PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	The Association Between Mentoring and Training Outcomes in Junior Doctors in Medicine: An Observational Study
AUTHORS	Ong, John; Swift, Carla; Magill, N; Ong, Sharon; Day, Anne; Al-Naeeb, Yasseen; Shankar, Arun

VERSION 1 – REVIEW

REVIEWER	Colin Mitchell Imperial College Healthcare NHS Trust UK
REVIEW RETURNED	21-Dec-2017

GENERAL COMMENTS	This is an interesting topic and the intention to add quantitative findings to the evaluation of mentoring is admirable and potentially valuable. Mentorship seems likely to be a valuable tool in improving medical training but it is not without cost and investigating it rigorously is highly relevant. The investigators use a selection of reasonable and valid outcome measures to look for differences between two groups of trainees - those who have been mentored as part of a voluntary RCP scheme, and a normal group who have not been part of this mentoring program. Only trainees from each group who responded to a voluntary survey were included. Significant differences were found in exam pass rates and ARCP outcomes as well as markers of satisfaction. It is inherent in an observational study design that only correlation can be shown, not causation. The authors also rightly mention the inherent selection bias in their two groups. The intervention group have found out about, volunteered for, and participated in a mentorship program, while the other group did not. Clearly this inserts a huge potential for selection bias - the intervention group seem very likely to be self-selected as a highly engaged, enthusiastic, informed and committed group of trainees, which would be a powerful explanation for the resulting finding of their superior performance compared to the norm. This does not mean that the mentorship intervention was ineffective, but it does inject a large amount of doubt into their interpretation of the findings. In a study like this, particularly with such a potential for confounding, I would hope to see significant efforts made to ameliorate this problem, for example by showing that the groups were (prior to the intervention) similar in ways relevant to the outcome measures. Unfortunately there is very little such information, and that which is present seems to suggest that the groups are genuinely different - for example in terms of age (intervention group was seemingly older) and the number of international medical graduates (fewer in the intervention group). The findings would be more significant if it were known that the intervention and control groups had similar prior academic

	achievements or ARCP pass rates, for example, or if there was some form of time/dose response to mentoring (ie the more you are mentored the better you get) suggesting an effect from mentoring even if the groups are different at baseline. An alternative strategy would be to try to incorporate trainees who applied for the mentorship scheme but then did not participate, to see if their outcomes were closer to the mentored group (suggesting selection bias) or the unmentored group (suggesting the intervention was the cause of the findings). Unfortunately I suspect this data would be difficult to obtain and the numbers small so even if such data is available it may not help much. Ultimately, in my opinion, the selection limitation is far more significant than the authors seem to be suggesting and in fact makes the findings inconclusive, which is a shame as clearly this evaluation has been done with the best of intentions and the paper is well written. It is let down by the study design, or lack of mitigation for the design's inherent limitations. It may be that the authors have additional data on their subjects or can obtain it post-hoc, which might allow these issues to be addressed directly. Alternatively the paper could be re-written to acknowledge the significance of the limitation although I suspect this would make the conclusions less noteworthy.

REVIEWER	Profesor Gregory Crawford Senior Consultant in Palliative Medicine, Northern Adelaide Local Health Network and Professor of Palliative Medicine, University of Adelaide, Australia
REVIEW RETURNED	27-Dec-2017

GENERAL COMMENTS	Thank you for this interesting paper. I did not gain a sense of what mentoring under this program actually entails. Some clearer description of the basic requirements of mentors and mentees would assist in understanding better the actual intervention being studied. I am concerned about the statistics for 2 very different sized arms of this study. The authors mention the potential for selection bias but it may be that higher functioning and achieving trainees are more open to mentoring. There was a style issue that I find difficult. There are sentences that are start with numbers, rather than words. There was no mention of whether ethics approval was sought or that a waiver was appropriate. This needs to be addressed.

REVIEWER	Idaira Rodriguez Santana Centre for Health Economics (University of York), United Kingdom
REVIEW RETURNED	23-Jan-2018

GENERAL COMMENTS	This is an interesting paper on the effect of mentoring programmes on training outcomes. Setting: The quantitative data and qualitative answers analysed in the paper come from a questionnaire send to the trainees, however relevant information about the method of data collection is missing. Authors should consider the inclusion of elements such us the type of questions (e.g multiple choice, open questions…), the method of

	contact, whether any incentives were given to participants or the number of reminders sent to participants. While low response rates to surveys are very common, the authors should include more discussion on the ways they think their sample characteristics differ from the target population (e.g. Do you have some descriptive statistics from the target population? How those differ from your sample stats?) Participants: Authors compare the outcomes from two mutually exclusive groups. The paper will benefit from more information on how students access that mentoring programme. Is the mentoring programme available to every student? If better achievers are more likely to be part of the mentored group then the differences in pass rates favouring mentored trainees are very likely to be due to the fact that they are better performers and not to the mentoring effect per se. Moreover, why were trainees from East of England chosen? How they may differ from trainees from other locations in the UK? Variables: The main outcome variables are described in detail in the text, but the authors say nothing about the other variables included in the study until the result section. For example, for the variable age, what is the reasoning behind the chosen age groups? If most students are concentrated in two central intervals, then having three empty categories is not very informative. Biases: As previously said, there might be self-selection into the mentoring group (are those doctors better achievers overall?) and response-bias (how respondents differ from the ones who didn't respond the questionnaire?). Moreover, if outcomes pass rates are also self-reported it might be that unsuccessful student are less likely to respond to the questionnaire. Authors acknowledge the existence of some of the biases in the discussion section, but a more detailed discussion on their implications and the direction of them is needed. Missing data: Is this a complete case analysis, or is there any strategy the authors used to deal with partially missing observations? Statistical Methods: The description of the statistical analysis is too concise, authors could explain the motivation for the chosen methods in more detail. The authors use a non-parametric statistical procedure chi-square that tell us how confident we can be about saying that the observed results differ from expected results (i.e. no differences between mentored and non-mentored group) and compute the confidence intervals for odd ratios using the Batista-Pike method (there is no information on how this method works, neither a reference to the relevant paper). Although authors highlight as a strength the novel quantitative data, the small sample size and the multiple biases present in the sample data question the reliability of the findings. Identifying the causal effect of social interactions (such as mentoring) by means of quantitative data is rather a challenging matter as it is very difficult to disentangle its effect from the effect of other confounders. If authors' main objective is providing evidence that would to justify the use of randomised control trials (in order to identify the causal effect of mentoring in postgraduate training outcomes), I think the study will benefit from being presented as a mixed methods study. The latter can be done by increasing the importance of the qualitative information in this study rather than solely focus on the quantitative aspects. Finally, regarding the use of RCT that might be unfeasible or very costly for this type of setting, authors might explore multivariate regression methods using observational data or survey data as a way of identifying causal effects from mentoring programmes.

