of interest" (SESOI) like this is crucial to ensure hypotheses are falsifiable, and they greatly increase the practical value of investigations like this one. In this particular case I think that considering "editorial significance" is an excellent and helpful idea and it would be nice to spark a discussion about what should be considered an editorially significant change. However, as far as I could see, not all of the SESOIs the authors set for themselves were followed up with the appropriate tests in the final manuscript. That would mean to test both a) if changes/differences are greater than zero and b) if they are significantly smaller than the SESOI. Unless I missed something, it seems that this was only done for the very first analysis (change of proportion of NPG in vivo studies reaching full compliance with the Landis 4 criteria before vs. after the intervention). In my view it would be important to add these tests against the SESOI for each confirmatory analysis (i.e., each comparison that is described in the preregistration). NB: The authors do explain the level of power they had to detect changes of 15% or more (15% is defined as the SESOI for many of the analyses in the preregistration), but this is not sufficient to conclude the absence of differences as large as 15% when no significant result is obtained in a null-hypothesis test (see e.g. Lakens, Scheel, & Isager, 2018).

	On a minor note, as happy as I am about Table 1 in the preregistration, I was wondering why the values set in there do not seem to be discussed in the text. I think it would be nice to add this to the final manuscript (i.e. mention and discuss "editorially significant changes" and explain the reasoning behind the SESOIs of 80% and 15% which are mentioned in the preregistration). 8. Matching in vivo and in vitro articles The authors report picking papers for the non-NPG control group by matching non-NPG papers to the NPG groups based on publication date and whether they reported in vitro or in vivo research. However, the final result are uneven group sizes before vs. after 1st May 2013. The authors write: "The difference in numbers for NPG and non-NPG before and after 1st May 2013 is because some of the NPG “before” papers matched best with publications in other journals published in the few months following May 2013." I find this problematic, because the crucial comparison for all groups is before vs. after intervention. Prioritising a match on the in vivo/in vitro status criterion over the before vs. after criterion therefore makes little sense to me - it means that a "match" is being established between two publications that do not need to be matched, and a match between two studies that ought to match is sacrificed. Given that none of the resulting subgroups is particularly small, one may argue that any resulting problems are negligible - but note that e.g. the difference in group size between NPG in vivo before and non-NPG in vivo before is 20%, which is substantial in my view and lowers statistical power to detect changes in the non-NPG group. 9. Reporting of results 9.1 First, all descriptive and test statistics are reported in the text, which I find very hard to read. Almost all of the reported numbers are also presented in tables 2-6, which in my view is a much better format for this. I would recommend to reduce redundancy between

	text and tables as much as possible by replacing most of the numbers in the text with references to the respective table, adding test statistics which are currently only reported in the text to the tables (Chi-square and df), and structuring the tables such that it becomes clear which of the preregistered analyses each result refers to. This would make the Results section much easier to comprehend, but it would also reduce additional sources of error. For example, the number of NPG in vitro studies after 1st May 2013 seems to be misreported as 176 in the text, although the (I assume) correct number 182 is apparent in Figure 1 and Table 3. On a related note, the abstract seems to contain a similar error: "The number of NPG publications meeting all relevant Landis 4 criteria increased from 0/203 prior to May 2013 to 31/181 (16.4%)" It seems to me that these should be 0/204 and 31/189, respectively, and they sum to 393, not 394. 9.2 The text in the Results section should be structured and labelled in the same way as the analyses are described in the Method section (e.g. by numbering them). 9.3 The size of the basic set of studies that compliant studies are compared against for the different checklist items varies wildly, and I found it hard to understand the reason for this. For example, for NPG in vivo studies there were 203 "before" and 189 "after" publications, but studies mentioning randomisation are reported as 14/169 before and 97/151 after. I assume that it is due to the fact that most items did not apply to all studies, and such studies were then excluded for the individual comparisons. In any case I believe it would be good to make this clearer to avoid confusion. 9.4 Significant changes in "statistical reporting" are claimed but not backed up with any test statistics (admittedly this feels somewhat ironic); these should be added, preferably also in form of a table rather than in the text.

	10. Discussion Tying back in with my first point, I think in the Discussion section the authors could make it a bit more clear what the results mean in relation to which outcomes could have reasonably been expected. I also think it would be great to pick up the notion of "editorially significant changes" from the preregistration and discuss this concept. One final minor comment is that regarding the authors' suggestion that machine learning approaches could potentially decrease bias in assessments like the one undertaken here and/or make the process more efficient: To this I would add that any coding unreliability the authors observed in their study hints at problems regarding how researchers report their work - in other words, if two trained reviewers cannot agree whether or not I report if my study was blinded, maybe the problem lies with me rather than the reviewers. So working on the end of reporting standards might be as important or even more important than finding new ways of coding research reports. 11. A list of additional (very) minor comments: - The manuscript still contains a significant number of typos, in particular misplaced punctuation (missing periods, double periods, superfluous spaces, double parentheses, etc). - "Poor replication of in vitro molecular and cellular biology studies has also been reported" could be read to imply that *in vivo* research is unreplicable (because this sentence follows a discussion of in vivo research). The word "also" makes this sentence slightly ambiguous, maybe consider rephrasing.

	- First line of the following paragraph: It would be good to add a comma at the beginning of the sentence to avoid confusion ("In May 2013, Nature Journals", otherwise it could be read as "In May, 2013 Nature Journals ...") - p. 4, first line: "completion of a checklist at the stage of manuscript revision was introduced)" --> An earlier sentence says that the checklist was introduced at the stage of manuscript acceptance, not revision. Which one is correct? - p. 4: "A matching pre-intervention publication then identified, as described above." --> should probably be "was then identified", or similar - p. 6, last sentence: The second part is not a grammatically correct sentence, please rephrase. - p. 8, 2nd paragraph, 2nd sentence: Please explain/define the word "reconciler" (I was initially struggling to understand this) - p. 10, 2nd paragraph, last sentence: "The prevalence of reporting the different items before and after is shown in table 2" I think this should be Table 4, not 2. - p. 11, 4th paragraph, 2nd sentence: According to my calculation, the 62% figure is 62.5%, which I assume should be rounded to 63%. But I might have made a mistake in my recalculation.

	- Figure 1: This figure is exceptionally helpful (no criticism). - Figure 2: Thank you for changing the colour scheme, this is a great improvement. - Figure 3: As mentioned in the paper, power for the interrupted time series analysis was substantially reduced due to the temporal distribution of the data, which in my view means that the results should be treated with caution. Therefore I would suggest to change Figure 3 to put more visual emphasis on the data points and less emphasis on the fitted lines. - Figure 4: Font size should be increased. - Tables: In the text, the comparison group is consistently labelled as "non-NPG" but in the tables it is referred to as "matched". I think text and tables should use the same label, in my view "non-NPG" is the better choice.

REVIEWER 3	***Michele Nuijten*** ***Tilburg University***
REVIEW RETURNED	25-05-18

GENERAL COMMENTS	The authors have addressed most of my comments and those of the other two reviewers in great detail, which, in my view, has greatly improved the manuscript. The only concern I still have is the way the comparison group is treated. In my eyes, the comparison articles have every

	characteristic of a control group: they are matched to the NPG publications, and they are analyzed to “establish whether any improvement at NPG was simply a reflection of a general improvement in similar articles across biomedical publishing” (from the cover letter of the authors). This seems to me to be an exact description of a classic control group: it is not sufficient to simply monitor a change over time in NPG publications, you need to compare this change to a change in other journals without such guidelines. If the objective of the authors is to only monitor if the reporting standards improved in NPG publications, then they wouldn’t need a comparison sample. However, it would then also be much harder to justify that the change was caused by the change in journal policy. On the other hand, the authors did go to great length to create a comparison group. Why do that if this group is not included in the formal analyses? In short, I am confused why the comparison group is there if it is not to serve as a formal control. If the goal is to see whether editorial policies had an effect on reporting standards, I would say the best, feasible design here would be a quasi-experimental design with a non-equivalent control group and a pre- and post-test (i.e., treating the comparison group as a control). I would be interested to hear if the authors agree, and if not, why. Signed,

	Michèle Nuijten

VERSION 2 – AUTHOR RESPONSE

12th October 2018

Dear Chris,

Manuscript ID bmjos-2017-000035 - "Did a change in Nature journals' editorial policy for life sciences research improve reporting? "

Many thanks for the exceptionally helpful comments from yourself and your editors. Unfortunately the revisions have taken longer than expected, but I hope we have been able to address the issues raised.

Some of the reviewers asked for more detailed information, and we have provided this as best we can. I understand that BMJ OS would rather not have supplementary materials because of difficulties in attributing and resolving DOIs. Therefore we have simply placed this material in line; we think it all contributes to the understanding of the paper, and so we hope that this is OK.

To help guide the reviewers we have set out this response as follows …

The reviewers' comments are given in 12 point right justified Times New Roman

Our responses are given in 12 point left justified italicised Times new Roman

The changes to the text, where shown, are given in 12 point left justified italicised Arial

Yours Sincerely,

Malcolm Macleod

On behalf of the NPQIP collaborrators

Editor in Chief Comments:

Comments to Author

The manuscript is substantially closer to acceptance, but requires a further round of revision, as outlined below in the detailed guidance from the associate editor. Please accept our apologies for the technical difficulties in handling your manuscript.

Associate Editor Comments:

Comments to Author

At the outset, please accept my apologies for the difficulties we have been experiencing with the manuscript handling system. Upon preparing the decision letter for your revised manuscript, we noticed that one of the original reviews (for the previous version of the manuscript) had not been conveyed, through not fault of the reviewer. In addition, one of the re-reviews submitted this round has unfortunately been irretrievably lost on the system. The current decision letter is therefore based on the original review, which the reviewer kindly updated based on your revised manuscript, the one successfully submitted re-review, and comments relayed directly to me from the reviewer whose review was lost. As editors we regret the inconvenience and delay this has caused for you and the reviewers. Based on these reviews, the manuscript is substantially closer to acceptance, but will require one more round of substantive revision to address the comments of the reviewer whose assessment was misplaced on the first round, and the additional comments of the other successfully submitted re-review. As you will see, the original reviewer offers a wide range of constructive suggestions for improvement, especially emphasising the importance of making clear where deviations from the protocol were required. In addition, in direction discussion with the editor, the reviewer whose re-review was lost on the system wished to convey general satisfication with the revision, but that some additional caveats are warranted in the interpretation of results for the sample size checklist, as "compliance" (defined most liberally) could require no sample size calculations at all. In addition to this reviewer's comment, I noticed upon reading the revision a possible discrepancy in the reporting of compliance rates in the Discussion (para 3: "For reports of in vivo research, compliance for randomisation, blinding, reporting of exclusions and sample size calculations in NPG articles reached 68%, 62%, 31% and 64% respectively") and the compliance rates reported in the Results (p9). In particular, if the numbers in the results are correct, then shouldn't these percentages be reported in the Discussion as 64%, 55%, 31%, and 58%, respectively? If I have misunderstood what these values are referring to then this would benefit from clarification in the Discussion to avoid readers making the same mistake.

Thank you. We hope we have clarified our reporting; the figures you draw from the results section for randomisation, blinding and sample size calculations arise for "compliant by reporting they did the thing", but also "compliant by virtue of saying they didn't do the thing". For a study which was "yes" for part 1, part 2 was not relevant, and the study was therefore excluded from the denominator for this question. There was a little variability in the assessments offered, in that the numbers are somewhat contradictory (you can see this in table 4, where the numbers where eg blinding is considered not relevant seems to be different for each of these questions). Because we cannot know whether the assessors were correct on part 1 or part 2, we present the numbers in the table and use the denominator identified for each individual question. We have clarified the textual expression of these results and added this as a further limitation,

"We encountered a further unexpected problem when assessing compliance with reporting of blinding, randomisation and with sample size calculations. These were assessed with pairs of questions, first did the study report doing it (yes/no/not relevant); and second, did they at least mention it (yes/no/not relevant). If a study was "yes" for the first question, assessors were instructed to score the second as not relevant. Therefore, the number scored as "not relevant" for the second question should represent the sum of those scored as "yes" and as "not relevant" for the first. This

was not always the case (for in vivo research occurring in 0.1%, 0.8% and 6% of assessments for sample size calculation, randomisation and blinding respectively but we did not become aware of this problem until after database lock. Any impact of this shortcoming is likely to be small. "

Also, please ensure that Figure 2 includes x-axis labels.

Thank you, fixed

Reviewer(s)' Comments to Author:

Comments
Please leave your comments for the authors below

#1 Submitted by : Reviewer 1

The authors have addressed the comments I raised in the previous round. I still have a minor concerns that some of the measures may not capture some unintended and potentially negative consequences of checklists. For example, the largest increase for sample size calculation was any mention of sample size without a formal power calculation, which could include "N = 12 was good enough for my PhD supervisor, so it's good enough here". This slightly unrealistic example is transparent, of course, but would entrench bad practice whilst still allowing articles to be "compliant". In the absence of training in what constitutes *good* reporting, checklists might only take us so far. This might warrant some discussion, but I leave that to the authors' discretion.

Thank you. We have added a comment in discussion to emphasise this point (p12),

"For each of the Landis 4 criteria, compliance was most often achieved by the authors reporting that they had not taken measures to reduce the risk of bias. While this is not ideal, we believe this represents an improvement, in terms of the usefulness of the research to those who wish to use it, from a situation where these issues are not reported at all."

#2 Submitted by : Reviewer 2

First, I should say that I think there was some kind of hiccup in the system, due to which the authors did not receive my review comments in the first round. It is possible that this had to do with me handing in quite late, for which I sincerely apologise.

Thank you, but the responsibility lies elsewhere! We are keen to support the journal through its teething problems.

Unfortunately this means that many of my initial concerns are still valid after reading the revised manuscript, although it contains many substantial improvements and clarifications, for which I thank the authors.

I am very sorry for putting the authors in the position of having to deal with such a large number of comments in the second round of reviews. My goal is to provide constructive feedback, and I therefore hope that my remarks can still be helpful. Substantial parts of the following text are identical

with what I submitted in the first round of reviews; I apologise if this leads to any redundancy in the information the authors or the editor have/has received.

In sum, I found the manuscript very interesting and relevant, and the revision offers many improvements. However, I think several aspects could still be explained more plainly, put in a clearer structure, and the relation between the published preregistration of this study and the present manuscript should be made as transparent as possible.
Below I list a number of major and minor comments, not necessarily in order of priority.

Thank you for your very helpful comments – we found them very useful, and have tried to respond to them as best we can

1. Intervention vs. outcome

After reading the initially submitted manuscript, the revised manuscript, and the published preregistration, I still fail to fully understand a) what the checklist that was implemented by NPG on 1st May 2013 looks like and b) how (if at all) the checklist used to measure the outcome of the present study differs from it. This could partly be due to the fact that I am not working in bio sciences or medicine, but since this study is of interest for meta researchers as well, I think that it should be written in a way to make it accessible to a broader audience. The revised manuscript refers to "Appendix 1" both for the NPG checklist and for the questions used for this study. Unfortunately I could not find an appendix in the current submission. The one in the previous submission contains one list which I presume represents the questions used in the study (i.e., not the NPG checklist), but I was not perfectly sure about it. This should be made much more transparent. Ideally, the checklist and the series of questions would be presented side-by-side.

