BMJ Open

Influence of external peer reviewer scores for funding applications on funding board decisions: a retrospective analysis of 1561 reviews

Lexy Sorrell,[1] Nicola Mcardle,[2] Taeko Becque,[3] Helen Payne,[2] Beth Stuart,[3] Sheila Turner,[2] Jeremy C Wyatt[4]

To cite: Sorrell L, Mcardle N, Becque T, *et al*. Influence of external peer reviewer scores for funding applications on funding board decisions: a retrospective analysis of 1561 reviews. *BMJ Open* 2018;**8**:e022547. doi:10.1136/bmjopen-2018-022547

► Prepublication history for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2018-022547).

Preliminary analyses of some of these data were presented at the 2017 JAMA Peer Review Congress in Chicago, the invited international workshop on peer review convened by NWO in Amsterdam and the HTAi Annual conference in Rome. Copies of abstracts can be sent on request.

Received 22 February 2018
Revised 23 August 2018
Accepted 18 October 2018

For numbered affiliations see end of article.

Correspondence to
Dr Sheila Turner;
s.turner@soton.ac.uk

ABSTRACT

Objectives To evaluate the influence of external peer reviewer scores on the National Institute for Health Research (NIHR) research funding board decisions by the number of reviewers and type of reviewer expertise.

Design Retrospective analysis of external peer review scores for shortlisted full applications for funding (280 funding applications, 1236 individual reviewers, 1561 review scores).

Setting Four applied health research funding programmes of NIHR, UK.

Main outcome measures Board decision to fund or not fund research applications.

Results The mean score of reviewers predicted funding decisions better than individual reviewer scores (area under the receiver operating characteristic (ROC) curve 0.75, 95% CI 0.69 to 0.81 compared with 0.62, CI 0.59 to 0.65). There was no substantial improvement in how accurately mean reviewer scores predicted funding decisions when the number of reviewers increased above 4 (area under ROC curve 0.75, CI 0.59 to 0.91 for four reviewers; 0.80, CI 0.67 to 0.92 for seven or more). Reviewers with differing expertise influenced the board's decision equally, including public and patient reviewers (area under ROC curves from 0.57, CI 0.47 to 0.66 for health economists to 0.64, CI 0.57 to 0.70 for subject-matter experts). The areas under the ROC curves were quite low when using reviewers' scores, confirming that boards do not rely solely on those scores alone to make their funding decisions, which are best predicted by the mean board score.

Conclusions Boards value scores that originate from a diverse pool of reviewers. On the basis of independent reviewer score alone, there is no detectable benefit of using more than four reviewer scores in terms of their influence on board decisions, so to improve efficiency, it may be possible to avoid using larger numbers of reviewers. The funding decision is best predicted by the board score.

Strengths and limitations of this study

- This study analysed data from a sizeable consecutive cohort of 280 applications for funding and 1561 individual reviews from a major national funder of health research (National Institute for Health Research).
- This study was conducted alongside a qualitative study eliciting the views of stakeholders in the peer review process which gives context to these results.
- The study analysed reviewer scores but did not consider the written comments from peer reviewers.
- The applications were assessed by different numbers of reviewers, giving a range of statistics and making our results harder to compare with some previous studies.

BACKGROUND

The National Institute for Health Research (NIHR)[1] is a significant, long-established UK funder of applied health research. The NIHR aims to select applications for research funding which are of the highest quality and address important health issues, providing much needed evidence for policy and practice. This selection process can be time consuming and costly in terms of human resource,[2 3] and the NIHR is committed to simplifying the research pathway, the time taken for scientific innovation to complete the journey from 'bench to bedside', and to improving the transparency and efficiency of the research funding process. To this end, the NIHR embarked on a programme of work (Push the pace project),[4] and this study is part of that wider piece of work.