REVIEWER	Dr Philp M Sedgwick St. George's, University of London, Cranmer Terrace, London SW17 0RE United Kingdom
REVIEW RETURNED	26-Jan-2018

GENERAL COMMENTS	Thank you for asking me to review this interesting study. Mentoring of junior doctors has obvious potential benefits for a doctor's career and the future welfare of patients. The points of interest that the authors need to pay attention to as follows: Study Design: Possibly of minor importance, I do not believe this is a cross-sectional study as there are two non-randomised groups, one of which received the intervention and the other did not, that were compared in a series of outcomes. The taxonomy of study designs is rigid, and it is difficult to ascribe a study design here. Obviously the researchers have not intervened in any way, and therefore the general taxonomy is one of observational. As the groups are not measured before and after the mentoring (and neither could they), it is not a before and after study. Ethical review: Whilst such a study presents limited ethical challenges, it is not clear if some form of ethical review was needed. The participants self-selected if they underwent the mentoring scheme, and we can infer informed consent if questionnaires were returned. Nonetheless, the authors should have considered ethical review before commencing; there was no mention in the manuscript. Statistical Analysis: Odds ratios were used to evaluate the association between characteristics and outcomes. As it would appear that no adjustment for confounding was made (through logistic regression, for example), the use of relative risks would have been more appropriate; in this instance it would have been possible since it was possible to estimate the population at risk. Nonetheless, I am not sure if I found some errors in the calculation of the presented odds ratios. On page 9 it states "The pass rate of the MRCP Part 1 exam is significantly higher in trainees receiving mentorship compared to non-mentored trainees; 84.0% (21/25) vs. 42.4% (36/85), $p < 0.01$ (OR=7.1, 95% CI 2.4 to 20.3). If I have interpreted the data correctly I calculated the unadjusted odds ratio to be 1.98. I did not subsequently check the remaining calculations, but i strongly advise the researchers to do so. Presentation of results: The results could have been presented more succinctly and perhaps in a more sophisticated way using a table, rather than a series of very simple pie charts and bar charts. The latter seem to take up a lot of space unnecessarily. I always find it strange when researchers feel the need to display the distribution of sex with it two categories using a pie chart; it is standard practice to display simple data such as this in tables.

	Generally this is a well written paper. The aims and objectives of the study were clearly identified and subsequently addressed. At times I found the text slightly muddled and more attention needs to be given to the flow of information. At times there were minor inaccuracies. This is easily fixed. I am not a great fan of titles and sub headings that are pre-loaded in outcome or inference. For example, I would prefer a title for the paper that incorporated the following, for example, "The Association between Mentoring and Training Outcomes in Junior Doctors in Medicine: A questionnaire survey." It would be good if the authors did not automatically assume that statistical significance implied significance generally. There should be a more informed approach to investigation and interpretation of the data. The authors did discuss limitations of their data. However, it would have been beneficial to do so in a more structured way. In particular the mentored group are a very self-selected and presumably motivated group.You might expect them to do well in the outcomes recorded. Not all of those junior doctors approached responded, and there is huge potential for non-response bias. Equally there is huge potential for response bias from the respondents in their answers. This makes it difficult to infer association although the researchers were generally careful in this respect. Nonetheless, I think it was very premature that on the basis of the results for this study it was proposed that randomised controlled trials should be undertaken. The authors did make a mistake in the Discussion regarding the limitations and possible source of bias; on page 12 "There were also more trainees aged 31 to 35 years in the mentored group compared to the non-mentored group and this may have occurred by chance or response bias." it should read "non-response bias" and not "response bias".

VERSION 1 – AUTHOR RESPONSE

Reviewer 1: Dr Collin Mitchell, Imperial College Healthcare NHS Trust, UK

Reviewer 1: “In a study like this, particularly with such a potential for confounding, I would hope to see significant efforts made to ameliorate this problem, for example by showing that the groups were (prior to the intervention) similar in ways relevant to the outcome measures. Unfortunately there is very little such information,…”

Response 1.1: We attempted this by choosing trainees from the RCP mentoring scheme (less East of England trainees), a national cohort representative of the UK CMT population which would ameliorate inter-deanery variability if any. East of England trainees were chosen for specific reasons discussed below (see Response 3.4). To improve the rigor of our study, we now provide logistic regression in our analyses. More information is now provide on Page 11-12, and the newly revised Figures. Voluntary surveys in pre- and post-intervention groups do not reduce the risk of self-selection and non-response bias.

Reviewer 1: "…and that which is present seems to suggest that the groups are genuinely different - for example in terms of age (intervention group was seemingly older) and the number of international medical graduates (fewer in the intervention group)."

Response 1.2: We now display the full composition data of our respondents in the revised Figure 1 and together with the logistic regression data provided on Page 11-12, demonstrate there are no significant differences between both groups.

Reviewer 1: "The findings would be more significant if it were known that the intervention and control groups had similar prior academic achievements or ARCP pass rates, for example, or if there was some form of time/dose response to mentoring (ie the more you are mentored the better you get) suggesting an effect from mentoring even if the groups are different at baseline."