Our apologies. We now have Appendix 1 (the checklist), and appendix 2 (the questions used to assess checklist compliance). We have elected to preserve the checklist in the form presented to authors by NPG, as we think this aids comprehension. Because some of the checklist items apply to both in vitro and in vivo research, the "mapping" is not linear, and we hope we have avoided adding complexity without contributing substantially to understanding.

This confusion is not just a technical issue, but a conceptual one. Was this study designed to test simple compliance with new publication requirements (i.e., intervention and outcome are based on the same instrument) or to test if new publication requirements have an effect on other outcomes which are not part of these same guidelines (i.e., intervention and outcome are based on different instruments), or both (i.e., the instrument used for the outcome measure contains the intervention checklist in addition to other items)? All of these questions are potentially valuable, but the conclusions will be different depending on what is being asked. For example, if the checklist used by the authors is more or less identical to the NPG checklist, I would be inclined to judge post-intervention compliance rates of less than 20% as outrageously low. If, however, what they measured went beyond the scope of the NPG checklist, I would agree that smaller changes could be considered a success. In the former case my conclusion would be "NPG publications do not adhere to NPG rules" (which should be very worrying for NPG), in the latter case it would be "NPG publication requirements have a positive, although not overwhelming, effect on transparent reporting".

The assessment was designed to assess compliance with the NPG checklist, no more. We think this is now clearer. 16% is indeed low; but it is substantially better than other journals (on average) are achieving, and we believe it does represent progress towards a goal of improved reporting. Indeed, partly as a result of this study MRM is working with Nature, Science, Cell, PLoS and others to articulate a Minimum Standards Framework for biomedical research which – it is hoped – will lead to further improvements. This will be described shortly in a Blog post, and it the timing works I would propose to add a reference to this in the discussion. It's also much more challenging to achieve compliance across 4 measures than for a single measure, and we have also discussed this.

As far as I understand, the checklist merely asks authors whether or not they report certain aspects of their study, and where they report it in their manuscript. The authors of the present study then coded if

these aspects were reported in a given manuscript (but not the responses to the checklist during the submission/publication process). As an example, the following hypothetical scenarios are possible:
a) The checklist asks "Do you report the randomisation procedure for assigning animals to experimental vs control group?" - Authors answer: "no"; coders check the manuscript and find: no, indeed they do not report this.
b) Authors answer: "no"; coders find: contrary to their indication, the authors do report this in their manuscript.
c) Authors answer: "yes"; coders find: yes, indeed the authors report this in their manuscript.
d) Authors answer: "yes"; coders find: no, the authors actually do not report this in their manuscript.

Or does the checklist require authors to state that they *have* reported all of these things? Whichever scenario is correct, I would encourage the authors to provide an example similar to the one I have attempted here in the beginning of the paper. I think it would make the purpose and scope of the study tremendously easier to understand (it took me several hours to get to this point). From what I understand, the answers to the NPG checklist are not compared to the coding outcomes - presumably the checklist answers were not available? This - i.e., the accuracy/honesty with which authors complete the checklist - might be another interesting question (I understand that it may be beyond the scope of the study, but pointing out the difference between that question and the actual focus of this study might make the manuscript much easier to comprehend).

Indeed – we did not have access to the checklists completed by the authors. These were not considered by NPG to be public documents, but I understand that, going forward, they will be, allowing the interesting research you propose to be conducted. We have added the sentence

"To do this we assessed whether – in the view of trained assessors – manuscripts reported the details required by the checklist. Importantly, we did not have access to the checklists completed by the authors."

As a slightly more minor point I found it difficult to understand how the authors categorised checklist items with regard to their research question. It seems clear that the "Landis 4" items were of primary interest, but the additional items, the way compliance with them was analysed (e.g. in conjunction vs. individually), and why, should be described more clearly. Again, my struggles here may be due to the fact that my background is in a different research field, but I find it important to make the manuscript accessible to researchers outside of biomedical areas.

We have tried to make this clearer

Similarly, the authors explain that single checklist items were broken down into several questions for the coding procedure. I could not find a statement about how compliance with an item was established: would the article in question have to be coded as compliant with each question for an item or just a subset?

We have tried to articulate more clearly how compliance could be achieved: For some items, this could be by asserting that (eg blinding) had been done, or stating that it had not been done. Where a number of in vivo experiments were reported, we required that all be compliant for the manuscript to be considered compliant.

2. Matching publications by country

Why were pre- and post-intervention NPG publications matched for country? Of course it is possible that this variable has an influence on the quality of reporting, but the same is true for many other variables which were not taken into account (e.g., number of authors, COI statements, research area...). The problem I have with this decision is that publications for which no match on the country variable could be established for the pre- and post-intervention group were excluded. This introduces a bias in favour of publications from very prolific countries (for less prolific countries, finding matching pre- and post-intervention publications will be less likely). Whether or not this has an impact on the results of this study cannot be determined without looking at the full dataset without such exclusions.

The reason for matching by country was that, at the time the study was being planned, there was a widely held view that laboratory research from some countries was poorly reported, and we wanted these to be equally represented at least in the NPG manuscripts where our primary outcome would be measured. We appreciate that this may have led to exclusion of manuscripts from less prolific countries; we have compared the distribution of countries with that in a search for the same NPG journals with a year of publication of 2011-2015 and applying a (non validated) filter to select in vivo research; essentially we find 6 countries (Sweden, Belgium, South Korea, Israel, Portugal and Finland) which – on the basis of their frequency in this larger set – might have contributed a total of 24 studies (12 pairs) in vivo studies (there were 394 in vivo studies in total). Overall, 77 countries contributing 11% of the literature were not included, either because they were not sampled, or they were sampled but a pair was not found

A second problem with this decision is that the same restriction was not made for the non-NPG sample (as a minor note, I found it somewhat confusing that the preregistered inclusion criteria for non-NPG publications do not mention the country variable, but the final manuscript implies that matching publications for country was attempted but found to not be possible).
I suggest that the authors discuss this potential pitfall and the risk of collider bias and include a table or figure showing/comparing country distributions in all subgroups.

Thank you – we have included an additional table 8 to this effect, and added the text

"We anticipated difficulty in identifying matching articles, and in particular in matching non NPG articles by country; we did not seek to do so."

3. Preregistration

Both adherence to and deviation from the published preregistration should be made transparent. Instances of discrepancies which struck me (I may not have noticed all of them):

3.1 Coder recruitment:
The preregistration states: "We will recruit individuals experienced in the critical appraisal of published materials (through for instance involvement with previous systematic reviews)," whereas the present manuscript says "We had no prior requirements for the skills required of these individuals". However, later the authors go on to say that coders were actually recruited from two different populations: "Each manuscript was scored by 2 individuals, one with experience in systematic review and risks of bias annotation and one recruited from outside this community." The recruitment procedure should be made clear and unambiguous, and adherence to/deviation from the preregistration must be made explicit.

Thank you. While we had no prior requirements for participation, we did consider it sensible to use the range of experience from volunteers identified through the avenues described and available to us to ensure that each manuscript was assessed by at least one assessor who was highly experienced; we were concerned that having 2 less experienced assessors on a single manuscript would give a greater risk of their both making the same error, which would be diminished by having a more senior assessor. We say

"We sought to recruit individuals with a background in medicine or biomedicine at graduate or undergraduate level who we believed should have experience in the critical appraisal of published materials. However, we also recruited two senior school students on Nuffield Research Placements in our group."

and

"Because of the range of expertise available we ensured that each manuscript was reviewed by at least one assessor highly experienced in systematic review and critical appraisal."

3.2 Coding process:
The preregistration states "Monitoring of outcome assessment after 10 % of manuscripts have been scored and adjudicated; we will review performance and if there are questions that are highly represented in those resulting in disagreements we will review the training materials and amend them as appropriate." I could not find any statement in the final manuscript about whether this actually took place, and if so, what the result was.

Thank you – we now describe what happened …

"We had intended to monitor outcome assessment after 10% of manuscripts had been scored and reconciled, but the reconciliation process lagged behind the outcome assessment, and this was not done."

3.3 Primary and secondary outcomes:
I thank the authors for clearing up the confusion about the switch of primary and secondary outcomes between the preregistration and the final paper. However, it is still tricky to comprehend the exact differences between preregistration and final analyses, especially for readers who have not read the preregistration. I strongly recommend listing the outcomes/analyses in the same way as has been done in the preregistration, ideally next to the preregistered ones (e.g. in a table).

Thank you. We've read through it all carefully again, and in fact the primary outcome measure used was exactly as proposed (the proportion of studies reporting in vivo research fully compliant with the Landis 4); what differs is that in the protocol we also predefined "editorially significant changes"; these were (for different items) either 80% compliance with the item in question; or a 15% improvement from baseline. Our text now says …

"Our primary outcome was the proportion of articles describing in vivo experiments published by NPG after May 2013 that meet all of the relevant Landis 4 criteria. This is described in the statistical analysis plan deposited on the Open Science Framework (osf.io/hc7fk) on 7th June 2017 prior to database lock and before we had derived any outcome information. Following discussion with the Nature Publications Group editorial team we also set out in the protocol[11] some predefined "editorially significant changes" – either reaching compliance of 80% or an increase of 15% in compliance."

and

"While our primary outcome measure was unchanged, when writing our data analysis plan (and prior to any data inspection or analysis), we did change our criterion for measuring success, from "whether compliance (with the Landis 4 criteria, for in vivo research) in the post intervention group of articles reached 80%" to "the change in proportion of articles describing in vivo research meeting the 4 Landis criteria". This was because our primary intention had been to observe any effect of a change in publication policy, and, with the benefit of hindsight, this was not captured in our original primary outcome, but we recognise this as a limitation in our findings. We note, however, that the primary outcome used reflects better the title of the study protocol than does the primary outcome measure proposed in that protocol."

3.4
The preregistration states "We will conduct sub-group analyses in groups defined by country of origin; categorisation of research; and whether the study is predominantly in silico; in vitro; in vivo; or involves human subjects," but I could not find these analyses in the final manuscript.

Thank you; because the number of countries of origin other than the USA, and the number of studies which were predominantly in silico (we used studies which used computer code as a surrogate for the upper extent of this) was small, we did not analyse these further. There were no differences in compliance with the primary outcome measure dependent on whether the study included human research, or whether they included both in vivo and in vitro research or in vivo research alone. This is now given in the text, viz…

Because the number of manuscripts with a country of origin other than the USA, and the number of studies which were predominantly in silico (we used studies which used computer code as a surrogate for the upper extent of this) was small, we did not analyse these further. There were no differences in compliance with the primary outcome measure dependent on whether the study included human research, or whether they included both in vivo and in vitro research or in vivo research alone.

4. Coding and reviewer recruitment
4.1 Were the 10 papers in the "Gold standard" pool to train coders part of the studied sample? If it was, which of the many reviews for these papers were used for the study, and how was this determined?

We used the consensus score of the 5 experienced assessors. This is now described

4.2 Did any coder fail to reach sufficient concordance for three consecutive papers in the 10-paper pool?

Of 205 individuals registered with the project, 70 started but did not complete training. We don't know whether they gave up because they were failing or because they lost interest in the project given the complexity of task. We now clarify this, saying…

"205 individuals registered with the project, of whom 109 started at least one training manuscript, 38 completed their training and 35 assessed at least one manuscript."

4.3 How many papers had to be reconciled?

All papers had to be reconciled for at least one item, but reconciliation was limited to that item, not extending to all items for that manuscript

4.4 What is meant by "The agreement between the initial pair of outcome assessors ranged from..."? Were there other assessors beside the "initial pair" and the reconciler?

Apologies; we present the agreement between the initial pair of outcome assessors. Apart from the initial assessors and a single reconciler there were no additional assessors. We now try to make this clear.

"Each item for each manuscript was therefore scored by 2 (if there was agreement) or 3 (if there was disagreement) reviewers, except for the 10 manuscripts which served as the gold standard, which had been scored by 5 experienced assessors."

5. Exploratory studies

Exploratory studies were not included in some or all of the analyses. It did not become entirely clear to me if this only referred to the Landis 4 items or to all outcomes. How many publications were excluded due to this criterion? Does this explain the varying group sizes for the individual outcome reports? Or were exploratory studies counted as complying with the checklist?

Explicity exploratory studies were not included in the original selection, and so did not contribute in any way to the analysis. We do not know how many otherwise suitable manuscripts were thus excluded because we did not complete a screening log (a practice sometimes used in clinical trials); but we suspect the number of explicitly exploratory published research is very low.

6. Power calculations

The power analyses reported in the preregistration are exceptionally detailed and well thought-out, which I see as a great strength of this study, and have been further improved in this revision by drawing a tighter link to the preregistration.

Thank you

7. Editorially significant changes

Another great strength of the preregistration is that the authors define the size of an "editorially significant change or prevalence" (Table 1). Explicitly setting a "smallest effect size of interest" (SESOI) like this is crucial to ensure hypotheses are falsifiable, and they greatly increase the practical value of investigations like this one. In this particular case I think that considering "editorial significance" is an excellent and helpful idea and it would be nice to spark a discussion about what should be considered an editorially significant change.

However, as far as I could see, not all of the SESOIs the authors set for themselves were followed up with the appropriate tests in the final manuscript. That would mean to test both a) if changes/differences are greater than zero and b) if they are significantly smaller than the SESOI. Unless I missed something, it seems that this was only done for the very first analysis (change of proportion of NPG in vivo studies reaching full compliance with the Landis 4 criteria before vs. after the intervention). In my view it would be important to add these tests against the SESOI for each confirmatory analysis (i.e., each comparison that is described in the preregistration).
NB: The authors do explain the level of power they had to detect changes of 15% or more (15% is defined as the SESOI for many of the analyses in the preregistration), but this is not sufficient to conclude the absence of differences as large as 15% when no significant result is obtained in a null-hypothesis test (see e.g. Lakens, Scheel, & Isager, 2018).

On a minor note, as happy as I am about Table 1 in the preregistration, I was wondering why the values set in there do not seem to be discussed in the text. I think it would be nice to add this to the final manuscript (i.e. mention and discuss "editorially significant changes" and explain the reasoning behind the SESOIs of 80% and 15% which are mentioned in the preregistration).

Thank you – we have added an additional table to indicate, for each item, whether the observed performance post-intervention was consistent with 80% compliance or with a 15% improvement in compliance. We have colour-coded this to make it more accessible, but would be happy to revert to plain text if this is preferred. We have also discussed this in the text.