Peer review is considered by many to be the gold standard for assessing the scientific quality and relevance of grant applications[5 6]; however, there is much discussion in the literature about the value, effectiveness and fairness of this process.[7–10] There are also many models for the peer review process. Peer reviewers may be internal or external to the funding process,[11] the internal reviewers commonly being members

of an independent decision-making board or committee, and the external peer reviewers working independently of that committee and funding organisation. The contribution from external reviewers often comprises several elements including independent scoring of applications and written comments expressing their views on various aspects of the application. Importantly, the expertise of external reviewers may cover gaps in the expertise of the standing board members.[12]

External peer review is integral to the NIHR process. This process comprises two stages. First, applications for funding are considered by a board and they are either shortlisted or rejected. Then applications which have been shortlisted submit a full application and these are sent out for external peer review. This is reviewed by individuals who are not members of the board and who do not attend the board meeting. The external peer review process involves a variety of external reviewers including clinicians, health economists, methodologists, public and patient reviewers and subject-matter experts who are invited to review and score applications shortlisted for funding, and to comment on their quality and relevance. A variety of factors influence the number of external reviews obtained for each application, including the topic, cost of the study, whether it is primary or secondary research, and time taken for reviewers to respond. Comments and scores obtained from the external reviewers are shared with the board members and the applicants before the board meeting where the applications are discussed and funding decisions made.

The NIHR includes patient and public reviewers in the peer review process; these include patients, potential patients, carers, people who use healthcare services and people who are part of organisations which support these people.[13] Opinions from patient and public reviewers are valued with particular regard to the importance of the research question posed. Comments and scores from all the peer reviewers are then considered by a standing funding board of experts when they are making decisions concerning the funding of those applications.

We acknowledge that reviewer scores are not the only part of the peer review report that influence board decisions as useful narrative comments are also included.[11 12] However, the objectives of our study were to investigate the influence of external peer reviewer scores on the funding decisions made by the board, how many reviewers are needed to review an application, and the relative value of peer review scores from reviewers with differing expertise. While external reviewers and board members assess the importance of the research question and the deliverability of the study, board members are also making a decision regarding funding.

METHODS

Data

The data include 280 full applications submitted to four NIHR research funding programmes (Efficacy and Mechanism Evaluation, Health Services and Delivery Research, Health Technology Assessment and Public Health Research) during 2015 and includes those for which a funding board decided to fund or reject the proposal. It excludes outline applications, a few full applications which were resubmissions and those for which the board decision was deferred to another meeting.

Each application had been scored by between three and nine peer reviewers, making 1236 reviewers in total. Some reviewers reviewed multiple applications (maximum of 6); the total number of reviews for all applications was 1561. Each reviewer awarded an application an overall score on a scale of 1–6, where a score of 1 indicated that the application was extremely poor and unsupportable, and 6 indicated that the application was excellent and could be funded without amendment. The board members considered all the individual reviewer scores and, after discussion, agreed a single score for each application and made a decision regarding funding, which could be fund, reject or invite for resubmission.

Peer reviewer variables analysed included the reviewer's experience with NIHR applying to these four programmes as a chief investigator or co-applicant (number of funded projects and number of times applied); the reviewer's role as assigned by NIHR staff: clinician (32%), methodologist (25%), public and patient reviewer (18%), subject matter expert (17%) and health economist (9%); and the score given to each application. The application variables analysed included the cost and duration of the application, whether the research was primary or secondary, how many reviewers reviewed the application, the funding programme and the board score post discussion (an average of typically 12–15 individuals scoring the application on the scale of 1 to 6; the original, individual scores are not recorded).

Analysis

The data were analysed using STATA V.14 (Stata Corp, College Station, Texas, USA).