Response 1.3: We agree with Dr Mitchell, however we point out this is difficult to achieve without significantly increasing the length of the survey. For example, degree with distinctions or honours are typically awarded to the top five to ten percent of a graduating cohort and ARCP pass rates are 100% in all Foundation Year 2 doctors recruited to CMT1 posts since successful completion of Foundation Year Training is a pre-requisite to joining CMT. Response to research surveys are notoriously poor and we decided a succinct questionnaire was more important to improve response rates and obtain sufficient numbers for statistical analyses. The analysis of time/dose responses is beyond the scope of this unfunded study. Such a study would need to investigate factors including length and frequency of interactions, communication medium and quality of the mentor-mentee interaction etc.

Reviewer 1: "An alternative strategy would be to try to incorporate trainees who applied for the mentorship scheme but then did not participate, to see if their outcomes were closer to the mentored group (suggesting selection bias) or the unmentored group (suggesting the intervention was the cause of the findings). Unfortunately I suspect this data would be difficult to obtain and the numbers small so even if such data is available it may not help much."

Response 1.4: A mentee's drive and ambition cannot be gauged by the disengagement or nonparticipation in the mentoring process. We have discussed the reasons why trainees disengage from mentoring in our manuscript. Theoretically, the "ideal" negative control group would be equally driven CMTs who sought mentorship with the RCP but were then matched by similar attributes and randomised to not receive mentoring for the purpose of this study. However, this is unethical and goes against current guidance on mentoring by the General Medical Council. We have now included this in the Discussion (Page 19-21) for readers to consider. However with the lack of an ideal negative control, we did attempt the suggested approach and as Dr Mitchell has correctly suspected, the number of responses to our survey from mentees who joined the RCP Mentoring Programme but did not take part in the mentoring process were too small for any meaningful analyses.

Reviewer 1: "It may be that the authors have additional data on their subjects or can obtain it post-hoc, which might allow these issues to be addressed directly. Alternatively the paper could be re-written to acknowledge the significance of the limitation although I suspect this would make the conclusions less noteworthy."

Response 1.5: Additional data has been provided as discussed above. Further to this, we include new comparisons to exam pass rates for all UK trainees taking the MRCP(UK) exams in 2017 to supports our observations. We have also improved the presentation of our data (revised Figure 1-4), expanded our discussion on the limitations of the study and the implication of the observed results on Page 19-21.

Reviewer 2: Prof Gregory Crawford, Palliative Medicine, University of Adelaide, Australia

Reviewer 2: “Thank you for this interesting paper. I did not gain a sense of what mentoring under this program actually entails. Some clearer description of the basic requirements of mentors and mentees would assist in understanding better the actual intervention being studied.”

Response 2.1: We have provided a generic description on Page 4-5. Mentoring is trainee led and therefore there will be variation between trainee-trainer interactions and activities.

Reviewer 2: “I am concerned about the statistics for 2 very different sized arms of this study. The authors mention the potential for selection bias but it may be that higher functioning and achieving trainees are more open to mentoring.”

Response 2.2: Our statistical methods have been checked by a statistician at King's College London, Mr Nick Magill, who is one of the co-authors of this manuscript. We acknowledge that our study may have been susceptible to self-selection bias and we discuss this in length in the section entitled “Limitations of the study and special considerations for future research.” (Page 19-21). However, we also point out that this is a separate research question not covered by the scope of this study i.e. “Does mentee intelligence and drive affect mentoring outcomes?”

Reviewer 2: “There was a style issue that I find difficult. There are sentences that are start with numbers, rather than words.”

Response 2.3: We have revised this accordingly.

Reviewer 2: “There was no mention of whether ethics approval was sought or that a waiver was appropriate. This needs to be addressed.”

Response 2.4: We have now described our ethics approval on Page 9 and Appendix 1.

Reviewer 3: Ms Idaira Rodriguez Santana, Centre for Health Economics, University of York

Reviewer 3: “Authors should consider the inclusion of elements such as the type of questions (e.g multiple choice, open questions…), the method of contact, whether any incentives were given to participants or the number of reminders sent to participants.”

Response 3.1: Thank you, we have now added this information on Pages 8-9.

Reviewer 3: “While low response rates to surveys are very common, the authors should include more discussion on the ways they think their sample characteristics differ from the target population (e.g. Do you have some descriptive statistics from the target population? How those differ from your sample stats?)”

Response 3.2: The matched MRCP and ARCP pass rates of the “target population” (we assume Ms Santana is referring to nationwide non-mentored trainees) is unknown. The purpose of the negative control group was to provide us with such information. We have now discussed the reasons why getting this background data is difficult (Page 19-21) and to our knowledge we are the first study to attempt assessing both MRCP and ARCP pass rates together. However, we now provide comparisons of MRCP pass rates to all UK candidates sitting the exams in 2017 therefore achieving better representation of the target group.

Reviewer 3: "The paper will benefit from more information on how students access that mentoring programme. Is the mentoring programme available to every student?"

Response 3.3: We have now added this information on Page 4-5.

Reviewer 3: "Moreover, why were trainees from East of England chosen? How they may differ from trainees from other locations in the UK?"

Response 3.4: East of England trainees were specifically chosen as a negative control group because at the time of the study, no mentoring programme for medicine was active within the region. In contrast, other deaneries had separate mentoring programmes for different stages of training (e.g. London deanery, Health Education England Thames Valley deanery). This would have made the standardisation of positive and negative controls difficult e.g. Career grade of mentors, level of training delivered to mentors, mentees from other mentoring programmes responding to our survey etc. This is now described on Page 7-8.

Reviewer 3: "The main outcome variables are described in detail in the text, but the authors say nothing about the other variables included in the study until the result section. For example, for the variable age, what is the reasoning behind the chosen age groups? If most students are concentrated in two central intervals, then having three empty categories is not very informative."

Response 3.5: The other variables include basic demographics (gender, age group) and career information (stage of training) which require little explanation. We saw no reason in decreasing intervals (or increasing number of age groups) because it serves no purpose to the study. With regards to the country of primary qualification, International Medical Graduates (IMGs) are usually observed to have lower pass rates in the MRCP exams. This is also observed in the US medical exams (USMLE). The supporting data is readily available online. This phenomenon is widely known within the field of postgraduate medical education and it has been previously studied. We did not discuss this phenomenon because it is complex, politically charged and not in the scope of the study. The ARCP pass rates of IMGs are not known (or unpublished).