8. Matching in vivo and in vitro articles

The authors report picking papers for the non-NPG control group by matching non-NPG papers to the NPG groups based on publication date and whether they reported in vitro or in vivo research. However, the final result are uneven group sizes before vs. after 1st May 2013. The authors write: "The difference in numbers for NPG and non-NPG before and after 1st May 2013 is because some of the NPG “before” papers matched best with publications in other journals published in the few months following May 2013." I find this problematic, because the crucial comparison for all groups is before vs. after intervention. Prioritising a match on the in vivo/in vitro status criterion over the before vs. after

criterion therefore makes little sense to me - it means that a "match" is being established between two publications that do not need to be matched, and a match between two studies that ought to match is sacrificed. Given that none of the resulting subgroups is particularly small, one may argue that any resulting problems are negligible - but note that e.g. the difference in group size between NPG in vivo before and non-NPG in vivo before is 20%, which is substantial in my view and lowers statistical power to detect changes in the non-NPG group.

There are at least two factors which might influence reporting quality; it was our view that differences between subject areas were large (see for instance the in vivo stroke community now compared with other communities); while the influence of a change in NPG policy on manuscripts published in non-NPG was, we thought, likely to be smaller. We acknowledge the loss of power which this has occasioned.

9. Reporting of results

9.1
First, all descriptive and test statistics are reported in the text, which I find very hard to read. Almost all of the reported numbers are also presented in tables 2-6, which in my view is a much better format for this. I would recommend to reduce redundancy between text and tables as much as possible by replacing most of the numbers in the text with references to the respective table, adding test statistics which are currently only reported in the text to the tables (Chi-square and df), and structuring the tables such that it becomes clear which of the preregistered analyses each result refers to. This would make the Results section much easier to comprehend, but it would also reduce additional sources of error. For example, the number of NPG in vitro studies after 1st May 2013 seems to be misreported as 176 in the text, although the (I assume) correct number 182 is apparent in Figure 1 and Table 3. On a related note, the abstract seems to contain a similar error: "The number of NPG publications meeting all relevant Landis 4 criteria increased from 0/203 prior to May 2013 to 31/181 (16.4%)" It seems to me that these should be 0/204 and 31/189, respectively, and they sum to 393, not 394.

Thank you; we have streamlined the reporting as you discuss, removing redundancy, and have corrected errors.

9.2
The text in the Results section should be structured and labelled in the same way as the analyses are described in the Method section (e.g. by numbering them).

Thank you, done

9.3
The size of the basic set of studies that compliant studies are compared against for the different checklist items varies wildly, and I found it hard to understand the reason for this. For example, for NPG in vivo studies there were 203 "before" and 189 "after" publications, but studies mentioning randomisation are reported as 14/169 before and 97/151 after. I assume that it is due to the fact that most items did not apply to all studies, and such studies were then excluded for the individual comparisons. In any case I believe it would be good to make this clearer to avoid confusion.

Thank you, comment also raised by associate editor, response copied below

We hope we have clarified our reporting; the figures you draw from the results section for randomisation, blinding and sample size calculations arise for "compliant by reporting they did the thing", but also "compliant by virtue of saying they didn't do the thing". For a study which was "yes" for part 1, part 2 was not relevant, and the study was therefore excluded from the denominator for this question. There was a little variability in the assessments offered, in that the numbers are somewhat contradictory (you can see this in table 4, where the numbers where eg blinding is considered not relevant seems to be different for each of these questions). Because we cannot know whether the assessors were correct on part 1 or part 2, we present the numbers in the table and use the

denominator identified for each individual question. We have clarified the textual expression of these results and added this as a further limitation,

"We encountered a further unexpected problem when assessing compliance with reporting of blinding, randomisation and with sample size calculations. These were assessed with pairs of questions, first did the study report doing it (yes/no/not relevant); and second, did they at least mention it (yes/no/not relevant). If a study was "yes" for the first question, assessors were instructed to score the second as not relevant. Therefore, the number scored as "not relevant" for the second question should represent the sum of those scored as "yes" and as "not relevant" for the first. This was not always the case (for in vivo research occurring in 0.1%, 0.8% and 6% of assessments for sample size calculation, randomisation and blinding respectively but we did not become aware of this problem until after database lock. Any impact of this shortcoming is likely to be small. "

9.4
Significant changes in "statistical reporting" are claimed but not backed up with any test statistics (admittedly this feels somewhat ironic); these should be added, preferably also in form of a table rather than in the text.

In Table 5 we report significant improvements in reporting in vivo research in the exact n, t-test defined as 1 or 2 sided, and demonstrating the data meet the assumptions of the test used; in reporting in vitro research in the exact n, t-test defined as 1 or 2 sided, and reporting whether the findings represent technical or biological replicates. IN the text we say ..

Statistical reporting (Table 5): For in vivo studies reported in NPG articles there were significant improvements in the reporting of exact numbers (from 46% to 69%, $X^2 = 22.07$, df = 1, adj p = 0.0004), of whether t-tests were defined as one or two sided (from 46% to 71%, $X^2 = 17.80$, df = 1, adj p = 0.003), and whether the assumptions of the test had been checked (from 9% to 27%, $X^2 = 18.58$, df = 1, adj p = 0.002). For in vitro experiments described in NPG articles there were significant improvements in the reporting of the exact numbers (from 32% to 70%, $X^2 = 12.60$, df = 1, adj p = 0.05); of whether data represented technical or biological replicates (from 57% to 75%, $X^2 = 13.29$, df = 1, adj p = 0.035); and whether t-tests were defined as one or two sided (from 47% to 72%, $X^2 = 16.18$, df = 1, adj p = 0.008). For in vivo and in vitro studies described in non-NPG articles there was no significant change in any of the items relating to statistical reporting.

10. Discussion

Tying back in with my first point, I think in the Discussion section the authors could make it a bit more clear what the results mean in relation to which outcomes could have reasonably been expected. I also think it would be great to pick up the notion of "editorially significant changes" from the preregistration and discuss this concept.

Thank you. We have added text, viz ..

"Prior to the study we identified achievement of 80% compliance, or an absolute improvement of 15% in the reporting of an item, as being the minimal change which would represent an important effect of an editorial intervention. In the NPG cohort, for 62 items the 95% confidence intervals of the observed change fell below 15% for 11 items, included 15% for 40 items, and were above 15% for 5 items. For 3 items there were insufficient data to calculate 95% confidence intervals, and for 3 items baseline performance already exceeded 85%.

Power calculations in primary research are often considered unfeasible, on the basis that prior to doing the study the effect size is not known. Our approach here – of identifying a smallest effect size of interest – is increasingly widely used, and has allowed us to demonstrate if any change observed might be as large as the smallest effect size of interest, is definitely that large, or is definitely not that large. We hope that those using our findings to guide their own improvements will find this helpful, and recommend the approach for use in future studies."

One final minor comment is that regarding the authors' suggestion that machine learning approaches could potentially decrease bias in assessments like the one undertaken here and/or make the process more efficient: To this I would add that any coding unreliability the authors observed in their study hints at problems regarding how researchers report their work - in other words, if two trained reviewers cannot agree whether or not I report if my study was blinded, maybe the problem lies with me rather than the reviewers. So working on the end of reporting standards might be as important or even more important than finding new ways of coding research reports.

Thank you. We have added

"For more complex items it is likely that machine learning approaches using for instance convoluted neural networks may be more successful, and this is a current focus of our research. We hope that, by making the dataset for this study available, this might be used for instance for distant supervised learning in such systems. However, the extent of disagreement between our trained assessors suggests that the language used to describe experiments in biomedicine is not altogether clear, and both machines and human may require greater clarity in reporting to fully understand published research.

11. A list of additional (very) minor comments:

- The manuscript still contains a significant number of typos, in particular misplaced punctuation (missing periods, double periods, superfluous spaces, double parentheses, etc).

Thank you – we've tried again to catch these

- "Poor replication of in vitro molecular and cellular biology studies has also been reported" could be read to imply that *in vivo* research is unreplicable (because this sentence follows a discussion of in vivo research). The word "also" makes this sentence slightly ambiguous, maybe consider rephrasing.

Thank you: we now say

"Poor replication of in vivo and in vitro research has been reported [7-9] and this has been attributed in part to poor descriptions of the experimental and analytical details."

- First line of the following paragraph: It would be good to add a comma at the beginning of the sentence to avoid confusion ("In May 2013, Nature Journals", otherwise it could be read as "In May, 2013 Nature Journals ...")

Thank you

- p. 4, first line: "completion of a checklist at the stage of manuscript revision was introduced)" --> An earlier sentence says that the checklist was introduced at the stage of manuscript acceptance, not revision. Which one is correct?

Thank you – at revision, now consistent ..

“Intervention Mandatory completion of a checklist during manuscript revision.”

and

“In May 2013, Nature Journals announced a change in editorial policy which required authors of submissions in the life sciences to complete a checklist, during manuscript revision, indicating whether or not they had taken certain measures which might reduce the risk of bias…”

- p. 4: "A matching pre-intervention publication then identified, as described above." --> should probably be "was then identified", or similar

Thank you, changed

- p. 6, last sentence: The second part is not a grammatically correct sentence, please rephrase.

Thank you, changed

- p. 8, 2nd paragraph, 2nd sentence: Please explain/define the word "reconciler" (I was initially struggling to understand this)

Thank you, we have explained this (we hope) more clearly, saying

“12 individuals also reconciled conflicting outcome assessments”

- p. 10, 2nd paragraph, last sentence: "The prevalence of reporting the different items before and after is shown in table 2" I think this should be Table 4, not 2.

Thank you, now actually table 5

- p. 11, 4th paragraph, 2nd sentence: According to my calculation, the 62% figure is 62.5%, which I assume should be rounded to 63%. But I might have made a mistake in my recalculation.

Thank you: I was taught (perhaps wrongly) to round to the nearest even number, but I’m happy to go along with different conventions

- Figure 1: This figure is exceptionally helpful (no criticism).

Thank you

- Figure 2: Thank you for changing the colour scheme, this is a great improvement.

Thank you

- Figure 3: As mentioned in the paper, power for the interrupted time series analysis was substantially reduced due to the temporal distribution of the data, which in my view means that the results should be treated with caution. Therefore I would suggest to change Figure 3 to put more visual emphasis on the data points and less emphasis on the fitted lines.

Thank you, done

- Figure 4: Font size should be increased.

Thank you

- Tables: In the text, the comparison group is consistently labelled as "non-NPG" but in the tables it is referred to as "matched". I think text and tables should use the same label, in my view "non-NPG" is the better choice.

Thank you, fixed

#3 Submitted by Michèle Nuijten

The authors have addressed most of my comments and those of the other two reviewers in great detail, which, in my view, has greatly improved the manuscript.

Thank you

The only concern I still have is the way the comparison group is treated. In my eyes, the comparison articles have every characteristic of a control group: they are matched to the NPG publications, and they are analyzed to “establish whether any improvement at NPG was simply a reflection of a general improvement in similar articles across biomedical publishing” (from the cover letter of the authors). This seems to me to be an exact description of a classic control group: it is not sufficient to simply monitor a change over time in NPG publications, you need to compare this change to a change in other journals without such guidelines.

If the objective of the authors is to only monitor if the reporting standards improved in NPG publications, then they wouldn’t need a comparison sample. However, it would then also be much harder to justify that the change was caused by the change in journal policy. On the other hand, the authors did go to great length to create a comparison group. Why do that if this group is not included in the formal analyses?

In short, I am confused why the comparison group is there if it is not to serve as a formal control. If the goal is to see whether editorial policies had an effect on reporting standards, I would say the best, feasible design here would be a quasi-experimental design with a non-equivalent control group and a pre- and post-test (i.e., treating the comparison group as a control). I would be interested to hear if the authors agree, and if not, why.

Thank you for the opportunity to attempt to clarify this further. Our primary purpose was to evaluate whether the change at NPG had led to improvement; a secondary question was whether that improvement simply reflected a general improvement in the biomedical literature. This is an important distinction because – from the perspective of the potential users of this work, i.e. biomedical journals – their stated ambition is to improve reporting quality in their own journal, and what happens to quality at other journals is a secondary concern. The vast majority of journals which contributed articles to the non-NPG cohorts would claim that they too had strategies to improve reporting (and indeed we do for some items see a general trend to improvement).

We do appreciate, however, that our presentation makes it difficult for readers to get a sense of how the improvement in the NPG cohort compares with that in the non NPG cohort. We therefore present a new figure (5) in which we show the 95% confidence intervals for the change in reporting in each cohort, to provide a visual summary. Because a formal statistical analysis was not proposed in our protocol or data analysis plan we would be very reluctant to perform one at this stage; although of course since our data are available online it would be possible for others to perform such an analysis.

Findings of a retrospective, controlled cohort study of the impact of a change in Nature journals' editorial policy for life sciences research on the completeness of reporting study design and execution.

Authors

The NPQIP Collaborative group

Corresponding author

Professor Malcolm Macleod: Malcolm.Macleod@ed.ac.uk, 07786 265166

Conflicts of interest

None declared.

Keywords

Risk of bias, reporting, methodological quality, study design, reporting guidelines

Acknowledgements

Funding

Laura and John Arnold Foundation.

Abstract

Objective: To determine whether a change in editorial policy, including the implementation of a checklist, has been associated with improved reporting of measures which might reduce the risk of bias.

Methods: The study protocol has been published at DOI: 10.1007/s11192-016-1964-8.
Design: Observational cohort study
Population Articles describing research in the life sciences published in Nature journals, submitted after May 1st 2013.
Intervention Mandatory completion of a checklist at the point of manuscript revision.
Comparators (1) Articles describing research in the life sciences published in Nature journals, submitted before May 2013; (2) Similar articles in other journals matched for date and topic.
Primary Outcome Change in proportion of Nature publications describing in vivo research published before and after May 2013 reporting the "Landis 4" items (randomisation, blinding, sample size calculation, exclusions).

We included 448 NPG papers (223 published before May 2013, 225 after) identified by an individual hired by NPG for this specific task, working to a standard procedure; and an independent investigator used Pubmed "Related Citations" to identify 448 non- NPG papers with a similar topic and date of publication in other journals; and then redacted all publications for time sensitive information and journal name. Redacted manuscripts were assessed by 2 trained reviewers against a 74 item checklist, with discrepancies resolved by a third.

Results: 394 NPG and 353 matching non-NPG publications described in vivo research. The number of NPG publications meeting all relevant Landis 4 criteria increased from 0/203 prior to May 2013 to 31/181 (16.4%) after (2-sample test for equality of proportions without continuity correction, $X^2 = 36.2$, df = 1, $p = 1.8 \times 10^{-9}$). There was no change in the proportion of non- NPG publications meeting all relevant Landis 4 criteria (1/164 before, 1/189 after). There were more substantial improvements in the individual prevalences of reporting of randomisation, blinding, exclusions and sample size calculations for in vivo experiments, and less substantial improvements for in vitro experiments.

Conclusions. There was a substantial improvement in the reporting of risks of bias in in vivo research in NPG journals following a change in editorial policy, to a level that to our knowledge has not been previously observed. However, there remain opportunities for further improvement.

Background

Few publications describing *in vivo* research report taking specific actions designed to reduce the risk that their findings are confounded by bias, and those that do not report such actions give inflated estimates of biological effects. Strategies and guidelines which might improve the quality of reports of *in vivo* research have been proposed, [1,2] and while these have been endorsed by a large number of journals there is evidence that this endorsement has not been matched by a substantial increase in the quality of published reports [3].