Agreement between reviewers

The intraclass correlation (ICC) was calculated to examine agreement between reviewers within applications for all outcomes. The ICC can only be calculated on groups of the same size; therefore, the ICC for applications with differing numbers of reviewers had to be calculated separately. An ICC of less than 0.4 is described as poor reliability, 0.4 to 0.59 is fair, 0.6 to 0.74 is good and above 0.75 is excellent reliability.[14]

Agreement between reviewer scores and board scores

Bland-Altman plots[15] were used to assess agreement between reviewer scores and board scores for all applications included in the study. For each application, the difference between the average board score and the average reviewer score was plotted against the mean of the same two quantities. The 95% limits of agreement, estimated by the mean difference±1.96 SDs of the difference,

provide an interval within which 95% of the differences between board and reviewer scores are expected to lie. If a linear relationship is found between the differences and the means of the board and reviewer scores, regression-based 95% limits of agreement can be computed.[16 17]

Multivariate analyses

Linear regressions were performed of the average board scores on the average reviewer score and all other application variables as well as the number of reviews per application on all relevant application variables. Each funding application had several reviewers; we anticipated that the reviewer scores for some applications might be similar. Multilevel modelling of reviewer scores allowing for clustering by application was therefore used to take account of application and reviewer variables when analysing reviewer scores. An ordered probit regression of reviewer scores on reviewer-level variables (number of applications as chief investigator, number of applications as co-investigator, number of reviews submitted and reviewer role) and application-level variables (application cost, funding stream and research type) with a random effect for application was performed.

Number of reviewers and influence on board decisions

The influence of peer reviewer scores on the board decision, using only applications that resulted in a fund or reject outcome (the very small number invited to resubmit were excluded), was measured by logistic regression with SEs adjusted for clustering. Both the mean and individual reviewers' scores were used to predict board decision, and predicted probabilities from the logistic regression were used to create receiver operating characteristic (ROC) curves. The area under the ROC curve summarises the accuracy of the reviewers' scores in predicting board decision. An area of 1 is perfect prediction, a score of 0.9 to 1 is excellent, 0.8 to 0.9 good, 0.7 to 0.8 fair, 0.6 to 0.7 poor and a score of 0.5 is no better than chance.

The number of peer reviewers used was explored by comparing area under ROC curve using the predicted probabilities from the logistic regression models for applications reviewed by different numbers of reviewers. These are compared to see if increasing the number of reviewers reviewing an application increases the influence of the reviewer scores on the board decision.

Table 2 Intraclass correlation (ICC) values for applications with 4, 5, 6 and 7 external reviewers

Number of reviewers	Number of applications	Average ICC	95% CI
4	40	0.35	−0.05 to 0.63
5	90	0.35	0.11 to 0.54
6	82	0.18	−0.13 to 0.43
7	51	0.41	0.12 to 0.63

Reviewers with specific categories of expertise and varying levels of NIHR application experience

The relative influence of peer reviewer scores from reviewers with different types of expertise and different levels of research experience was measured by logistic regression, using the area under the ROC curve to compare scores from different types of reviewer with the board's decision.

Patient and public involvement

Patients and members of the public were not involved in the design of this study; however, the scores from patient and public representatives who performed peer review of applications for funding were included in the data analysed, and results for this group are presented below.

RESULTS

Data

A summary of the application variables is given in table 1.

Analysis

The mean number of external reviewers per application was 5.6, with scores on average one point higher than the board. The average board scores varied more than the reviewer average scores, with SD of 1.00 and 0.53, respectively. The applications covered a wide range in terms of cost and duration, and a linear regression of the average board scores on the average reviewer score and all other application-level variables (n=280) gave the following statistically significant effects:

- Average board score increased by 0.85 (0.67 to 1.03) for each unit increase in average reviewer score ($p<0.001$).

Table 1 Summary statistics of application variables for all applications in the study n=280

Variable	Mean	Median	SD	Minimum	Maximum
Board score	3.7	3.8	1.00	1.5	5.8
Reviewer average score	4.7	4.7	0.53	2.75	5.6
Total reviewers per application	5.6	6	1.13	3	9
Application cost (£)	881 000	605 000	979 000	50 000	11 500 000
Project duration (months)	35.1	32	17.7	6	148