Reviewer 3: “As previously said, there might be self-selection into the mentoring group (are those doctors better achievers overall?) and response-bias (how respondents differ from the ones who didn’t respond the questionnaire?).”

Response 3.6: To answer these questions, the survey has to be made compulsory for all trainees to complete but there are ethical considerations preventing this. We now discuss this in length on Page 19-21. For a better representation of candidates who sat the MRCP exams, we have now compared our results to the national MRCP(UK) pass rates in 2017, when this survey was conducted (see revised Figure 2).

Reviewer 3: “Authors acknowledge the existence of some of the biases in the discussion section, but a more detailed discussion on their implications and the direction of them is needed.”

Response 3.7: We have now provided a more extensive discussion on Page 19-21.

Reviewer 3: “Is this a complete case analysis, or is there any strategy the authors used to deal with partially missing observations?”

Response 3.8: This partly relates to response 3.6. We have now stated our exclusion criteria on Page 9-10.

Reviewer 3: “The description of the statistical analysis is too concise, authors could explain the motivation for the chosen methods in more detail.”

Response 3.9: We have now provided a more informative description on Page 10-11. We used chi-squared tests to test the association between mentoring and binary outcomes when n > 5 in a 2x2 contingency table. When n ≤ 5 in the contingency table, the Fisher's exact test was used to calculate p-values for better accuracy. We report marginal ORs and risk ratios. The Koopman asymptotic method [12] was used to calculate the confidence intervals of the relative risk (RR) and the Baptista-Pike method was used to calculate confidence intervals for the Odd's Ratio (OR) [13].
We also performed logistic regression in order to calculate conditional (adjusted) odds ratios. These can be compared to the unadjusted odds ones to assess the amount of confounding. We explained the use of age and country of primary qualification in the main text. Graphpad 7.0 was used to perform the chi-squared analyses and Medcalc Version 18 was used to perform the logistic regression.

Reviewer 3: "The authors use a non-parametric statistical procedure chi-square that tell us how confident we can be about saying that the observed results differ from expected results (i.e. no differences between mentored and non-mentored group) and compute the confidence intervals for odd ratios using the Batista-Pike method (there is no information on how this method works, neither a reference to the relevant paper)."

Response 3.10:We have now provided references for the Koopman asymptotic used to calculate the confidence intervals of the relative risk [12] and the Baptista-Pike methods used to calculate the confidence intervals of the Odds Ratio [13].

Reviewer 3: "I think the study will benefit from being presented as a mixed methods study. The latter can be done by increasing the importance of the qualitative information in this study rather than solely focus on the quantitative aspects."

Response 3.11:The main aim of our paper was to evaluate quantitative data. As an internal check, we also collected matched qualitative results which was used to evaluate congruency in response and support our conclusion of a positive association. We have now expanded and emphasized the importance of our qualitative results on Page 8. There is an abundance of qualitative studies in current literature on mentoring so we have not emphasized this further.

Reviewer 3: "Finally, regarding the use of RCT that might be unfeasible or very costly for this type of setting, authors might explore multivariate regression methods using observational data or survey data as a way of identifying causal effects from mentoring programmes."

Response 3.12:We appreciate that there are several methods and complexities in investigating this further. We have now amended our statement and assume a broader stance to allow other groups to determine how best to investigate the effects of mentoring given the challenges we faced (Page 3 and Page 21)

Reviewer 4: Dr Philip M Sedgwick, School of Education, St. George's University of London

Reviewer 4: "Possibly of minor importance, I do not believe this is a cross-sectional study as there are two non-randomised groups, one of which received the intervention and the other did not, that were compared in a series of outcomes. The taxonomy of study designs is rigid, and it is difficult to ascribe a study design here. Obviously the researchers have not intervened in any way, and therefore the general taxonomy is one of observational. As the groups are not measured before and after the mentoring (and neither could they), it is not a before and after study."

Response 4.1: We have amended our manuscript to describe an observational study.

Reviewer 4: “Ethical review: Whilst such a study presents limited ethical challenges, it is not clear if some form of ethical review was needed. The participants self-selected if they underwent the mentoring scheme, and we can infer informed consent if questionnaires were returned. Nonetheless, the authors should have considered ethical review before commencing; there was no mention in the manuscript.”

Response 4.2: We have now described our ethical approval of the study on Page 9 and Appendix 1.

Reviewer 4: “Odds ratios were used to evaluate the association between characteristics and outcomes. As it would appear that no adjustment for confounding was made (through logistic regression, for example), the use of relative risks would have been more appropriate; in this instance it would have been possible since it was possible to estimate the population at risk. Nonetheless, I am not sure if I found some errors in the calculation of the presented odds ratios. On page 9 it states "The pass rate of the MRCP Part 1 exam is significantly higher in trainees receiving mentorship compared to non-mentored trainees; 84.0% (21/25) vs. 42.4% (36/85), $p < 0.01$ (OR=7.1, 95% CI 2.4 to 20.3). If I have interpreted the data correctly I calculated the unadjusted odds ratio to be 1.98. I did not subsequently check the remaining calculations, but i strongly advise the researchers to do so.”

Response 4.3: We have gone through our calculations again and the unadjusted OR is correctly reported as 7.1. However, we did identify other minor inaccuracies which we have now corrected. As recommended by Dr Sedgwick we have now reported the relative risk as well as included analysis of binary outcomes using logistic regression in order to adjust for possible confounding. In light of this, we have also reported the unadjusted odds ratios to enable an assessment of the amount of confounding (by comparing the unadjusted and adjusted ORs).

Reviewer 4: The results could have been presented more succinctly and perhaps in a more sophisticated way using a table, rather than a series of very simple pie charts and bar charts. The latter seem to take up a lot of space unnecessarily. I always find it strange when researchers feel the need to display the distribution of sex with it two categories using a pie chart; it is standard practice to display simple data such as this in tables.

Response 4.4: We have now amended Figure 1 to include the demographic and composition datain a table.