Poor replication of *in vitro* molecular and cellular biology studies has also been reported [4,5] and this has been attributed in part to poor descriptions of the experimental and analytical details.

In May 2013 Nature Journals announced a change in editorial policy which required authors of submissions in the life sciences to complete a checklist, at the time of manuscript acceptance, indicating whether or not they had taken certain measures which might reduce the risk of bias and to report key experimental and analytical details; and in their submission to detail where in the manuscript these issues were addressed [6].

The aim of this study was to determine whether the implementation of this checklist for submissions has been associated with improved reporting of measures that might reduce the risk of bias. To establish whether any observed change in quality was a simply a secular trend occurring across all journals we matched each included publication with a publication in a similar subject area published at around the same time by a different publisher.

Methods

The methods are described in detail in the published study protocol [7], and the data analysis plan and analysis code were articulated prior to database lock and registered on the Open Science Framework (https://osf.io/mqet6/#). The complete study dataset including PMIDs (but not, for copyright reasons, the source pdfs) of included articles is available on Figshare (10.6084/m9.figshare.5375275).

In this observational cohort study we aimed to determine whether the implementation of a checklist for submissions has been associated with improved reporting of measures which might reduce the risk of bias. The study populations comprised (1) Published articles accepted for publication in Nature journals, which described research in the life sciences and which were submitted after May 1st 2013, at which time the mandatory completion of a checklist at the stage of manuscript revision, was introduced. This checklist required authors to indicate where details relating to study design could be found in the manuscript at the point of manuscript revision. and before November 1st 2014; (2) Published articles accepted for publication in Nature journals in the months preceding May 2013, which describe research in the life sciences; and (3) manuscripts from other journals matched for subject area and time of publication. We measured the change in the reporting of items included in the checklist.

Identification of relevant manuscripts

NPG publications: One individual was specifically employed by Nature to select studies which (a) described in vivo or in vitro research; (b) was published in Nature, Nature Neurology, Nature Immunology, Nature Cell Biology, Nature Chemical Biology, Nature Biotechnology, Nature Methods, Nature Medicine or Nature Structural and Molecular Biology. First, they identified papers accepted for publication with an initial submission date later than May 1st, 2013. Beginning with the then-current issues (volume corresponding to year 2015), they worked backwards in time, ensuring the submission date was after 1st May 2013, collecting papers until they had 40 Nature papers and 20 each from other titles ("Post intervention" group). They then used a similar process to identify papers submitted for publication before 1st May 2013, matched for journal and for country of origin, starting with the May 2013 issue and working backwards, ensuring that the date of submission was after 1st May 2011 ("pre-intervention" group). Where no match could be found with a submission date after 1st May 2011 (i.e. in a two year period) then the non-matched post intervention publication was excluded from analysis and a replacement post intervention publication selected, as above, with a matching pre-intervention publication then identified, as described above. Publications describing research involving only human subjects were not to be included. A Nature editorial administrator independent of publishing decisions reviewed manuscripts selection against the inclusion criteria and found some (less than10%) had been included incorrectly; they replaced these with manuscript pairs that they selected according to the inclusion algorithm. The published files corresponding to the publication pdfs (including the extended methods section, extended data and other supplementary materials) were used to generate pdfs for analysis. These were provided to a member of our research team (RM) at a different institution who used Adobe Acrobat to redact information relating to author names or affiliations, dates, volumes or page numbers; and the reference list; to minimise awareness of outcome assessors to whether the manuscript was pre- or post- intervention.

Non- NPG publications: The same individual was responsible for identifying matching publications in other journals. They identified the NPG publication in PubMed and searched for "related citations" with the same calendar month of publication, selecting the first that was not published in an NPG Journal that also matched for whether it reported in vivo research, in vitro research, or both. If no matching related citation was found the extended the window of publication by 2 months, continuing until a matching publication was found. Because of a limited number of potential matching publications it was not possible to match non NPG manuscripts by country. The individual making this selection paid no further part in the study.

Outcome assessment

The Nature checklist focussed on transparency in reporting and availability of materials and code, reflected in 10 items. We designed a series of questions (Appendix 1) to establish whether a given publication met or did not meet the requirements of the checklist. Where a manuscript described both in vivo and in vitro research, the series of questions was completed for each. Where there is more than one in vitro experiment or more than one in vivo experiment the question was considered in aggregate; that is, all experiments had to meet the requirements of the checklist item for it to be considered compliant.

Five researchers experienced in systematic review and risk of bias annotation scored the same 10 publications using our series of questions. Disagreements were resolved by group discussion, to arrive at a set of “Gold standard” answers for these 10 publications. We also used this experience to write a training guide for those seeking to use the checklist. We then used social media platforms and mailing lists to recruit outcome assessors. We had no prior requirements for the skills required of these individuals, but most had a background in medicine or biomedicine at graduate or undergraduate level; two were senior school students on Nuffield Research Placements in our group. After reviewing the training materials outcome assessors were invited sequentially to score publications from the “Gold standard” pool until their concordance with the Gold standard responses was 80% overall, and was 100% for the components of the primary outcome measure, for three successive publications. At this point we considered them to be trained. The training platform remains available for continuing professional development, at https://ecrf1.clinicaltrials.ed.ac.uk/npqip/Review/TrainingCover.

Pdf files of included manuscripts were uploaded to a bespoke website. Trained assessors were presented with manuscripts for scoring in random order. Each manuscript was scored by 2 individuals, one with experience in systematic review and risks of bias annotation and one recruited from outside this community. Disagreement between assessors were reconciled by a third, experienced individual who was not one of the original reviewers, who could see the responses previously given but not who the initial reviewers were.

Statistical analysis plan.

Given our focus on the reporting of measures to reduce the risks of bias we took as our primary outcome measure a composite measure of the proportion of publications meeting the relevant measures identified by Landis et al as being most important for transparency in reporting in vivo research. These are covered by items 2, 3 4 and 5 of the checklist and relate to the reporting of randomisation; of the blinded assessment of outcome; of sample size calculations; and of whether the manuscript described whether samples or animals were excluded from analysis. Importantly, checklist compliance did not require for example that the study was randomised; but rather that the authors stated whether or not it was randomised. The evaluation principle was to determine if someone with reasonable domain-knowledge could understand the parameters of experimental design sufficiently to inform interpretation. It has been argued that these measures might not be as relevant for exploratory studies, and for these we recorded the item as “not relevant”. We defined exploratory studies as those where hypothesis testing inferential statistical analyses were not reported. Where an item was not relevant for a publication (for instance with studies using transgenic animals where group allocation had been achieved by Mendelian randomisation) we considered compliance as meeting the remaining relevant criteria. Where a publication described both in vivo and in vitro experiments we analysed each type of experiment separately.

Our primary outcome was the change in the proportion of publications describing in vivo experiments published by NPG before and after May 2013 that meet all of the relevant Landis 4 criteria. We used the two-sample proportion test (prop.test) in R without the Yates continuity correction and two sided hypothesis testing to be sensitive to the possibility that performance might have declined rather than improved. Secondary

outcomes were whether the proportion of publications describing in vivo experiments published by NPG after May 2013 which meet all four of the Landis 4 criteria was 80% or higher (Wald test; wald.ptheor.test, RVAideMemoire in R); the change in the proportion of publications describing in vitro experiments published by NPG before and after May 2013 which meet all four of the Landis 4 criteria (two sample proportion test as above); and the change of proportions in adequate reporting of statistical analysis details, individual Landis criteria, and descriptions of animals; reagents and their availability; sequence, structure or computer code deposition; and items relating to the involvement of human subjects or materials in included studies.. For the matching publications from non-NPG journals the secondary outcomes were the change in the proportion of publications describing in vivo experiments published before and after May 2013 which met all of the Landis 4 criteria (two sample proportion test); whether the proportion of publications describing in vivo experiments published after May 2013 which met all four of the Landis 4 criteria was 80% or higher (Wald test); the change in the proportion of publications describing in vitro experiments published before and after May 2013 which meet all four of the Landis 4 criteria 4 (two sample proportion test); and the change of proportions in adequate reporting of statistical analysis details, individual Landis criteria, and descriptions of animals; reagents and their availability; sequence, structure or computer code deposition; and items relating to the involvement of human subjects or materials in included studies. For each of these outcomes we compared the changes observed in NPG publications with that observed in non NPG publications. For each secondary analysis we used Holm Boneferroni correction using the p.adjust option for prop.test in R to account for the number of comparisons drawn, as described in Appendix B of the Data Analysis Plan. We also used interrupted time series analysis for each checklist item in an attempt to distinguish a discrete “shift” in performance from an upward “drift”, as described in the data analysis plan. A number of tertiary outcomes are described in the study protocol and statistical analysis plan and are reported in the supplementary material.

Power Calculations

In planning the study we performed power calculations in STATA. The power to detect changes in reporting depended on the baseline performance; with baseline prevalence of compliance of 10% we had 80% power to detect an absolute increase of 13% to 23% at a significance level of $p<0.01$; with baseline compliance of 50% we had 80% power to detect an absolute increase of 16% to 66% at a significance level of $p<0.01$. For secondary outcomes we had lower statistical power, but after correction for the number of comparisons made we had at worse 67% power to detect a 15% improvement in the reporting of any individual item.

Results

896 publications were identified and uploaded for outcome ascertainment, 448 in each cohort. 2 non-NPG manuscripts were excluded because they did not meet the inclusion criteria, and we identified 4 NPG and 9 non-NPG manuscripts included more than once. 444 NPG publications and 437 non-NPG publications underwent outcome assessment. One NPG publication and one non-NPG publication were adjudged at the time of outcome assessment to report neither in vivo nor in vitro research and so were excluded. The analysis is therefore based on 443 NPG publications (219 before and 224 after 1st May 2013) and 436 non-NPG publications (194 before and 242 after 1st May 2013) (Figure 1). The difference in numbers for NPG and non-NPG before and after 1st May 2013 is because some of the NPG “before” papers matched best with

publications in other journals published in the few months following May 2013. Overall, 43% of matched pairs had dates of publication within 1 month, 54% within 2 months, 64% within 3 months and 81% within 6 months of each other (range -11 to +22 months). 239 publications described only in vivo research, 132 described only in vitro research, and 508 described both. The source journals are given in Table 1; in total 198 different titles contributed matching publications (median manuscripts per publication 1, range 1 – 47). The PMIDs of included publications are listed in the data supplement.

205 individuals registered with the project, of whom 38 completed their training and 35 assessed at least one manuscript. 12 also served as reconcilers, and the web interface was programmed to ensure that they were not offered for reconciliation manuscripts that they had previously adjudicated. Including reconciliation, the median number of manuscripts scored was 13 (range 1 to 441). The agreement between the initial pair of outcome assessors ranged from being no better than chance at 50% (in vivo studies, Implementation of statistical methods and measures: “Is the variance similar (difference less than two-fold) between the groups that are being statistically compared?”) to 98% (in vivo studies, “Does the study report the species?”). Median agreement was 82%. (IQR 68 – 89%).

Reporting of the Landis 4 items: The proportion of NPG in vivo studies reaching full compliance with the Landis 4 criteria increased from 0% (0/204) to 16.3% (31/190) ($X^2 = 36.1$, df = 1, $p = 1.8 \times 10^{-9}$), but remained significantly lower than the target of 80% (95% CI 11.7% to 22.3%, Wald test v 80% t = -15.4, $p = 2.2 \times 10^{-16}$).

For randomisation to experimental group, the preferred standard is that the manuscript describes which method of randomization was used to determine how samples or animals were allocated to experimental groups, although manuscripts were also compliant if they included a statement about randomization even if no randomization was used. The proportion of NPG in vivo studies reporting randomisation was 1.8% (3/170, 95% CI 0.6 to 5.3%) before and 11.2% (19/170, 95% CI 7.2 to 16.9%) after ($\chi^2 = 12.4$, df = 1, adj p = 0.054). The proportion of studies mentioning randomisation even where it was not reported increased from 8.3% (14/169, 95% CI 5.0 to 13.5) to 64.2% (97/151, 95% CI 56.3 to 71.5%)($\chi^2 = 110.2$, df = 1, adj $p = 3.2 \times 10^{-14}$). Figure 2(a) shows change in the proportion of studies meeting these criteria before and after the change in editorial policy.

For blinding, the preferred standard is that the manuscript describes whether the investigator was blinded to the group allocation during the experiment and/or when assessing the outcome, although manuscripts were also compliant if they included a statement about blinding even if no blinding was done. The proportion of NPG in vivo studies reporting blinding during group allocation or outcome assessment or both increased from 4% (8/198, 95% CI 2.0 to 7.9%) to 22.8% (42/184, 95% CI 17.3 to 29.4%)($X^2 = 29.6$, df = 1, adj $p = 7.6 \times 10^{-6}$). The proportion of studies mentioning blinding even where it was not reported increased from 1.6% (3/182, 95% CI 0.5 to 5.0%) to 55.3% (73/132, 95% CI 46.8 to 65.6%)($X^2 = 120.1$, df = 1, adj $p < 3.2 \times 10^{-14}$). Figure 1(b) shows change in the proportion of studies meeting these criteria before and after the change in editorial policy.

The proportion of studies reporting animals excluded from analysis increased from 13.9% (28/202, 95% CI 9.7 to 19.3%) to 30.7% (58/189, 95% CI 24.5 to 36.7%)(X^2 = 16.1, df = 1, adj p = 0.008). Figure 1(c) shows change in the proportion of studies meeting these criteria before and after the change in editorial policy.

For sample size calculations, the preferred standard is that the manuscript describes how the sample size was chosen to ensure adequate power to detect a pre-specified effect size, although manuscripts were also compliant if they included a statement about sample size estimate even if no statistical methods were used. The proportion of studies reporting an a priori sample size calculation increased from 2.0% (4/196, 95% CI 0.8 to 5.3%) to 14.8% (27/182, 95% CI 10.4 to 20.8%)(X^2 = 20.5, df = 1, adj p = 0.0008). The proportion of studies mentioning sample size even where a sample size calculation was not reported increased from 1.6% (3/192, 95% CI 0.5 to 4.7%) to 58.4% (90/154, 95% CI 50.5 to 66.0%)(X^2 = 140.7, df = 1, adj p < 3.2×10^{-14}). Figure 1(d) shows change in the proportion of studies meeting these criteria before and after the change in editorial policy.