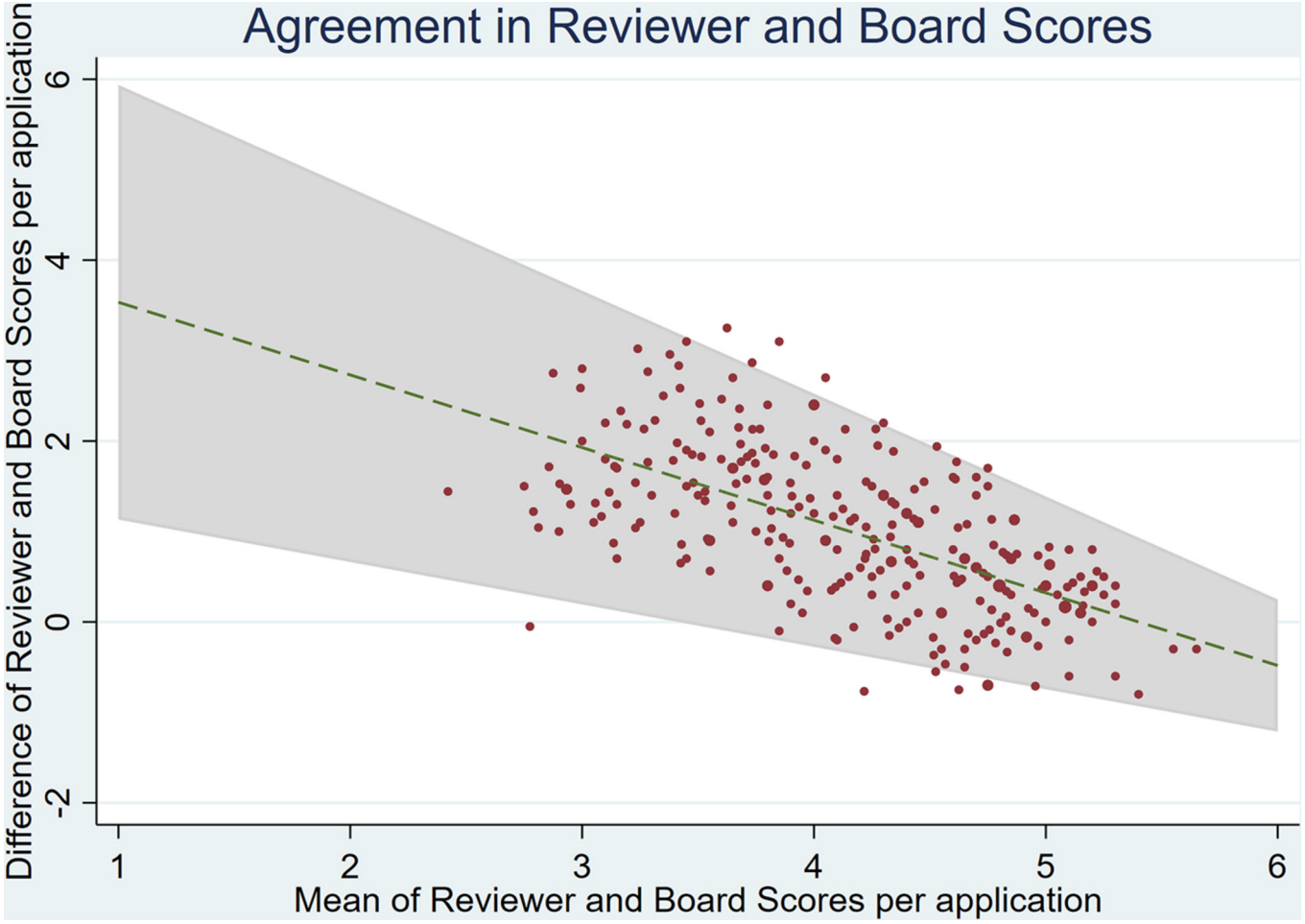

Figure 1 Adjusted Bland-Altman plot of reviewer average scores and board average scores (each dot represents one application, larger dots indicate multiple applications at the same value, the dashed green line is the mean difference and the shaded grey area is the 95% limits of agreement).