Reviewer 4: Generally this is a well written paper. The aims and objectives of the study were clearly identified and subsequently addressed. At times I found the text slightly muddled and more attention needs to be given to the flow of information. At times there were minor inaccuracies. This is easily fixed. I am not a great fan of titles and sub headings that are pre-loaded in outcome or inference. For example, I would prefer a title for the paper that incorporated the following, for example, "The Association between Mentoring and Training Outcomes in Junior Doctors in Medicine: A questionnaire survey."

Response 4.5: We have revised the title and some relevant sub-headings in the manuscript to address these minor issues. We have also made changes to improve the flow of the manuscript.

Reviewer 4: “It would be good if the authors did not automatically assume that statistical significance implied significance generally. There should be a more informed approach to investigation and interpretation of the data.”

Response 4.6: We have made our text more conservative. We clearly report the associations observed in our results (e.g. Page 3, Page 21 and other changes within the manuscript). We did not assume statistical significance implied significance in general, however we now explain that our

qualitative data within the questionnaire served as an internal check and validated what we observed statistically by respondents (Page 8). The wealth of qualitative data in literature already report on the general effects of mentoring on trainees (though quantitative data is lacking).

Reviewer 4: “The authors did discuss limitations of their data. However, it would have been beneficial to do so in a more structured way.“

Response 4.7: We have now amended our discussion (Page 19-21) to improve the discussions of the limitations of this study.

Reviewer 4: “Nonetheless, I think it was very premature that on the basis of the results for this study it was proposed that randomised controlled trials should be undertaken.”

Response 4.8: We understand Dr Sedgwick’s point of view. Though further research on this important topic is needed, we have now edited our manuscript to allow readers and other groups to consider how best to proceed given the challenges we faced in designing and conducting the study (Page 3, Page 19-21).

Reviewer 4: “The authors did make a mistake in the Discussion regarding the limitations and possible source of bias; on page 12 "There were also more trainees aged 31 to 35 years in the mentored group compared to the non-mentored group and this may have occurred by chance or response bias." it should read "non-response bias" and not "response bias".”

Reviewer 4.9: This text has now been removed for reasons described in response 1.2.

In summary, we are pleased that all four reviewers found our manuscript interesting and that it touches on an important but under-studied topic in UK postgraduate medical education. We have now made the changes recommended by the reviewers to improve the strength of our observations and the quality of our manuscript. We believe our quantitative findings are reinforced by our matched qualitative data and all are in keeping with the wealth of qualitative evidence in current literature. We also discuss the limitations of the study in greater length which has been influenced heavily by challenges in data collection, resource, ethics, and logistical dilemmas in gathering information from CMTs nationwide. The aim of this study was to assess quantitatively if a positive association existed between mentoring and better training outcomes in the UK. To our knowledge, this has not been previously attempted and this study is also the first UK-specific study that incorporates important and clinically relevant aspects such as MRCP(UK) and ARCP pass rates. Therefore we believe our study, which provides early data and a discussion of the challenges in research in this area, will be informative to other groups with similar interests within the UK. Once again we thank all involved for their time and input, and we look forward to your favourable reply.

VERSION 2 – REVIEW

REVIEWER	Colin Mitchell Imperial College Healthcare NHS Trust
REVIEW RETURNED	23-Mar-2018

GENERAL COMMENTS	This is a substantially improved paper and again, I applaud the attempt to investigate mentoring in a novel and rigorous way. Also thank you for the thorough response letter which made this process much more straightforward. My only remaining quibble would be the final paragraph, which I feel still slightly overstates the findings. For the reasons you now elucidate, you were not in fact able to

	demonstrate a positive association between mentoring and training outcomes - you were able to demonstrate an association between being in a mentorship program and training outcomes (ie it could have been the mentoring itself, or it could have been self-selection). I would also suggest the word 'demonstrate' in the final sentence would be better replaced with 'investigate' as this avoids the assumption of a causal link.

REVIEWER	Gregory Crawford University of Adelaide South Australia
REVIEW RETURNED	14-Mar-2018

GENERAL COMMENTS	Thank you for so comprehensively addressing all the suggestions.

REVIEWER	Colin Mitchell Imperial College Healthcare NHS Trust UK
REVIEW RETURNED	21-Dec-2017

GENERAL COMMENTS	This is an interesting topic and the intention to add quantitative findings to the evaluation of mentoring is admirable and potentially valuable. Mentorship seems likely to be a valuable tool in improving medical training but it is not without cost and investigating it rigorously is highly relevant. The investigators use a selection of reasonable and valid outcome measures to look for differences between two groups of trainees - those who have been mentored as part of a voluntary RCP scheme, and a normal group who have not been part of this mentoring program. Only trainees from each group who responded to a voluntary survey were included. Significant differences were found in exam pass rates and ARCP outcomes as well as markers of satisfaction. It is inherent in an observational study design that only correlation can be shown, not causation. The authors also rightly mention the inherent selection bias in their two groups. The intervention group have found out about, volunteered for, and participated in a mentorship program, while the other group did not. Clearly this inserts a huge potential for selection bias - the intervention group seem very likely to be self-selected as a highly engaged, enthusiastic, informed and committed group of trainees, which would be a powerful explanation for the resulting finding of their superior performance compared to the norm. This does not mean that the mentorship intervention was ineffective, but it does inject a large amount of doubt into their interpretation of the findings. In a study like this, particularly with such a potential for confounding, I would hope to see significant efforts made to ameliorate this problem, for example by showing that the groups were (prior to the intervention) similar in ways relevant to the outcome measures. Unfortunately there is very little such information, and that which is present seems to suggest that the groups are genuinely different - for example in terms of age (intervention group was seemingly older) and the number of international medical graduates (fewer in the intervention group). The findings would be more significant if it were known that the intervention and control groups had similar prior academic achievements or ARCP pass rates, for example, or if there was some form of time/dose response to mentoring (ie the more you are mentored the better you get) suggesting an effect from mentoring even if the groups are different at baseline.