For NPG in vitro studies, the proportion reaching full compliance with the Landis 4 criteria was 0% (0/159) before and 3.3% (6/176) after (X^2 = 6.8, df = 1, Holm Bonferroni adjusted p = 1.00). The proportion of studies reporting randomisation was 0% (0/149) before and 2.9% (5/173, 95% CI 1.2 to 6.8%) after (X^2 = 4.4, df = 1, adj p=1.00). The proportion of studies mentioning randomisation even where it was not reported increased from 0% (0/149) to 15.6% (97/151, 95% CI 10.8 to 21.9%)(X^2 = 25.3, df = 1, p = 6.9×10^{-5}). The proportion of studies reporting blinding during group allocation or outcome assessment or both was 3.9% (6/155, 95% CI 1.8 to 8.4%) before and 8.9% (16/179, 95% CI 5.6 to 14.1) after (X^2 = 3.467, df = 1, p=1.00). The proportion of studies mentioning blinding even where it was not reported increased from 0.7% (1/150, 95% CI 0.1 to 4.6%) to 15.9% (25/157, 95% CI 11.0 to 22.5) (X^2 = 23.0, df = 1, p=0.0002). The proportion of studies reporting exclusions from analysis was 8.2% before (13/159, 95% CI 4.8 to 13.6%) and 15.9% (29/182, 95% CI 11.3 to 22.0%) after (X^2 = 4.73, df = 1, p =1.00). The proportion of studies reporting an a priori sample size calculation was 1.3% (2/155, 95% CI 0.3 to 5.0%) before and 7.9% (14/177, 95% CI 5.1 to 13.5%) after (X^2 = 8.7106, df = 1, p = 1.00). The proportion of studies mentioning sample size even where a sample size calculation was not reported increased from 3.3% (5/153, 95% CI 1.4 to 7.6%) to 28.5% (47/165, 95% CI 22.1 to 35.8%)(X^2=36.9, df = 1, p=1.8×10^{-7}).

The proportion of matching (non-NPG) in vivo studies reaching full compliance with the Landis 4 criteria was 1% (1/164) before and 1% (1/189) after (X^2 = 0.01, df = 1, adj p = 1.00), and for in vitro studies, the proportion of non-NPG studies reaching full compliance with the Landis 4 criteria was 0% (0/134) before and 1% (1/165) after (X^2 = 0.8, df = 1, adj p = 1.00). The prevalence of reporting the different items before and after is shown in table 2; there was no significant change in reporting of any of the individual Landis 4 criteria for either in vivo or in vitro research.

Statistical reporting: For in vivo studies reported in NPG manuscripts there were significant improvements in the reporting of exact numbers (from 46% to 69%), of whether t-tests were defined as one or two sided (from 46% to 71%), and whether the assumptions of the test had been checked (from 9% to 27%). For in vitro

experiments described in NPG manuscripts there were significant improvements in the reporting of the exact numbers (from 32% to 70%); of whether data represented technical or biological replicates (from 57% to 75%); and whether t-tests were defined as one or two sided (from 47% to 72%). For in vivo and in vitro studies described in non-NPG publications there was no significant change in any of the items relating to statistical reporting.

Other checklist items: For reporting of details of animals used, reporting of animal species and strain was high even before the change in editorial policy. There was no significant change in reporting any of these items in NPG- and non-NPG manuscripts, or in the reporting of details of antibodies used. For in vitro research, there was an increase in the proportion of studies in NPG manuscripts reporting recent mycoplasma testing of the cell lines used (from 1% to 26%) but not for non-NPG manuscripts (1% before, 1% after). For reporting and availability of accession data (eg DNA or protein sequence deposition) and computer code there were no significant changes for either NPG or non-NPG publications. Finally, there were no significant changes in the reporting of items relating to human subjects or the use of human materials, but for most items the number of publications for which these were relevant was very low indeed.

We were also interested in whether changes in reporting had occurred as a step change at the time of the change in editorial policy; whether there was an initial improvement with then a return to previous performance; or if there was an ongoing improvement in reporting. To address these we conducted an interrupted time series analysis, to estimate the rate of change before the intervention; any step change at the time of the intervention; and the rate of change after the intervention. We grouped publications in 3 month periods starting November 2011, and for each quarter calculated the proportional compliance with the criteria in question. Because publications were not evenly distributed across time the analysis is of substantially reduced power, but the fitted lines for overall compliance and for each component of the Landis checklist for in vivo research are shown in Figure 3. It appears that with the exception of sample size calculation there is a continuing improvement over time in both NPG and non NPG publications; for sample size calculations the improvement is only seen in NPG publications. Figure 4 shows radar charts of compliance for each checklist item in NPG ans non NPG manuscripts before and after May 2013.

Discussion

The change in editorial policy at NPG was associated with major improvements in reporting of randomisation, blinding, exclusions from analysis and sample size calculations. For the highly challenging primary outcome measure, full compliance increased from zero to 16%. This falls short of the target compliance of 80%, but should be seen in the context firstly that only 1 of 1073 publications from 2009-10 from leading UK institutions achieved this standard[8]; and secondly that overall compliance of 80% would require compliance with individual items of around 95%.

The checklist relates to transparency in reporting, and manuscripts were judged to be compliant if they either reported measures to address that risk of bias, or reported that such measures were not taken. For reports of in vivo research, compliance for randomisation, blinding, reporting of exclusions and sample size calculations

in NPG publications reached 68%, 62%, 31% and 64% respectively. For non NPG publications the performance was 12%, 5%, 12% and 3%. The figures for NPG publications are similar to those recently reported for in vivo research published in the journal “Stroke” [9], which began requiring reporting of such details following the publication of good practice guidelines in 2009 [10]; and where performance was found to be substantially higher than for in vivo research published in other American Heart Association journals.

For reports of in vitro research, compliance was substantially lower. There have been few systematic attempts to measure the quality of reporting of measures to reduce the risks of bias in vitro research, and our findings suggest that, both in NPG and non NPG journals, this remains low. There were improvements in reporting randomisation, blinding and sample size calculations in NPG descriptions of in vitro research, but only to 18%, 23% and 34% respectively. For non NPG the equivalent figures were 3%, 1% and 1%. There were no significant changes in the reporting of exclusion of in vitro data, with post intervention compliance of 16% in NPG publications and 6% in non NPG publications.

For other checklist items, changes in performance were less dramatic, but there appeared to be incremental improvements across most of the items measured, although few of these breached our rather parsimonious adjustment for multiple testing. In spite of substantial attention given to the importance of reporting the sex of experimental animals this was only done in 52% of post intervention NPG studies and in 36% of non NPG studies.

Ours is an observational study, and it is possible that other (related or unrelated) changes were responsible for much if not all of the differences seen. These changes were not observed at other journals (at least not when taken in aggregate), and so it is likely that alternative causal factors would relate to NPG editorial policy and practice. While we are not aware of any other relevant changes in editorial policy occurring at a relevant time, it is likely that this change in editorial policy was accompanied by increased attention given to the importance of the quality of reporting by both in house editorial staff and external peer reviewers. It is not possible to determine whether these might have caused the changes seen. However, a randomised controlled study of the effect of ARRIVE checklist completion on the quality of reporting of in vivo research at PLoS One will report shortly.

During the course of the study we encountered some difficulties that we had not expected. We had thought that it would be straightforward to distinguish between an in vivo experiment and an in vitro experiment, but we had to develop an operational approach which defined that experiment on the basis of the subject at the time that the experimental intervention occurred; so a tissue slice experiment involving tissues from animals exposed to treatment or control we considered in vivo; while a similar experiment applying drugs directly to the slice we considered to be an in vitro experiment.

Further, there were some checklist items where agreement between outcome assessors was very low – for instance, for the question of whether for in vivo research the difference in variance between groups being compared was less than two fold, the agreement was no better than would be expected by chance alone. We

recommend that the development of publication checklists should include an assessment of inter-observer variation by potential users of the checklist for each checklist item; low agreement might indicate that the item should be rephrased or reframed, or that more explanatory text is required.

Finally, our work shows the challenge of assessing even a relatively limited number of publications against a relatively straightforward checklist. We are delighted that so many collaborators (from 6 continents) agreed to participate, and are very grateful to them. However, even with their help the outcome assessment and reconciliation took 17 months. This is too slow to be useful for instance for quality improvement activity, where more rapid feedback would allow more rapid adjustments in response to performance. We have tested the use of text analytics using regular expressions to automatically ascertain reporting of measures to reduce the risk of bias, and for some such risks of bias the approach achieves sensitivities and specificities above 80%. However, for more complex items it may be that machine learning approaches using for instance convoluted neural networks may be more successful, and this is a current focus of our research. We hope that, by making the dataset for this study available, this might be used for instance for distant supervised learning in such systems.

Conclusions

Introduction of a checklist lead to substantial improvements in the quality of reporting in NPG publications that was not seen in matched manuscripts from other publishers, and this improvement appears to be ongoing. However, there is still substantial room for improvement, and this suggests that measures such as mandatory author checklists need to be supplemented by other approaches.

Authorship: the NPQIP consortium

Study steering committee: Malcolm Macleod (University of Edinburgh, Chief Investigator and Chair), Emily Sena (University of Edinburgh), David Howells (University of Tasmania).

Study management committee: Malcolm Macleod (University of Edinburgh, Chief Investigator and Chair), Emily Sena (University of Edinburgh), David Howells (University of Tasmania), Veronique Kiermer (Nature, until mid 2015), Sowmya Swaminathan (Nature, from mid 2015).

Redaction and identification of publications: Hugh Ash, Rosie Moreland (Imperial College, London)

Authoring and testing of training materials: Cadi Irvine, Paula Grill, Monica Dingwall, Emily Sena, Gillian Currie, Malcolm Macleod (University of Edinburgh)

Programming and data management: Jing Liao, Chris Sena (University of Edinburgh)

Outcome assessors: Paula Grill (272), Monica Dingwall (258), Malcolm Macleod (229), Cadi Irvine (179), Cilene Lino de Oliveira (170), Daniel-Cosmin Marcu (113), Fala Cramond(96), Sulail Rajani (93), Andrew Ying (81), Hanna Vesterinen (31), Roncon Paolo (28), Kaitlyn Hair (26), Marie Soukupova (23), Devon C. Crawford (17), Kimberley Wever (16), Mahajabeen Khatib (16), Ana Antonic (13), Thomas Ottavi (13), Xenios Milidonis (12), Klara Zsofia Gerlei (10), Thomas Barrett (10), Ye Liu (10), Chris Choi (9), Evandro Araújo De-Souza (8), Alexandra Bannach-Brown (8), Peter-Paul Zwetsloot (5), Kasper Jacobsen Kyng (5), Sarah McCann (4), Emily Wheater (4), Aaron Lawson McLean (1), Marco Casscella (1), Alice Carter (1), Privjyot Jheeta (1), Emma Eaton (1).

Reconciliation: Alexandra Bannach-Brown (199), Malcolm Macleod (197), Monica Dingwall (167), Paula Grill (161), Kaitlyn Hair (97), Cilene Lino de Oliveira (40), Sulail Rajani (9), Daniel-Cosmin Marcu (8), Cadi Irvine (3), Fala Cramond (1).

Data analysis: Paula Grill, Jing Liao, Malcolm Macleod

Writing Committee: Malcolm Macleod, David Howells, Jing Liao, Paul Grill, Emily Sena

Disclaimer: the opinions expressed in this article are the authors' own and do not reflect the view of any employing agency including the U.S. National Institutes of Health, the U.S. Department of Health and Human Services, or the United States Government."

Funding

The study was funded by a grant from the Laura and John Arnold Foundation, who played no role in the design, conduct or analysis of the study or in decisions regarding publication or dissemination.

Role of Nature in data analysis and data ownership

The study dataset belongs to the investigators, and all decisions relating to data analysis and publication were be taken by the steering committee and were independent of Nature. NPG were invited to correct any errors of fact in a draft version of the manuscript.

Reference List

(1) Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. Improving Bioscience Research Reporting: The ARRIVE Guidelines for Reporting Animal Research. PLoS Biol 2010 Jun 29;8(6):e1000412.

(2) Landis SC, Amara SG, Asadullah K, Austin CP, Blumenstein R, Bradley EW, et al. A call for transparent reporting to optimize the predictive value of preclinical research. Nature 2012 Oct 11;490(7419):187-91.

(3) Baker D, Lidster K, Sottomayor A, Amor S. Two years later: journals are not yet enforcing the ARRIVE guidelines on reporting standards for pre-clinical animal studies. PLoS Biol 2014 Jan;12(1):e1001756.

(4) Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. Nature 2012 Mar 28;483(7391):531-3.

(5) Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? Nat Rev Drug Discov 2011 Aug 31;10(9):712-c1.

(6) Anon. Announcement: Reducing our Irreproducibility. Nature 496[7446], 398. 24-4-2013.

(7) Cramond F, Irvine C, Liao J, Howells D, Sena E, Currie G, et al. Protocol for a retrospective, controlled cohort study of the impact of a change in Nature journals' editorial policy for life sciences research on the completeness of reporting study design and execution. Scientometrics 2016;108:315-28.

(8) Macleod MR, Lawson MA, Kyriakopoulou A, Serghiou S, de WA, Sherratt N, et al. Risk of Bias in Reports of In Vivo Research: A Focus for Improvement. PLoS Biol 2015 Oct 13;13(10):e1002273.

(9) Ramirez FD, Motazedian P, Jung RG, Di SP, MacDonald ZD, Moreland R, et al. Methodological Rigor in Preclinical Cardiovascular Studies: Targets to Enhance Reproducibility and Promote Research Translation. Circ Res 2017 Jun 9;120(12):1916-26.

(10) Macleod MR, Fisher M, O'Collins V, Sena ES, Dirnagl U, Bath PM, et al. Good laboratory practice: preventing introduction of bias at the bench. Stroke 2009 Mar;40(3):e50-e52.

Table and figure legends

Figure 1: Manuscripts initially included, and reasons for exclusion, and type of experiments described.

Figure 2: Compliance with each Landis criteria for in vivo experiments for NPG and non NPG manuscripts before and after 1st May 2013.

Figure 3: Interrupted time series analysis for overall Landis compliance and compliance with Landis components in in vivo experiments reported in NPG and non NPG manuscripts. Quarter 6 began on 1st May 2013.

Figure 4: Radar plots for compliance with individual components of the NPG checklist before (red) and after (green) 1st May 2013 for (a) statistical reporting, in vivo research; (b) statistical reporting, in vitro research; (c) reporting of details of animals used; and (d) reporting of reagents used. * adjusted $p < 0.05$ for change between “before” and “after”.

Table 1: Sources of manuscripts included in the study

Table 2: Primary outcome: Compliance with Landis 4 guidelines, in vivo research: n, number meeting criteria: N, total number of studies: %, percent meeting criteria: CI, 95% confidence interval of that percentage: p, significance level (two sample proportion test): n.s., not significant at $p<0.05$.

Table 3: Secondary outcome: Full Landis compliance, in vitro research: n, number meeting criteria: N, total number of studies: %, percent meeting criteria: CI, 95% confidence interval of that percentage: Adj p, adjusted significance level (two sample proportion test (prop.test) followed by Holm Bonferroni correction (p.adjust.methods)): n.s., not significant at $p<0.05$.