- Average board score decreased by −0.19 (−0.31 to −0.07) for each £1 000 000 increase in application cost (p=0.002).

Agreement between reviewers

The ICC can only be calculated for applications with the same number of reviewers, so it has been calculated multiple times. ICCs of reviewer scores within applications are shown in table 2; the agreement between reviewers reviewing an application is judged low for each number of reviewers.

Agreement between reviewer scores and board scores

The standard Bland-Altman plot shows a linear relationship between the difference of the board and reviewer scores and the mean of the board and reviewer scores. The difference in mean scores between reviewers and the board is greater for lower scoring applications than higher scoring applications.

A linear regression is used to adjust the 95% limits of agreement accordingly (figure 1). This shows that higher scoring applications have higher agreement than lower scoring applications.

Multivariate analyses

A linear regression of the number of reviews per application showed that the number of reviews obtained varied by the type and cost of research, changing by −0.98 (−1.36 to −0.60) reviews for secondary research (predominantly evidence synthesis and systematic reviews) compared with primary research (p<0.001) and increasing slightly by 0.20 (0.06 to 0.34) reviews for each £1 000 000 increase in application cost (p=0.006).

No relevant statistically significant results were obtained through multilevel modelling of reviewer scores on application-level and reviewer-level covariates, that is, none of the included variables was a statistically significant predictor of reviewer score.

Number of reviewers and influence on board decisions

The logistic regression model with a binary outcome of fund or reject against individual reviewer scores gives the area under the ROC curve as 0.62, 95% CI 0.59 to 0.65. Table 3 shows how the area under the ROC curve changed with differing numbers of reviewers. However, only small differences are seen in area under the curve between applications having four to seven or more reviewers.

Table 3 Area under the receiver operating characteristic curve for logistic regression of fund or reject decision from reviewer mean scores with different numbers of external reviewers for 263 applications that resulted in an outcome of fund or reject

Number of reviewers	N (applications)	AUC	95% CI
All applications with fund or reject outcome	263	0.62	0.59 to 0.65
4	38	0.75	0.59 to 0.91
5	82	0.76	0.66 to 0.87
6	79	0.68	0.56 to 0.80
7–9	57	0.80	0.67 to 0.92

AUC, area under the curve.

Reviewers with specific categories of expertise and varying levels of NIHR application experience

Using reviewers' mean scores per application increased the area under ROC curve to 0.75, 95% CI 0.69 to 0.81 (table 4). As the areas under the ROC curves were quite low when using reviewers' scores, this demonstrated that the board does not rely entirely on scores to make the funding decision. The board score best predicts the funding decision, with the area under the ROC curve for the board score of 0.97, 95% CI 0.95 to 0.99. These results can be seen in table 4.

Table 4 Area under the receiver operating characteristic curve for logistic regression of fund or reject decisions from reviewer scores with different roles and NIHR application experience (as expected, the AUC is highest for board scores)

	N	AUC	95% CI
Individual reviewer scores	1467	0.62	0.59 to 0.65
Mean reviewer scores	263	0.75	0.69 to 0.81
Board scores	263	0.97	0.95 to 0.99
Type of reviewer			
Clinical reviewer	470	0.60	0.55 to 0.65
Health economist	130	0.57	0.47 to 0.66
Methodologist	356	0.61	0.56 to 0.66
Public contributor	257	0.64	0.58 to 0.70
Subject-matter expert	254	0.64	0.57 to 0.70
Research experience			
No applications to NIHR programmes	591	0.61	0.58 to 0.66
Applied but unsuccessful	335	0.58	0.53 to 0.64
One or two funded applications	317	0.63	0.57 to 0.69
Three or more funded applications	224	0.65	0.59 to 0.72

AUC, area under the curve; NIHR, National Institute for Health Research.

All types of reviewer (including public reviewers) influenced the board decision. Although scores from all reviewer groups had similar levels of influence on the board, those reviewers with unsuccessful NIHR applications may have a lower influence than those who have either been funded or have not applied; however, all CIs overlap.