	An alternative strategy would be to try to incorporate trainees who applied for the mentorship scheme but then did not participate, to see if their outcomes were closer to the mentored group (suggesting selection bias) or the unmentored group (suggesting the intervention was the cause of the findings). Unfortunately I suspect this data would be difficult to obtain and the numbers small so even if such data is available it may not help much. Ultimately, in my opinion, the selection limitation is far more significant than the authors seem to be suggesting and in fact makes the findings inconclusive, which is a shame as clearly this evaluation has been done with the best of intentions and the paper is well written. It is let down by the study design, or lack of mitigation for the design's inherent limitations. It may be that the authors have additional data on their subjects or can obtain it post-hoc, which might allow these issues to be addressed directly. Alternatively the paper could be re-written to acknowledge the significance of the limitation although I suspect this would make the conclusions less noteworthy.

REVIEWER	Dr Philp M Sedgwick Institute for Medical & Biomedical Education St. George's, University of London Tooting London SW17 0RE
REVIEW RETURNED	11-Apr-2018

GENERAL COMMENTS	Thank you for asking me to review the submission of the revised manuscript "The Association Between Mentoring and Training Outcomes in Junior Doctors in Medicine: An Observational Study". Many of the initial thoughts and concerns regarding the initial submission have been addressed. Whilst this is a generally well written manuscript, there are some points that the authors need to pay attention This is a questionnaire survey (i.e. observational study design). However, on Page 7 the authors refer to those trainees that received mentoring as "positive controls" and those trainees that did not as "negative controls". This is incorrect use of terminology as far as I am aware - such terms relate to participants in a clinical trial that receive the control treatment, with a positive control being an active drug and a negative one otherwise e.g. a placebo. Page 11 The paragraph "Statistical and Qualitative Analyses" is confused. In particular the second and third sentences are not accurate with regards the application of the Chi-Squared and Fisher's Exact test. These should be rewritten. It would also be useful, if only because it is standard practice, to indicate the critical level of statistical significance and indicate that traditional statistical hypothesis testing (two-sided alternative) was undertaken. The two groups of trainees (mentored and otherwise) were compared to the UK 2017 cohort of trainees on several variables (Figure 2 A and B). No details were provided as to the statistical tests that were used to achieve these comparisons - the groups are unlikely to be independent. Furthermore, pairwise comparisons were performed which were inappropriate since it would increase the probability of Type I errors. Several tables are referred to as Figures, and were often

	concatenated with graphs into a single figure so that an individual figure might consist of for example, one table and several figures. This seemed excessive and unnecessary, and possibly done to increase the number of tables and figures that could be incorporated into the manuscript. Despite the excessive number of tables and figures, the results of the logistic regression were not presented in a table (or at least they were not obvious). Such information is essential. Nonetheless, because of the small number of respondents it is not obvious how feasible a regression method would have been in terms of the accuracy of the estimates. The Discussion is interesting and generally well-balanced. Generally the authors tended to imply association and avoid inferring causality, this being an observational However, it is frustrating that the authors consistently assumed that the presence of statistical significance automatically inferred practical significance or importance. This is problematic since one cannot be inferred from the other, and in particular because of the large number of statistical tests the probability of Type I errors was high. There was no apparent adjustment for multiple hypothesis testing.

VERSION 2 – AUTHOR RESPONSE

We thank all four reviewers once again for their constructive feedback and suggestions for our manuscript. We have now performed a further revision to address the remaining points raised by each reviewer where appropriate. While we agree with most of the feedback given, we present our arguments in instances of disagreement and leave it up to readers to draw their own conclusions. Please see our responses below.

Reviewer 1: Dr Colin Mitchell, Imperial College Healthcare NHS Trust, UK

Reviewer 1: “My only remaining quibble would be the final paragraph, which I feel still slightly overstates the findings. For the reasons you now elucidate, you were not in fact able to demonstrate a positive association between mentoring and training outcomes - you were able to demonstrate an association between being in a mentorship program and training outcomes (ie it could have been the mentoring itself, or it could have been self-selection).”

Response 1.1: We would argue that we did demonstrate mentoring is associated with better outcomes and this was also clearly reported by the respondents who acknowledged they were active in the mentoring programme and that it contributed to their career progression etc. The only difference between the mentored and the non-mentored groups was that the aforementioned group received mentoring. We did not however investigate how mentoring was delivered in the mentoring programme as it is not within the scope of the research question posed. Although we mention selfselection as being a potential source of bias, there is no evidence that this is actually the case. If self selected respondents felt the effect of mentoring was negligible this would have been reflected in our qualitative results.

Reviewer 1: “I would also suggest the word 'demonstrate' in the final sentence would be better replaced with 'investigate' as this avoids the assumption of a causal link.”

Response 1.2: This has been changed accordingly, thank you.

Reviewer 2: Prof Gregory Crawford, Palliative Medicine, University of Adelaide, Australia

Reviewer 2: No further changes suggested.

Response: Thank you.

Reviewer 3: Ms Idaira Rodriguez Santana, Centre for Health Economics, University of York

Reviewer 3: "…the small sample size, the lack of information on prior academic achievements from the control and intervention groups, the selection bias and response bias make the quantitative findings from the paper less noteworthy. I'll leave to the editor the decision on whether the quantitative findings are worth of being published".

Response 3.1: There is no robust evidence that academic achievements in medical school correlates with career progression and clinical performance i.e. does being an average student in medical school make a junior doctor clinically inferior to a junior doctor with an undergraduate distinction? We urge caution in making such assumptions. As mentioned previously, we determined that the yield from extensively interrogating an academic history is not as beneficial as attaining an adequate sample size which Ms Rodriguez Santana has already eluded to. We have already acknowledged the response rates are low but this is typical of any nonincentivised and non-compulsory healthcare related survey.

Also, Ms Rodriguez Santana is incorrect in assuming that selection bias and response bias factually exist in our results. This is only demonstrable if both cohorts in their entirety complete the questionnaire and results are compared to our results. It is more accurate to describe the studies are at risk to such biases which we have already done so under "limitations of the study". We leave it up to readers to interpret the results and come to their own conclusion. Once again, we state that this study is first UK study to attempt to address the important issue of mentoring in British postgraduate medical training to such an extent. Furthermore, groups with more resources or that are better placed may find our study informative for planning future studies.

Reviewer 3: "Nonetheless, I still think that the qualitative findings are the most valuable output of this work and that need to be emphasised."