Table 4: Compliance with individual Landis 4 items, in vivo and in vitro research: n, number meeting criteria: N, total number of studies: %, percent meeting criteria: CI, 95% confidence interval of that percentage: Adj p, adjusted significance level (two sample proportion test (prop.test) followed by Holm Bonferroni correction (p.adjust.methods)): n.s., not significant at $p<0.05$: n.t. not tested ($n<10$ for one of the comparisons)

Table 5: Secondary outcome: statistical items, in vivo and in vitro experiments: n, number meeting criteria: N, total number of studies: %, percent meeting criteria: CI, 95% confidence interval of that percentage: Adj p, adjusted significance level (two sample proportion test (prop.test) followed by Holm Bonferroni correction (p.adjust.methods)): n.s., not significant at $p<0.05$.

Table 6: Other secondary outcomes: n, number meeting criteria: N, total number of studies: %, percent meeting criteria: CI, 95% confidence interval of that percentage: Adj p, adjusted significance level (two sample

proportion test (prop.test) followed by Holm Bonferroni correction (p.adjust.methods)): n.s., not significant at $p<0.05$: n.t. not tested ($n<10$ for one of the comparisons).

Did a change in Nature journals' editorial policy for life sciences research improve reporting?

Authors

The NPQIP Collaborative group

Corresponding author

Professor Malcolm Macleod: Malcolm.Macleod@ed.ac.uk, 07786 265166

Conflicts of interest

None declared.

Keywords

Risk of bias, reporting, methodological quality, study design, reporting guidelines

Acknowledgements

Funding

Laura and John Arnold Foundation.

Abstract

Objective: To determine whether a change in editorial policy, including the implementation of a checklist, has been associated with improved reporting of measures which might reduce the risk of bias.

Methods: The study protocol has been published at DOI: 10.1007/s11192-016-1964-8.
Design: Observational cohort study
Population Articles describing research in the life sciences published in Nature journals, submitted after May 1st 2013.
Intervention Mandatory completion of a checklist at the point of manuscript revision.
Comparators (1) Articles describing research in the life sciences published in Nature journals, submitted before May 2013; (2) Similar articles in other journals matched for date and topic.
Primary Outcome Change in proportion of Nature articles describing in vivo research published before and after May 2013 reporting the "Landis 4" items (randomisation, blinding, sample size calculation, exclusions).

We included 448 Nature Publication Group (NPG) articles (223 published before May 2013, 225 after) identified by an individual hired by NPG for this specific task, working to a standard procedure; and an independent investigator used Pubmed "Related Citations" to identify 448 non- NPG articles with a similar topic and date of publication from other journals; and then redacted all articles for time sensitive information and journal name. Redacted articles were assessed by 2 trained reviewers against a 74 item checklist, with discrepancies resolved by a third.

Results: 394 NPG and 353 matching non-NPG articles described in vivo research. The number of NPG articles meeting all relevant Landis 4 criteria increased from 0/203 prior to May 2013 to 31/181 (16.4%) after (2-sample test for equality of proportions without continuity correction, $X^2 = 36.2$, df = 1, $p = 1.8 \times 10^{-9}$). There was no change in the proportion of non- NPG articles meeting all relevant Landis 4 criteria (1/164 before, 1/189 after). There were more substantial improvements in the individual prevalences of reporting of randomisation, blinding, exclusions and sample size calculations for in vivo experiments, and less substantial improvements for in vitro experiments.

Conclusions. There was an improvement in the reporting of risks of bias in in vivo research in NPG journals following a change in editorial policy, to a level that to our knowledge has not been previously observed. However, there remain opportunities for further improvement.

Background

Few articles describing *in vivo* research report taking specific actions designed to reduce the risk that their findings are confounded by bias[1][1], and those that do not report such actions give inflated estimates of biological effects[2 3][2,3]. Strategies and guidelines which might improve the quality of reports of *in vivo* research have been proposed, [4 5][4,5] and while these have been endorsed by a large number of journals there is evidence that this endorsement has not been matched by a substantial increase in the quality of published reports [6][6].

Poor replication of *in vitro* molecular and cellular biology studies has also been reported[7-9]{Scott, 2008 #427} [11][7,8] and this has been attributed in part to poor descriptions of the experimental and analytical details.

In May 2013 Nature Journals announced a change in editorial policy which required authors of submissions in the life sciences to complete a checklist, at the time of manuscript acceptance, indicating whether or not they had taken certain measures which might reduce the risk of bias and to report key experimental and analytical details; and in their submission to detail where in the manuscript these issues were addressed [10][9]. The development of this checklist was prompted in part by a consensus statement[5] [5] setting out key aspects of study design and conduct which were necessary to allow the reader to assess the validity of the findings presented; it identified these as randomisation; blinding; sample size estimation and data handling (the "Landis 4"). The Nature Journals' checklist also included items relating to figures and statistical representation of data; reagents used; species, strain, and sex of experimental animals, reporting of relevant ethical approvals; consent (for research involving human subjects); data deposition; and availability of any bespoke computer code. The full checklist is given in Appendix 1.

The aim of this study was to determine whether the implementation of this checklist for submissions has been associated with improved reporting of measures that might reduce the risk of bias. To establish whether any observed change in quality was a simply a secular trend occurring across all journals we matched each included publication with a publication in a similar subject area published at around the same time by a different publisher.

Methods

The methods are described in detail in the published study protocol [11][10], and the data analysis plan and analysis code were articulated prior to database lock and registered on the Open Science Framework ((DOI 10.17605/OSF.IO/HC7FK).). The complete study dataset including PMIDs and data descriptors (but not, for copyright reasons, the source pdfs) of included articles is available on Figshare (10.6084/m9.figshare.6226718).

In this observational cohort study we aimed to determine whether the implementation of a checklist for submissions has been associated with improved reporting of measures which might reduce the risk of bias. The study populations comprised (1) Published articles accepted for publication in Nature journals, which described research in the life sciences and which were submitted after May 1st 2013 (when the mandatory

completion of a checklist at the stage of manuscript revision was introduced) and before November 1st 2014; (2) Published articles accepted for publication in Nature journals in the months preceding May 2013, which describe research in the life sciences; and (3) articles from other journals matched for subject area and time of publication. We measured the change in the reporting of items included in the checklist.

Identification of relevant articles

We included studies which described in vivo (articles that contain at least one non-human animal experiment, including rodents, flies, worms, zebrafish etc.) or in vitro research

NPG articles: One individual was specifically employed by Nature Publishing Group to select studies which (a) described in vivo or in vitro research; and (b) was published in Nature, Nature Neurology, Nature Immunology, Nature Cell Biology, Nature Chemical Biology, Nature Biotechnology, Nature Methods, Nature Medicine or Nature Structural and Molecular Biology. First, the individual identified papers accepted for publication with an initial submission date later than May 1st, 2013. Beginning with the then-current issues (volume corresponding to year 2015), they worked backwards in time, ensuring the submission date was after 1st May 2013, collecting papers with the intention of identifying 40 Nature papers and 20 from each of the other 8 titles (i.e. 200 papers in total) ("Post intervention" group). They then used a similar process to identify papers submitted for publication before 1st May 2013, matched for journal and for country of origin (based on the address of the corresponding author), starting with the May 2013 issue and working backwards, ensuring that the date of submission was after 1st May 2011 ("pre-intervention" group). Where no match could be found with a submission date after 1st May 2011 (i.e. in a two year period) then the non-matched post intervention publication was excluded from analysis and a replacement post intervention publication selected, as above. A matching pre-intervention publication then identified, as described above. Articles describing research involving only human subjects were excluded. A Nature editorial administrator independent of publishing decisions reviewed articles selected against the inclusion criteria and found some (less than 10%) had been included incorrectly; they replaced these with manuscript pairs that they selected according to the inclusion algorithm described above. The published files corresponding to the publication pdfs (including the extended methods section, extended data and other supplementary materials) were used to generate pdfs for analysis. These were provided to a member of our research team (RM) at a different institution who used Adobe Acrobat to redact information relating to author names or affiliations, dates, volumes or page numbers; and the reference list; to minimise awareness of outcome assessors to whether the manuscript was pre- or post-intervention.

Non- NPG articles: The same member of our research team (RM) was responsible for identifying matching articles in other journals. Using PubMed, she entered the Nature Publishing Group publication title to retrieve the relevant record. She then added the "related citations for PubMed" result to the search builder. In the second line search field of the search builder she searched for "Date of publication" in the same calendar month and year, and performed the search. In the results returned she started with the first result returned and established whether it was published in a Nature Publication Group Journal (given in Appendix 2). If it was not, she applied the study inclusion criteria (in vivo or in vitro research or both, as defined above), ensuring that there was a match on the in vivo/in vitro status between the index Nature Publishing Group publication and the non-Nature Publishing Group publication. Where these criteria were met she selected the

publication for the study and retrieved the pdf, through open access, online institutional subscription, interlibrary loan, or by request from the authors. If the first related citation did not fulfil these criteria, she moved to the next, until an appropriate publication was found. If an appropriate publication was not found, she repeated these steps but with the date of publication used in the search extended by 1 month earlier and 1 month later. If this process did not identify an eligible publication, she again extended the search by a month in each direction, and continued until a matching publication was found. She then recorded the difference in calendar months between the date of publication of the index NPG article and the date of publication of the matching non-NPG article. Because of a limited number of potential matching articles it was not possible to match non NPG articles by country.

She then used Adobe Acrobat to redact information relating to author names or affiliations, dates, volumes or page numbers; and the reference list; to minimise awareness of outcome assessors to whether the manuscript was pre- or post- intervention.

The individual making this selection and redacting information from articles paid no further part in the study. In total 896 articles were selected for analysis.

Outcome assessment

The Nature checklist focussed on transparency in reporting and availability of materials and code, reflected in 10 items. We designed a series of questions (Appendix 1) to establish whether a given publication met or did not meet the requirements of the checklist. We did this to aid outcome assessors, because many checklist items included more than one embedded criteria. For instance, the section on “Figures and Statistical Representation of Data” was operationalised to 12 individual “present/absent/not applicable” responses. Where a manuscript described both in vivo and in vitro research, the series of questions was completed for each. Where there was more than one in vitro experiment or more than one in vivo experiment the question was considered in aggregate; that is, all in vitro experiments had to meet the requirements of the checklist item for the article to be considered compliant in reporting of in vitro experiments, and all in vivo experiments had to meet the requirements of the checklist item for the article to be considered compliant in reporting of in vivo experiments. [12]

Five researchers experienced in systematic review and risk of bias annotation scored a set of 10 articles using our series of questions. Disagreements were resolved by group discussion, to arrive at a set of “Gold standard” answers for these 10 articles. We also used this experience to write a training guide for outcome assessors. We then used social media platforms and mailing lists to recruit outcome assessors. We had no prior requirements for the skills required of these individuals, but most had a background in medicine or biomedicine at graduate or undergraduate level; two were senior school students on Nuffield Research Placements in our group. After they had reviewed the training materials, outcome assessors were invited to score articles from the “Gold standard” pool, presented in random order, until their concordance with the Gold standard responses was 80% overall, and was 100% for the components of the primary outcome measure, for three successive articles. At this point we considered them to be trained. The training platform remains available for continuing professional development, at .

Pdf files of included articles were uploaded to the study website. Trained assessors were presented with articles for scoring in random order. Each manuscript was scored by 2 individuals, one with experience in systematic review and risks of bias annotation and one recruited from outside this community. Disagreement between assessors were reconciled by a third, experienced individual who was not one of the original reviewers, who could see the responses previously given but not who the initial reviewers were.

Statistical analysis plan.

Given our focus on the reporting of measures to reduce the risks of bias we took as our primary outcome measure a composite measure of the proportion of articles meeting the relevant measures identified by Landis et al in 2012 as being most important for transparency in reporting in vivo research. These are covered by items 2, 3 4 and 5 of the checklist and relate to the reporting of randomisation; of the blinded assessment of outcome; of sample size calculations; and of whether the manuscript described whether samples or animals were excluded from analysis. Importantly, checklist compliance did not require for example that the study was randomised; but rather that the authors stated whether or not it was randomised. The evaluation principle was to determine if someone with reasonable domain-knowledge could understand the parameters of experimental design sufficiently to inform interpretation. It has been argued that these measures might not be as relevant for exploratory studies, and for these we recorded the item as “not relevant”. We defined exploratory studies as those where hypothesis testing inferential statistical analyses were not reported. Where an item was not relevant for a publication (for instance with studies using transgenic animals where group allocation had been achieved by Mendelian randomisation) we considered compliance as meeting all of the relevant criteria. Where a publication described both in vivo and in vitro experiments we analysed each type of experiment separately.

Our primary outcome was the change in the proportion of articles describing in vivo experiments published by NPG before and after May 2013 that meet all of the relevant Landis 4 criteria. This is described in the statistical analysis plan deposited on the Open Science Framework (osf.io/hc7fk) on 7th June 2017 prior to database lock and before we had derived any outcome information. It differs from the primary outcome measure described in the original published protocol [11][11] (where it was the first secondary outcome). We did this because we realised during preparation of the data analysis plan that what was most important was not the achievement of some arbitrary level of performance, but rather whether the intervention was associated with an important improvement in performance.

We used the two-sample proportion test (prop.test) in R without the Yates continuity correction and two sided hypothesis testing to be sensitive to the possibility that performance might have declined rather than improved. Secondary outcomes were whether the proportion of articles describing in vivo experiments published by NPG after May 2013 which meet all four of the Landis 4 criteria was 80% or higher (the original primary outcome; Wald test; wald.ptheor.test, RVAideMemoire in R); the change in the proportion of articles describing in vitro experiments published by NPG before and after May 2013 which meet all four of the Landis 4 criteria (two sample proportion test as above). Further secondary outcomes were the change of proportions in adequate reporting of statistical analysis details, individual Landis criteria, and descriptions of animals;

reagents and their availability; sequence, structure or computer code deposition; and items relating to the involvement of human subjects or materials in included studies.. For the matching articles from non-NPG journals the secondary outcomes were the change in the proportion of articles describing in vivo experiments published before and after May 2013 which met all of the Landis 4 criteria (two sample proportion test); whether the proportion of articles describing in vivo experiments published after May 2013 which met all four of the Landis 4 criteria was 80% or higher (Wald test); the change in the proportion of articles describing in vitro experiments published before and after May 2013 which meet all four of the Landis 4 criteria 4 (two sample proportion test); and the change of proportions in adequate reporting of statistical analysis details, individual Landis criteria, and descriptions of animals; reagents and their availability; sequence, structure or computer code deposition; and items relating to the involvement of human subjects or materials in included studies. For each of these outcomes we compared the changes observed in NPG articles with that observed in non NPG articles. For each secondary analysis we used Holm Bonferroni correction using the p.adjust option for prop.test in R to account for the number of comparisons drawn, as described in Appendix B of the Data Analysis Plan. We also used interrupted time series analysis for each checklist item in an attempt to distinguish a discrete “shift” in performance from an upward “drift”, as described in the data analysis plan. A number of tertiary outcomes are described in the study protocol and statistical analysis plan and are reported in the supplementary material.