DISCUSSION

Statement of the principal findings

Low agreement between reviewer scores is to be expected because they are assigning scores from a different perspective. Our results show that the mean reviewer score predicts the board decision on an application better than individual reviewer scores. However, the Board does not rely entirely on reviewer scores to make their funding decisions. Reviewers tend to award more generous scores than the Board, probably in part because they assess proposals individually against rather broad criteria. However, the Board has access to all the reviewer comments plus the applicant responses to those comments as well as their own opinions and therefore are probably more confident in giving a wider range of scores as they have more context in which to make a judgement. The Bland-Altman plot shows that the Board and external reviewers agree more strongly on applications with a high score. Variability for high-quality and very low-quality applications is probably less than applications with a mix of strengths and weaknesses. However, due to the two-stage application process, there are unlikely to be many proposals where all reviewers agree that the application is weak as these applications have already been rejected at an earlier stage. Therefore, the lower scoring proposals in this study are likely to be those with a mix of strengths and weaknesses and thus may have one or two low scores combined with several high scores, for example the methodology may be weak where everything else is strong.[18] Board scores will take into account these weaknesses, which could explain the decreased agreement for lower scoring proposals.

Where there are more than four reviewers per application, this circumstance does not appear to increase the influence of the external reviewers on the board decision; therefore, there may be no advantage in having more than four reviewers. This has implications regarding the resources needed for the peer review process, as much time is spent by NIHR staff finding appropriate reviewers, by those reviewers completing reviews and by board members considering their comments. Limiting or reducing the number of peer reviewers without an adverse effect on board decision-making would increase efficiency. Scores from all types of reviewer (clinical reviewers, health economists, methodologists, public reviewers and subject-matter experts) have a similar influence on the board decision. There is also no evidence of reviewers with more experience of applying to NIHR for

funding influencing the board's decision. The agreement between reviewers reviewing an application is low for each number of reviewers. This is unsurprising, as NIHR staff deliberately choose reviewers with different expertise to assess each application, so they are likely to view the applications from different perspectives, all of which are valued.

Strengths and limitations of the study

This study investigated the influence of peer reviewer scores from peer reviewers with a variety of expertise, on a large consecutive cohort of applications from a major national funder. This study also benefits from being conducted alongside a qualitative study eliciting the views of stakeholders in the peer review process,[12] which gives context to our results. Considering the findings from these two studies together enhances our overall understanding of the peer review process and of the possible changes which might be made to enhance the efficiency of the process.

A weakness of the study was that, as different numbers of reviewers were used to review applications, and the agreement statistics such as the ICC were calculated on each category of numbers of reviewers, this gave a range of values which makes it hard to compare with previous work. A more complex approach to modelling using resampling might have allowed us to explore this but was beyond the scope of this study.

The ICCs were also calculated for each reviewer role; however, this divided the applications into more categories for ICC calculation and the numbers within each were too small to report. For some of the calls for research, applications were in competition with each other, so it was possible that applications may have been rated highly but were still not funded.

Strengths and weaknesses in relation to other studies, discussing important differences in results

Few studies have previously looked at the reliability of external peer review of funding applications[19 20] or evaluation by boards and panels.[5 20] Some studies have looked at inter-rater reliability where a selected cohort of reviewers or boards[21] have evaluated multiple applications. We have found no similar studies analysing data from several external reviewers scoring each application.

Previous studies have found that inter-reviewer differences can be reduced by consensus discussion,[5 22] and that a range of reviewer agreement levels within applications is not unusual.[5 23 24] In our case, we deliberately obtain contrasting reviewer perspectives, giving a comprehensive and balanced view of the application. Training for reviewers has been suggested[12 23] and, although this is not necessarily the aim, training may improve inter-rater agreement. There is limited evidence about the benefits of including patient and public peer reviewers.[22] Our study included patient and public reviewers and found that their scores were as influential as scores from other types of reviewers regarding the funding decision. Previous work[22] found that patient reviewers scored more harshly than scientists, while in our study public and patient reviewers scored highly compared with the other reviewers. This difference may be because the patient reviewers in Fleurence *et al*'s study were able to discuss applications with scientists and subsequently revise their scores; this did not apply to any of our external peer reviewers.