Response 3.2: We have added a further paragraph on our qualitative findings on Page 16-17.

Reviewer 3: "Authors could explore some relationships between the qualitative and quantitative findings. For example, they could test whether there are any differences in postgraduate outcomes between those doctors who reported 'negative' or 'positive' feedback"

Response 3.3: Mentees who report a negative experience in mentoring make up only 12% (3/25). Further sub-analyses would not contribute anything further to this paper due to this small sample size. The remaining doctors who reported a positive experience have a higher MRCP Part 1 pass rate compared to the 2017 UK average – we have now added this to our results section on Page 16-17.

Reviewer 3: "It is a common practice in quantitative studies to define the all variables included in the analysis, irrespectively of how obvious their meaning might be for the reader. Moreover, authors describe the meaning of a Significant Event (SE) in the results section (page 13). This should be done in advance and together with the description of the other variables included in the analysis."

Response 3.4: We now define "Significant Event" in our methods section on Page 7. The other variables included in the analysis have already been described on Page 10 and 11, thank you.

Reviewer 3: "Regarding the age variable, I still think that to get parsimonious model it would be a better idea to reduce the number of age categories (e.g. younger than 30 and older than 30)"

Response 3.5: We have now done so as advised. Please see Page 11 and updated Table 1.

Reviewer 3: “I cannot find the table with the results from the logistic regression. I think that authors should clarify what variables are included in the logistic regression”

Response 3.6: We have now included a logistic regression results table in our updated Table 3. Variables included in the logistic regression were previously mentioned on Page 11 of our manuscript.

Reviewer 3: “They should also give a point estimate of the effect of the mentoring programme on the three dependent outcomes, after controlling for age, gender, IMG, etc.”

Response 3.7: The logistic regression provided include OR values, for MRCP Part 1, MRCP Part 2 (written) and MRCP Part 2 (PACES). We now described these point estimates in our results section on Pages 13-14.

Reviewer 3: “The numbering of the groups in Figures 1 (B) and 2 (A) and (B) is not consistent. If mentoring group is defined as group (1) in the first figure, authors should stick to that categorization.”

Response 3.8: Our figures have now been revised and we have updated the numbering of our groups accordingly, thank you (see new Table 2).

Reviewer 4: Dr Philip M Sedgwick, School of Education, St. George’s University of London

Reviewer 4: “This is a questionnaire survey (i.e. observational study design). However, on Page 7 the authors refer to those trainees that received mentoring as "positive controls" and those trainees that did not as "negative controls". This is incorrect use of terminology as far as I am aware - such terms relate to participants in a clinical trial that receive the control treatment, with a positive control being an active drug and a negative one otherwise e.g. a placebo.”

Response 4.1: We respectfully disagree with Dr Sedgwick. A ‘positive control’ group is defined as a group, either in the basic or clinical sciences, where an intervention is administered and evaluated for its effect (compared to a negative control). The questionnaire was a means for us to collect information and the intervention administered was mentoring. However to avoid any confusion amongst readers, we have now changed the groups to "mentored group" and "control group" or "nonmentored group".

Reviewer 4: “Page 11 The paragraph "Statistical and Qualitative Analyses" is confused. In particular the second and third sentences are not accurate with regards the application of the Chi-Squared and Fisher's Exact test. These should be rewritten.”

Response 4.2: Thank you, this has now been re-written (Page 10).

Reviewer 4: “It would also be useful, if only because it is standard practice, to indicate the critical level of statistical significance and indicate that traditional statistical hypothesis testing (two-sided alternative) was undertaken.”

Response 4.3: We have now included this on Page 10.

Reviewer 4: “The two groups of trainees (mentored and otherwise) were compared to the UK 2017 cohort of trainees on several variables (Figure 2 A and B). No details were provided as to the statistical tests that were used to achieve these comparisons - the groups are unlikely to be independent.”

Response 4.4: We have now removed this comparison from Table 2. Response 4.5 also partly relates to this point.

Reviewer 4: “Furthermore, pairwise comparisons were performed which were inappropriate since it would increase the probability of Type I errors.”

Response 4.5: Pairwise comparisons were done because it was not possible to separate our cohort from the 2017 cohort completely and we believed the probability of Type I errors i.e. false positives, is still negligible e.g. MRCP Part 1 Pass rates: 21 mentored vs. 2065 total. For "theoretical completeness", we have now removed comparisons between subgroups and the 2017 cohort and let readers draw their own conclusions from the data provided. We have also changed chi squared pair wise comparisons to chi squared test for associations to reduce the number of hypothesis tests, see Table 1.

Reviewer 4: “Several tables are referred to as Figures, and were often concatenated with graphs into a single figure so that an individual figure might consist of for example, one table and several figures. This seemed excessive and unnecessary, and possibly done to increase the number of tables and figures that could be incorporated into the manuscript.”

Response 4.6: We had included more figures and tables after the first draft at the request of Reviewers 1 and 3 who asked for more data of our cohorts. We have now reviewed our figures and separated the tables from these. We have also provided one as a supplementary file. BMJ Open does not have any restriction on the number of figures we are able to submit.

Reviewer 4: “Despite the excessive number of tables and figures, the results of the logistic regression were not presented in a table (or at least they were not obvious). Such information is essential.”

Response 4.7: A logistic regression results table has now been added in Table 3, thank you.

Reviewer 4: “Nonetheless, because of the small number of respondents it is not obvious how feasible a regression method would have been in terms of the accuracy of the estimates.”

Response 4.8: We performed a logistic regression at the suggestion of Reviewer 3 in the previous review. Our statistician has recommended we present our logistic regression results as well.

Reviewer 4: “The Discussion is interesting and generally well-balanced. Generally the authors tended to imply association and avoid inferring causality, this being an observational However, it is frustrating that the authors consistently assumed that the presence of statistical significance automatically inferred practical significance or importance. This is problematic since one cannot be inferred from the other, and in particular because of the large number of statistical tests the probability of Type I errors was high.”