Power Calculations

Power calculations were performed in STATA prior to commencement of the study. For the primary outcome measure we approximated required sample sizes using power calculations for a one sided two sample Chi squared test in STATA seeking a significance level of $p<0.01$ and with varying estimates of compliance with the Landis 4 criteria in the pre-intervention group. With 200 articles in each group we had 80% power to detect an increase from 10% to 21%, or from 20% to 34%, or from 30% to 45%, or from 40% to 56%, or from 50% to 66%. We wanted to detect an absolute difference of 10% or more, and thought that compliance with the Landis 4 criteria in the pre-intervention group would be around 10%, so thought that having 200 studies in each group would be sufficient.

For the primary outcome measure proposed in the original study protocol (that compliance with the Landis 4 criteria in the post-intervention group reached 80%), 200 studies in each group would be sufficient to reject the alternative hypothesis if the observed compliance was 72% or lower and again, we considered this to be sufficient.

For individual checklist items, after correcting for multiple comparisons, statistical power again depends on the level of reporting in the pre intervention group. Where this was between 15% and 85%, with 200 studies per group we would have 80% power to detect an absolute increase of 15% in the reporting of each item. We considered this to be the minimal increase that would represent an important improvement in reporting. The power calculations are described in greater detail in the study protocol[11] [11].

Results

896 articles were identified and uploaded for outcome ascertainment, 448 in each cohort. 2 non-NPG articles were excluded because they did not meet the inclusion criteria, and we identified 4 NPG and 9 non-NPG

articles which had been included more than once. 444 NPG articles and 437 non-NPG articles underwent outcome assessment. One NPG publication and one non-NPG publication were adjudged at the time of outcome assessment to report neither in vivo nor in vitro research and so were excluded. The analysis is therefore based on 443 NPG articles (219 before and 224 after 1st May 2013) and 436 non-NPG articles (194 before and 242 after 1st May 2013) (Figure 1). The difference in numbers for NPG and non-NPG before and after 1st May 2013 is because some of the NPG "before" articles matched best with articles in other journals published in the few months following May 2013. Specifically, 26 NPG pre-intervention articles were matched with other papers published an average of 3.2 months after May 2013 (max 8 months), and 6 NPG post-intervention articles were matched with other papers published 1,2,9,11,12 and 215 months before May 2013. Overall, 43% of matched pairs had dates of publication within 1 month, 54% within 2 months, 64% within 3 months and 81% within 6 months of each other (range -11 to +22 months). 239 articles described only in vivo research, 133 described only in vitro research, and 507 described both. 494 papers were completely matched for in vivo and in vitro status, 276 were partially matched (one member of matched pair reporting in vivo and in vitro research, the other reporting only in vitro or only in vivo research), and 36 were mismatched (one reporting only in vivo research, the other reporting only in vitro research). The source journals are given in Table 1; in total 198 different titles contributed matching articles (median of 1 article per source journal, range 1 – 47). The PMIDs of included articles are listed in the data supplement.

205 individuals registered with the project, of whom 38 completed their training and 35 assessed at least one manuscript. 12 also served as reconcilers, and the web interface was programmed to ensure that they were not offered for reconciliation articles that they had previously adjudicated. Including reconciliation, the median number of articles scored was 13 (range 1 to 441). The agreement between the initial pair of outcome assessors ranged from being no better than chance at 50% (in vivo studies, Implementation of statistical methods and measures: "Is the variance similar (difference less than two-fold) between the groups that are being statistically compared?") to 98% (in vivo studies, "Does the study report the species?"). Median agreement was 82%. (IQR 68 – 89%). Two articles were identified during manuscript preparation as having been incorrectly recorded at datalock as reporting both in vivo and in vitro research, where in fact they only reported in vitro research, and one article had been incorrectly recorded as reporting both in vivo and in vitro research, where in fact it only reported in vivo research,.

Reporting of the Landis 4 items: The proportion of NPG in vivo studies reaching full compliance with the Landis 4 criteria increased from 0% (0/203) to 16.4% (31/189) ($X^2 = 36.1$, df = 1, $p = 1.8 \times 10^{-9}$), but remained significantly lower than the target of 80% (95% CI 11.6% to 22.6%, Wald test v 80% z = -15.4, $p = 2.2 \times 10^{-16}$).

For randomisation to experimental group, the preferred standard is that the manuscript describes which method of randomization was used to determine how samples or animals were allocated to experimental groups, although articles were also compliant if they included a statement about randomization even if no randomization was used. The proportion of NPG in vivo studies reporting randomisation was 1.8% (3/170, 95% CI 0.6 to 5.3%) before and 11.2% (19/170, 95% CI 7.2 to 16.9%) after ($\chi^2 = 12.4$, df = 1, adj p = 0.054). The proportion of studies mentioning randomisation even where it was not reported increased from 8.3%

(14/169, 95% CI 5.0 to 13.5) to 64.2% (97/151, 95% CI 56.3 to 71.5%)(χ^2 = 110.2, df = 1, adj p = 3.2×10^{-14}). Figure 2(a) shows change in the proportion of studies meeting these criteria before and after the change in editorial policy.

For blinding, the preferred standard is that the manuscript describes whether the investigator was blinded to the group allocation during the experiment and/or when assessing the outcome, although articles were also compliant if they included a statement about blinding even if no blinding was done. The proportion of NPG in vivo studies reporting blinding during group allocation or outcome assessment or both increased from 4% (8/198, 95% CI 2.0 to 7.9%) to 22.8% (42/184, 95% CI 17.3 to 29.4%)(X^2 = 29.6, df = 1, adj p = 7.6×10^{-6}). The proportion of studies mentioning blinding even where it was not reported increased from 1.6% (3/182, 95% CI 0.5 to 5.0%) to 55.3% (73/132, 95% CI 46.8 to 65.6%)(X^2 = 120.1, df = 1, adj p < 3.2×10^{-14}). Figure 2(b) shows change in the proportion of studies meeting these criteria before and after the change in editorial policy.

The proportion of studies reporting animals excluded from analysis increased from 13.9% (28/202, 95% CI 9.7 to 19.3%) to 30.7% (58/189, 95% CI 24.5 to 36.7%)(X^2 = 16.1, df = 1, adj p = 0.008). Figure 2(c) shows change in the proportion of studies meeting these criteria before and after the change in editorial policy.

For sample size calculations, the preferred standard is that the manuscript describes how the sample size was chosen to ensure adequate power to detect a pre-specified effect size, although articles were also compliant if they included a statement about sample size estimate even if no statistical methods were used. The proportion of studies reporting an a priori sample size calculation increased from 2.0% (4/196, 95% CI 0.8 to 5.3%) to 14.8% (27/182, 95% CI 10.4 to 20.8%)(X^2 = 20.5, df = 1, adj p = 0.0008). The proportion of studies mentioning sample size even where a sample size calculation was not reported increased from 1.6% (3/192, 95% CI 0.5 to 4.7%) to 58.4% (90/154, 95% CI 50.5 to 66.0%)(X^2 = 140.7, df = 1, adj p < 3.2×10^{-14}). Figure 2(d) shows change in the proportion of studies meeting these criteria before and after the change in editorial policy.

For NPG in vitro studies, the proportion reaching full compliance with the Landis 4 criteria was 0% (0/159) before and 3.3% (6/176) after (X^2 = 6.8, df = 1, Holm Bonferroni adjusted p = 1.00). The proportion of studies reporting randomisation was 0% (0/149) before and 2.9% (5/173, 95% CI 1.2 to 6.8%) after (X^2 = 4.4, df = 1, adj p=1.00). The proportion of studies mentioning randomisation even where it was not reported increased from 0% (0/149) to 15.6% (97/151, 95% CI 10.8 to 21.9%)(X^2 = 25.3, df = 1, p = 6.9×10^{-5}). The proportion of studies reporting blinding during group allocation or outcome assessment or both was 3.9% (6/155, 95% CI 1.8 to 8.4%) before and 8.9% (16/179, 95% CI 5.6 to 14.1) after (X^2 = 3.467, df = 1, p=1.00). The proportion of studies mentioning blinding even where it was not reported increased from 0.7% (1/150, 95% CI 0.1 to 4.6%) to 15.9% (25/157, 95% CI 11.0 to 22.5) (X^2 = 23.0, df = 1, p=0.0002). The proportion of studies reporting exclusions from analysis was 8.2% before (13/159, 95% CI 4.8 to 13.6%) and 15.9% (29/182, 95% CI 11.3 to 22.0%) after (X^2 = 4.73, df = 1, p =1.00). The proportion of studies reporting an a priori sample size calculation was 1.3% (2/155, 95% CI 0.3 to 5.0%) before and 7.9% (14/177, 95% CI 5.1 to 13.5%) after (X^2 = 8.7106, df = 1, p = 1.00). The proportion of studies mentioning sample size even where a sample size calculation was not

reported increased from 3.3% (5/153, 95% CI 1.4 to 7.6%) to 28.5% (47/165, 95% CI 22.1 to 35.8%)(X^2=36.9, df = 1, $p=1.8x10^{-7}$).

The proportion of matching (non-NPG) in vivo studies reaching full compliance with the Landis 4 criteria was 1% (1/164) before and 1% (1/189) after (X^2 = 0.01, df = 1, adj p = 1.00), and for in vitro studies, the proportion of non-NPG studies reaching full compliance with the Landis 4 criteria was 0% (0/133) before and 1% (1/165) after (X^2 = 0.8, df = 1, adj p = 1.00). The prevalence of reporting the different items before and after is shown in table 2; there was no significant change in reporting of any of the individual Landis 4 criteria for either in vivo or in vitro research.

Statistical reporting: For in vivo studies reported in NPG articles there were significant improvements in the reporting of exact numbers (from 46% to 69%), of whether t-tests were defined as one or two sided (from 46% to 71%), and whether the assumptions of the test had been checked (from 9% to 27%). For in vitro experiments described in NPG articles there were significant improvements in the reporting of the exact numbers (from 32% to 70%); of whether data represented technical or biological replicates (from 57% to 75%); and whether t-tests were defined as one or two sided (from 47% to 72%). For in vivo and in vitro studies described in non-NPG articles there was no significant change in any of the items relating to statistical reporting.

Other checklist items: For reporting of details of animals used, reporting of animal species and strain was high even before the change in editorial policy. There was no significant change in reporting any of these items in NPG- and non-NPG articles, or in the reporting of details of antibodies used. For in vitro research, there was an increase in the proportion of studies in NPG articles reporting recent mycoplasma testing of the cell lines used (from 1% to 26%) but not for non-NPG articles (1% before, 1% after). For reporting and availability of accession data (eg DNA or protein sequence deposition) and computer code there were no significant changes for either NPG or non-NPG articles. Finally, there were no significant changes in the reporting of items relating to human subjects or the use of human materials, but for most items the number of articles for which these were relevant was very low indeed.

We were also interested in whether changes in reporting had occurred as a step change at the time of the change in editorial policy; whether there was an initial improvement with then a return to previous performance; or if there was an ongoing improvement in reporting. To address these we conducted an interrupted time series analysis, to estimate the rate of change before the intervention; any step change at the time of the intervention; and the rate of change after the intervention. We grouped articles in 3 month periods starting November 2011, and for each quarter calculated the proportional compliance with the criteria in question. Because articles were not evenly distributed across time the analysis is of substantially reduced power, but the fitted lines for overall compliance and for each component of the Landis checklist for in vivo research are shown in Figure 3. It appears that with the exception of sample size calculation there is a continuing improvement over time in both NPG and non NPG articles; for sample size calculations the

improvement is only seen in NPG articles. Figure 4 shows radar charts of compliance for each checklist item in NPG and non NPG articles before and after May 2013.

Discussion

The change in editorial policy at NPG was associated with improvements in reporting of randomisation, blinding, exclusions from analysis and sample size calculations. For the highly challenging primary outcome measure, full compliance increased from zero to 16%. This falls short of the target compliance of 80%, but should be seen in the context firstly that only 1 of 1073 articles from 2009-10 from leading UK institutions achieved this standard[1][1]; and secondly that overall compliance of 80% would require compliance with individual items of around 95%.

It is notable that even with considerable investment in designing and implementing a checklist, and working with authors to encourage its completion, that compliance remains so low. This stands rather in contrast to the belief that "all" that is required to ensure transparency in reporting is that journals "insist" that authors do the right thing. Securing transparency in research reports is a complex challenge, and experience in other fields (MM is also clinical lead for a clinical Neurology service) suggests such challenges require a range of complementary approaches with commitment from all stakeholders, might best be achieved through formal improvement activity, and often take multiple attempts to achieve and sustain change.

The checklist relates to transparency in reporting, and articles were judged to be compliant if they either reported measures to address that risk of bias, or reported that such measures were not taken. For reports of in vivo research, compliance for randomisation, blinding, reporting of exclusions and sample size calculations in NPG articles reached 68%, 62%, 31% and 64% respectively. For non NPG articles the performance was 12%, 5%, 12% and 3%. The figures for NPG articles are similar to those recently reported for in vivo research published in the journal "Stroke"[13 14] [12], which began requiring reporting of such details following the publication of good practice guidelines in 2009[15] [13]; and where performance was found to be substantially higher than for in vivo research published in other American Heart Association journals[14].

While we saw improvements in the transparency of reporting, the observed improvements in experimental design were much more modest. However, peer review may not ensure the quality of published work[16] [14], as evidenced for in vivo research by poor reporting of measures to reduce risks of bias[1] [1]. We believe that the ultimate responsibility for assessing research quality (and therefore the validity of the findings presented) rests with the reader, and transparency in reporting is fundamental to this assessment.

For reports of in vitro research, compliance was substantially lower. There have been few systematic attempts to measure the quality of reporting of measures to reduce the risks of bias in vitro research, and our findings suggest that, both in NPG and non NPG journals, this remains low. There were improvements in reporting randomisation, blinding and sample size calculations in NPG descriptions of in vitro research, but only to 18%, 23% and 34% respectively. For non NPG the equivalent figures were 3%, 1% and 1%. There were no

significant changes in the reporting of exclusion of in vitro data, with post intervention compliance of 16% in NPG articles and 6% in non NPG articles.

For other checklist items, changes in performance were less dramatic, but there appeared to be incremental improvements across most of the items measured, although few of these breached our rather parsimonious adjustment for multiple testing. In spite of substantial attention given to the importance of reporting the sex of experimental animals this was only done in 52% of post intervention NPG studies and in 36% of non NPG studies.
17

Ours is an observational study, and it is possible that other (related or unrelated) changes were responsible for much if not all of the differences seen. These changes were not observed at other journals (at least not when taken in aggregate), and so it is likely that alternative causal factors would relate to NPG editorial policy and practice. While we are not aware of any other relevant changes in editorial policy occurring at a relevant time, it is likely that this change in editorial policy was accompanied by increased attention given to the importance of the quality of reporting by both in house editorial staff and external peer reviewers. It is not possible to determine whether these might have caused the changes seen. However, a randomised controlled study of the effect of ARRIVE checklist completion on the quality of reporting of in vivo research at PLoS One will report shortly.