The optimum number of reviewers is a frequently debated question,[25] the concern being that using too few may impact negatively on the quality of decision-making, whereas too many consumes unnecessary resources. Our findings based on scores alone indicate that more than four external peer reviewers may not be needed. Snell[25] used peer review scores from a postdoctoral fellowships competition to show that five reviewers per application represented a practical optimum number of reviewers. Findings from qualitative work conducted alongside this study[12] suggested merits in developing a more proportionate approach across all funding programmes by varying the numbers of reviewers according to the amount of funding requested by the application.

Meaning of the study: possible explanations and implications for clinicians and policy-makers

The NIHR peer review process is viewed as worthwhile and important to the funding boards, external reviewers and applicants.[12] Our findings indicate that the NIHR should continue to encourage participation from public and patient reviewers, as their expertise and opinions on new potential research is valued. Improvements to the peer review process, however, may enhance efficiency, and there may be little benefit from having a large number of reviewers (eg, six or more) to review applications. Limiting the number of reviewers would reduce the workload for stakeholders including board members and funding agency staff, thereby conserving the overall resource for research.[26]

Unanswered questions and future research

Future work to elucidate the interactions and discussions by boards regarding funding decisions for applications would provide greater insight into the peer review process. Analysing the discordance between reviewer scores would enable us to explore further the importance of individual reviewer scores and how they influence the funding board. Future research focusing on the relationship between the funding process (eg, reviewer and board scores) and the outcomes and long-term impact of research and the influence of PPI input on these would be very informative.

CONCLUSIONS

This study investigated the influence of external peer reviewer scores on funding board decisions, and our results indicate that having more than four reviewers per application does not increase this. Reviewers with

different types of expertise, including public and patient reviewers, influenced the board's decision equally.

Author affiliations
[1]School of Computing, Electronics and Mathematics, University of Plymouth, Plymouth, UK
[2]National Institute for Health Research, Evaluation, Trials and Studies Coordinating Centre (NETSCC), University of Southampton, Southampton, UK
[3]Primary Care and Population Sciences, University of Southampton, Southampton, UK
[4]Wessex Institute, Faculty of Medicine, University of Southampton, Southampton, UK

Acknowledgements Colleagues at National Institute for Health Research Evaluation, Trials and Studies Coordination Centre, especially Andy Masters, Neil Challis, Shameer Gandhi and Tim Ellwood for their assistance in collating the data and Geoff Frampton and Jonathan Shepherd for their comments.

Contributors The study was conceived and designed by JCW, NCM, ST and HP; undertaken by NCM, JCW and LS. LS led the writing guided by JCW, NCM, ST, BS and TB. All authors read and approved the final manuscript.

Funding This research was supported by the NIHR Evaluation, Trials and Studies Coordinating Centre through its Research on Research Programme.

Disclaimer The views and opinions expressed are those of the authors and do not necessarily reflect those of the Department of Health or of NETSCC.

Competing interests The authors have no competing financial interests; however, NCM, HP, ST and JCW were employed by the University of Southampton to work at least part time for the NIHR at the time when the study was conducted.

Patient consent Not required.

Ethics approval Advice about the need for ethics approval for this study was sought from the University of Southampton's Ethics and Research Governance Online (ERGO) service, and it was agreed that approval was not required for this study as it was a process evaluation using routine data, no patients were contacted as part of this study and no NHS premises were used to identify our study participants.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement All data requests should be submitted to the corresponding author for consideration. Access to available anonymised data may be granted following review.