When writing our data analysis plan, and prior to any data inspection or analysis, we substituted our “first” secondary outcome (the change in proportion of articles describing in vivo research meeting the 4 Landis criteria) for our original (in the published study protocol) primary outcome (whether compliance in the post intervention group of articles reached 80%). This was because our primary intention had been to observe any effect of a change in publication policy, and, with the benefit of hindsight, this was not captured in our original primary outcome, but we recognise this as a limitation in our findings. We note, however, that the primary outcome used reflects better the title of the study protocol than does the primary outcome measure proposed in that protocol.

For our comparator group we chose similar articles with a similar date of publication identified using the PubMed “related citations” tool. The journals in which these works were published will vary in the attention which they have given to transparency in reporting, and it may be that for some journals there have been changes similar to those observed in the Nature Publishing Group articles. While we might have restricted our comparator group to journals more similar to NPG articles (for instance by Impact Factor, or extent of editorial intervention) this would have meant lower fidelity of matching by subject area or date of publication or both, and we considered these factors to be more important. For this reason, our findings for NPG articles cannot be interpreted as showing improved reporting compared with similar articles in similar journals. The representation of such “similar” journals in the comparator group is too small to allow meaningful conclusions to be drawn.

During the course of the study we encountered some difficulties that we had not expected. We had thought that it would be straightforward to distinguish between an in vivo experiment and an in vitro experiment, but we had to develop an operational approach which defined that experiment on the basis of the subject at the time that the experimental intervention occurred; so a tissue slice experiment involving tissues from animals exposed to treatment or control we considered in vivo; while a similar experiment applying drugs directly to the slice we considered to be an in vitro experiment.

Our matching on whether studies reported in vitro or in vivo research, or both, was also reasonable in the majority of cases. Differences will have emerged where, as described above, articles were initially categorised with one set of characteristics (in vitro, in vivo or both) and matched accordingly, but later judged to have different characteristics. Our matching for date of publication worked reasonably well, with the exception of the inclusion of one comparator article published in 1995, 215 months before its “matching” NPG article. We had not anticipated that matching articles be so difficult to identify, so our matching rules did not have an upper limit of difference in date of publication. Since the comparator group do not contribute to our primary outcome, and the matching is generally good, we do not think that these mismatches devalue our findings to any appreciable extent.

Further, there were some checklist items where agreement between outcome assessors was very low – for instance, for the question of whether for in vivo research the difference in variance between groups being compared was less than two fold, the agreement was no better than would be expected by chance alone. We recommend that the future development of publication checklists should include an assessment of inter-observer variation by potential users of the checklist for each checklist item; low agreement might indicate that the item should be rephrased or reframed, or that more explanatory text is required.

Finally, our work shows the challenge of assessing even a relatively limited number of articles against a relatively straightforward checklist. We are delighted that so many collaborators (from 6 continents) agreed to participate, and are very grateful to them. However, even with their help the outcome assessment and reconciliation took 17 months. This is too slow to be useful for instance for quality improvement activity, where more rapid feedback would allow more rapid adjustments in response to performance. We have tested the use of text analytics using regular expressions to automatically ascertain reporting of measures to reduce the risk of bias, and for some such risks of bias the approach achieves sensitivities and specificities above 80%[18][15]. However, for more complex items it may be that machine learning approaches using for instance convoluted neural networks may be more successful, and this is a current focus of our research. We hope that, by making the dataset for this study available, this might be used for instance for distant supervised learning in such systems.

Conclusions

Introduction of a checklist lead to substantial improvements in the quality of reporting in NPG articles that was not seen in matched articles from other publishers, and this improvement appears to be ongoing. However,

there is still substantial room for improvement, which suggests that measures such as mandatory author checklists need to be supplemented by other approaches.

Data and code availability

The data are available at Figshare, along with the data dictionary (10.6084/m9.figshare.6226718). The analysis code is available at the open science framework (DOI 10.17605/OSF.IO/HC7FK).

Authorship: the NPQIP consortium

Study steering committee: Malcolm Macleod (University of Edinburgh, Chief Investigator and Chair), Emily Sena (University of Edinburgh), David Howells (University of Tasmania).

Study management committee: Malcolm Macleod (University of Edinburgh, Chief Investigator and Chair), Emily Sena (University of Edinburgh), David Howells (University of Tasmania), Veronique Kiermer (Nature, until mid 2015), Sowmya Swaminathan (Nature, from mid 2015).

Redaction and identification of articles: Hugh Ash, Rosie Moreland (Imperial College, London)

Authoring and testing of training materials: Cadi Irvine, Paula Grill, Monica Dingwall, Emily Sena, Gillian Currie, Malcolm Macleod (University of Edinburgh)

Programming and data management: Jing Liao, Chris Sena (University of Edinburgh)

Outcome assessors: Paula Grill (272), Monica Dingwall (258), Malcolm Macleod (229), Cadi Irvine (179), Cilene Lino de Oliveira (170), Daniel-Cosmin Marcu (113), Fala Cramond(96), Sulail Rajani (93), Andrew Ying (81), Hanna Vesterinen (31), Roncon Paolo (28), Kaitlyn Hair (26), Marie Soukupova (23), Devon C. Crawford (17), Kimberley Wever (16), Mahajabeen Khatib (16), Ana Antonic (13), Thomas Ottavi (13), Xenios Milidonis (12), Klara Zsofia Gerlei (10), Thomas Barrett (10), Ye Liu (10), Chris Choi (9), Evandro Araújo De-Souza (8), Alexandra Bannach-Brown (8), Peter-Paul Zwetsloot (5), Kasper Jacobsen Kyng (5), Sarah McCann (4), Emily Wheater (4), Aaron Lawson McLean (1), Marco Casscella (1), Alice Carter (1), Privjyot Jheeta (1), Emma Eaton (1).

Reconciliation: Alexandra Bannach-Brown (199), Malcolm Macleod (197), Monica Dingwall (167), Paula Grill (161), Kaitlyn Hair (97), Cilene Lino de Oliveira (40), Sulail Rajani (9), Daniel-Cosmin Marcu (8), Cadi Irvine (3), Fala Cramond (1).

Data analysis: Paula Grill, Jing Liao, Malcolm Macleod

Writing Committee: Malcolm Macleod, David Howells, Jing Liao, Paul Grill, Emily Sena

Disclaimer: the opinions expressed in this article are the authors' own and do not reflect the view of any employing agency including the U.S. National Institutes of Health, the U.S. Department of Health and Human Services, or the United States Government."

Funding

The study was funded by a grant from the Laura and John Arnold Foundation, who played no role in the design, conduct or analysis of the study or in decisions regarding publication or dissemination.

Role of Nature in data analysis and data ownership

The study dataset belongs to the investigators, and all decisions relating to data analysis and publication were be taken by the steering committee and were independent of Nature. NPG were invited to correct any errors of fact in a draft version of the manuscript.

Table and figure legends

Figure 1: Articles initially included, and reasons for exclusion, and type of experiments described.

Figure 2: Compliance with each Landis criteria for in vivo experiments for NPG (top two panels of each quartet) and non NPG articles (lower 2 panels) before and after 1st May 2013. A Randomisation; B Blinding; C Sample size calculation; and D Reporting of exclusions. For A-C Black represents studies where compliance was achieved by reporting that the measure was taken; green that compliance was achieved by describing that the measure was not taken; and white represents that compliance was not achieved. For D Black represents that the exclusions were reported, and white that exclusions were not reported.

Figure 3: Interrupted time series analysis for overall Landis compliance and compliance with Landis components in in vivo experiments reported in NPG and non NPG articles overall, and individually for randomisation, blinding, reporting of animals excluded from analysis and sample size calculations. .

Figure 4: Radar plots for compliance with individual components of the NPG checklist before (red) and after (green) 1st May 2013 for (a) statistical reporting, in vivo research; (b) statistical reporting, in vitro research; (c) reporting of details of animals used; and (d) reporting of reagents used. * adjusted $p < 0.05$ for change between “before” and “after”.

Table 1: Sources of articles included in the study

Table 2: Primary outcome: Compliance with Landis 4 guidelines, in vivo research: n, number meeting criteria: N, total number of studies: %, percent meeting criteria: CI, 95% confidence interval of that percentage: p, significance level (two sample proportion test): n.s., not significant at $p<0.05$.

Table 3: Secondary outcome: Full Landis compliance, in vitro research: n, number meeting criteria: N, total number of studies: %, percent meeting criteria: CI, 95% confidence interval of that percentage: Adj p, adjusted significance level (two sample proportion test (prop.test) followed by Holm Bonferroni correction (p.adjust.methods)): n.s., not significant at $p<0.05$.

Table 4: Compliance with individual Landis 4 items, in vivo and in vitro research: n, number meeting criteria: N, total number of studies: %, percent meeting criteria: CI, 95% confidence interval of that percentage: Adj p, adjusted significance level (two sample proportion test (prop.test) followed by Holm Bonferroni correction (p.adjust.methods)): n.s., not significant at $p<0.05$: n.t. not tested ($n<10$ for one of the comparisons)

Table 5: Secondary outcome: statistical items, in vivo and in vitro experiments: n, number meeting criteria: N, total number of studies: %, percent meeting criteria: CI, 95% confidence interval of that percentage: Adj p, adjusted significance level (two sample proportion test (prop.test) followed by Holm Bonferroni correction (p.adjust.methods)): n.s., not significant at $p<0.05$.

Table 6: Other secondary outcomes: n, number meeting criteria: N, total number of studies: %, percent meeting criteria: CI, 95% confidence interval of that percentage: Adj p, adjusted significance level (two sample proportion test (prop.test) followed by Holm Bonferroni correction (p.adjust.methods)): n.s., not significant at p<0.05: n.t. not tested (n<10 for one of the comparisons).

Reference List

(1) Macleod MR, Lawson MA, Kyriakopoulou A, Serghiou S, de WA, Sherratt N, et al. Risk of Bias in Reports of In Vivo Research: A Focus for Improvement. PLoS Biol 2015 Oct 13;13(10):e1002273.

(2) Macleod MR, van der Worp HB, Sena ES, Howells DW, Dirnagl U, Donnan GA. Evidence for the efficacy of NXY-059 in experimental focal cerebral ischaemia is confounded by study quality. Stroke 2008 Oct;39(10):2824-9.

(3) Rooke ED, Vesterinen HM, Sena ES, Egan KJ, Macleod MR. Dopamine agonists in animal models of Parkinson's disease: a systematic review and meta-analysis. Parkinsonism Relat Disord 2011 Jun;17(5):313-20.

(4) Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. Improving Bioscience Research Reporting: The ARRIVE Guidelines for Reporting Animal Research. PLoS Biol 2010 Jun 29;8(6):e1000412.

(5) Landis SC, Amara SG, Asadullah K, Austin CP, Blumenstein R, Bradley EW, et al. A call for transparent reporting to optimize the predictive value of preclinical research. Nature 2012 Oct 11;490(7419):187-91.

(6) Baker D, Lidster K, Sottomayor A, Amor S. Two years later: journals are not yet enforcing the ARRIVE guidelines on reporting standards for pre-clinical animal studies. PLoS Biol 2014 Jan;12(1):e1001756.

(7) Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. Nature 2012 Mar 28;483(7391):531-3.

(8) Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? Nat Rev Drug Discov 2011 Aug 31;10(9):712-c1.

(9) Anon. Announcement: Reducing our Irreproducibility. Nature 496[7446], 398. 24-4-2013.

(10) Cramond F, Irvine C, Liao J, Howells D, Sena E, Currie G, et al. Protocol for a retrospective, controlled cohort study of the impact of a change in Nature journals' editorial policy for life sciences research on the completeness of reporting study design and execution. Scientometrics 2016;108:315-28.

(11) Cramond F, Irvine C, Liao J, Howells D, Sena E, Currie G, et al. Protocol for a retrospective, controlled cohort study of the impact of a change in Nature journalsâ€™ editorial policy for life sciences research on the completeness of reporting study design and execution. Scientometrics 2016;108(1):315-28.

(12) Ramirez FD, Motazedian P, Jung RG, Di SP, MacDonald ZD, Moreland R, et al. Methodological Rigor in Preclinical Cardiovascular Studies: Targets to Enhance Reproducibility and Promote Research Translation. Circ Res 2017 Jun 9;120(12):1916-26.

(13) Macleod MR, Fisher M, O'Collins V, Sena ES, Dirnagl U, Bath PM, et al. Good laboratory practice: preventing introduction of bias at the bench. Stroke 2009 Mar;40(3):e50-e52.

(14) Smith R. Peer review: a flawed process at the heart of science and journals. J R Soc Med 2006 Apr;99(4):178-82.

(15) Bahor Z, Liao J, Macleod MR, Bannach-Brown A, McCann SK, Wever KE, et al. Risk of bias reporting in the recent animal focal cerebral ischaemia literature. Clinical Science 2017;131(20):2525-32.

1. Macleod MR, McLean AL, Kyriakopoulou A, et al. Risk of bias in reports of in vivo research: a focus for improvement. *PLoS biology* 2015;13(10):e1002273.
2. 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* 2008;39(10):2824-29.
3. Rooke EDM, Vesterinen HM, Sena ES, et al. Dopamine agonists in animal models of Parkinson's disease: a systematic review and meta-analysis. *Parkinsonism & related disorders* 2011;17(5):313-20.
4. Kilkenny C, Browne WJ, Cuthill IC, et al. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLoS Biol* 2010;8(6):e1000412. doi: 10.1371/journal.pbio.1000412
5. 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.
6. 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 Biol* 2014;12(1):e1001756.
7. 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.
8. Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? *Nat Rev Drug Discov* 2011;10(9):712-7c1.
9. Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. *Nature* 2012;483(7391):531-33.
10. Anon. Announcement: Reducing our Irreproducibility. Nature, 2013:398-98.
11. Cramond F, Irvine C, Liao J, et al. Protocol for a retrospective, controlled cohort study of the impact of a change in Nature journals' editorial policy for life sciences research on the completeness of reporting study design and execution. *Scientometrics* 2016;108(1):315-28.
12. Mogil JS, Macleod MR. No publication without confirmation. *Nature* 2017;542(7642):409-11.
13. Minnerup J, Zentsch V, Schmidt A, et al. Methodological Quality of Experimental Stroke Studies Published in the Stroke Journal: Time Trends and Effect of the Basic Science Checklist. *Stroke* 2016;47(1):267-72.
14. Ramirez FD, Motazedian P, Jung RG, et al. Methodological Rigor in Preclinical Cardiovascular Studies: Targets to Enhance Reproducibility and Promote Research Translation. *Circ Res* 2017;120(12):1916-26.
15. Macleod MR, Fisher M, O'Collins V, et al. Good laboratory practice: preventing introduction of bias at the bench. *Stroke* 2009;40(3):e50-e52.
16. Smith R. Peer review: a flawed process at the heart of science and journals. *J R Soc Med* 2006;99(4):178-82.
17. Blanco D, Biggane AM, Cobo E, et al. Are CONSORT checklists submitted by authors adequately reflecting what information is actually reported in published papers? *Trials* 2018;19(1):80. doi: 10.1186/s13063-018-2475-0
18. Bahor Z, Liao J, Macleod MR, et al. Risk of bias reporting in the recent animal focal cerebral ischaemia literature. *Clinical Science* 2017;131(20):2525-32.