REFERENCES

1. National Institute for Health Research. https://www.nihr.ac.uk/ (accessed Aug 2017).
2. Schroter S, Groves T, Højgaard L. Surveys of current status in biomedical science grant review: funding organisations' and grant reviewers' perspectives. *BMC Med* 2010;8:62.
3. Bonetta L. Enhancing NIH grant peer review: a broader perspective. *Cell* 2008;135:201–4.
4. National Institute for Health Reseach (NIHR). Push the pace. http://www.nihr.ac.uk/about-us/how-we-are-managed/boards-and-panels/push-the-pace.htm (accessed Aug 2017).
5. Fogelholm M, Leppinen S, Auvinen A, *et al*. Panel discussion does not improve reliability of peer review for medical research grant proposals. *J Clin Epidemiol* 2012;65:47–52.
6. Reinhart M. Peer review of grant applications in biology and medicine. Reliability, fairness, and validity. *Scientometrics* 2009;81:789–809.
7. Marsh HW, Jayasinghe UW, Bond NW. Improving the peer-review process for grant applications: reliability, validity, bias, and generalizability. *Am Psychol* 2008;63:160–8.
8. Mayo NE, Brophy J, Goldberg MS, *et al*. Peering at peer review revealed high degree of chance associated with funding of grant applications. *J Clin Epidemiol* 2006;59:842–8.
9. Guthrie S, Ghiga I, Wooding S. What do we know about grant peer review in the health sciences? [version 1; referees: 1 approved, 1 approved with reservations]. 2017 https://f1000research.com/articles/6-1335/v1 (accessed Jan 2018).
10. Bornmann L. Scientific peer review. *Annual Review of Information Science and Technology* 2011;45:197–245.
11. Abdoul H, Perrey C, Amiel P, *et al*. Peer review of grant applications: criteria used and qualitative study of reviewer practices. *PLoS One* 2012;7:e46054.
12. Turner S, Bull A, Chinnery F, *et al*. Evaluation of stakeholder views on peer review of NIHR applications for funding: a qualitative study. *In press BMJ Open* 2018.
13. NIHR. Patients and the public. https://www.nihr.ac.uk/patients-and-public/ (accessed Jan 2018).
14. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess* 1994;6:284–90 http://doi.apa.org/getdoi.cfm?doi=.
15. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;1:307–10.
16. Mander A. BATPLOT: Stata module to produce Bland-Altman plots accounting for trend. 2012 https://econpapers.repec.org/RePEc:boc:bocode:s448703 (accessed Jan 2018).
17. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res* 1999;8:135–60.
18. Eblen MK, Wagner RM, RoyChowdhury D, *et al*. How criterion scores predict the overall impact score and funding outcomes for national institutes of health peer-reviewed applications. *PLoS One* 2016;11:e0155060.
19. Benda WGG, Engels TCE. The predictive validity of peer review: a selective review of the judgmental forecasting qualities of peers, and implications for innovation in science. *Int J Forecast* 2011;27:166–82.
20. Demicheli V, Di Pietrantonj C. Peer review for improving the quality of grant applications. *Cochrane Database Syst Rev* 2007;2:MR000003.
21. Clarke P, Herbert D, Graves N, *et al*. A randomized trial of fellowships for early career researchers finds a high reliability in funding decisions. *J Clin Epidemiol* 2016;69:147–51.
22. Fleurence RL, Forsythe LP, Lauer M, *et al*. Engaging patients and stakeholders in research proposal review: the patient-centered outcomes research institute. *Ann Intern Med* 2014;161:122–30.
23. Sattler DN, McKnight PE, Naney L, *et al*. Grant peer review: improving inter-rater reliability with training. *PLoS One* 2015;10:e0130450.
24. Lobb R, Petermann L, Manafo E, *et al*. Networking and knowledge exchange to promote the formation of transdisciplinary coalitions and levels of agreement among transdisciplinary peer reviewers. *J Public Health Manag Pract* 2013;19:E9–E20.
25. Snell RR. Menage a quoi? optimal number of peer reviewers. *PLoS One* 2015;10:14:e0120838.
26. Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. *Lancet* 2009;374:86–9.