Response 4.9: We reiterate again that we did not assume statistical significance inferred practical significance based on our statistical results alone. In fact, we bring to attention that the positive association with mentoring observed within our quantitative data is similar to our qualitative data and also the extensively reported qualitative data in current literature. In addition, we also highlight the size of the effect of mentoring on MRCP part 1 exam pass rates, where the estimated effect was large and most obvious (OR=9.56). Our stance is that the quantitative data is congruent with these other points of reference and not a standalone observation from which we made an inference. Our statistician is in agreement.

Reviewer 4: “There was no apparent adjustment for multiple hypothesis testing.”

Response 4.10: Our hypothesis tests are exploratory and we therefore we did not consider adjusting for multiple testing to be necessary. Our approach is supported by other studies such as Rothman K (now cited as reference 15) which reported that making adjustments for multiple comparisons can lead to an increased number of errors of interpretation when data being evaluated are actual observations. We have now included this justification in the manuscript on Page 10.

VERSION 3 – REVIEW

REVIEWER	Colin Mitchell Imperial College NHS Healthcare Trust UK
REVIEW RETURNED	25-May-2018

GENERAL COMMENTS	Thanks again for your responsiveness. There is obviously some debate as to the significance and meaningfulness of your findings but given the inherent limitations of the project you have done good work to get them to a publishable level. Despite the accepted limitations, I agree this is a useful step in progressing the literature relating to a (probably) important intervention.

REVIEWER	Idaira Rodriguez Santana Centre for Health Economics, University of York
REVIEW RETURNED	04-Jun-2018

GENERAL COMMENTS	The authors have done a great job addressing all the comments from the reviewers. I think that now the limitations of the study are clearly stated and there is sufficient information for readers to judge the reliability of the findings. Just a minor comment regarding the logistic regression table. In order to interpret the results, authors need to clarify which is the omitted outcome when the explanatory variable is categorical (i.e. age and primary degree). This could be done as a footnote in the regression table.

REVIEWER	Dr Philip M Sedgwick St. George's, University of London, Cranmer Terrace, London SW17 0RE United Kingdom
REVIEW RETURNED	29-Jun-2018

GENERAL COMMENTS	This is an interesting study that explores an important area. Thank you for asking me to review the manuscript. I believe this manuscript has now undergone two major revisions following peer review, and this is third review, with each stage involving several reviewers. It is greatly improved as a result of that feedback. This now presents the potential ethical dilemma as to whether the manuscript represents original work of the authors. In particular my feedback has focused on study design, research methods, plus statistical analysis; the original manuscript was deficient in these areas indicating lack of appreciation and understanding. In my previous review I raised the need to undertake some form of regression analysis to take account for potential confounding, whilst also adjusting P-values to minimise the probability of Type I errors occurring through multiple testing. The former has been addressed (in part) whilst the later has not at all. In my opinion this is a serious

omission that affects the validity of the study conclusions.

The authors' rationale for not undertaking adjustment for potential Type I errors is based on a paper by Rothman in 1990. In particular, they claim that since this is an exploratory study it would not be wise to do so since it may miss potential associations that could be explored in a larger study. I am not sure I agree with that rationale. Rothman's paper is an old one; more recently concern has been widely expressed about null hypothesis statistical testing and the validity of scientific findings. If the researchers wish to treat their study as an exploratory study then they should not treat hypothesis testing as a recipe that can be simply followed. Most, if not all of the results upon which the study conclusions are based seem to originate from bivariate statistical tests analysis, and therefore do not take into account potential confounding. This is problematic in itself. The significant P-values are "abbreviated" to "P<0.05", "P<0.01", or "P<0.001". I am not sure if this is journal style but in my opinion it should avoided, and the exact or asymptotic P-values presented not least since it allows the reader to assess the statistical strength of any association. Nonetheless, the statistically significance results upon which the main conclusions are based are all presented (I believe) as "P<0.05"; hence the exact or asymptotic P-value is between 0.01 and 0.05. Because of the number of statistical tests performed such statistical significance is therefore most likely to be a Type I error, and hence this study would not demonstrate any positive findings if adjustments were made. To avoid any adjustment for potential Type I errors, and represent the conclusions as shown in the abstract, is misleading.

Conclusions (Abstract): The authors conclude that "Further studies are needed to investigate the causative effects of mentoring in postgraduate medical training within the UK." This is misleading since it is not possible to infer causation from this study i.e. a questionnaire (observational) design.

Results
Page 16, Paragraph 3
The authors state "Logistic regression demonstrated that age and the country of primary qualification did not have any significant influence on the effects observed in mentoring for all components of the MRCP(UK) exam." I presume that authors are suggesting that the effects of mentoring are independent of age and country of primary qualification (having undertaken a logistic regression analysis). However, based on the table presented (Table 3), the reader cannot conclude this since it requires presentation of both the unadjusted and adjusted odds ratios.

Page 16, Paragraph 4
The authors state "The ARCP review provides a comprehensive assessment of a trainee's progress in the core medical training educational curriculum and personal clinical practice. In our study, 97 trainees (24 mentored, 73 non-mentored) out of 110 had an ARCP within 12 months. The ARCP pass rate (Outcome 1s) was observed to be significantly higher in mentored trainees compared to non-mentored trainees; 95.8% (23/24) vs. 69.9% (51/73), p<0.05 (OR=9.9, 95% CI 1.5 - 107 and RR=1.4, 95% CI 1.1 - 1.7)."

I would suggest that the OR presented in the final sentence is incorrect. I calculate it to be 1.37, and not 9.9. However, it is not obvious (and no reasons were given) as to why both the RR and OR

	were presented. Minor Points There is a tendency to change tense throughout the entire script. At times the script would benefit from greater attention to punctuation in order to enhance reading. At times words are apparently missing and the text does not make intuitive sense. For example in the paragraph headed “Exclusion criteria” (Page 10, Paragraph 1), the first sentence reads “The first half of the survey collected demographic data therefore surveys with less than 50% of answered questions were not interpretable.”

VERSION 3 – AUTHOR RESPONSE

Thank you for your recent correspondence regarding our manuscript and your acceptance of it for publication in BMJ Open.

We have now added a footnote to our logistic regression table as per Reviewer 3's recommendation and this can be found on page 14. We believe all points raised by the reviewers have been fully addressed now, barring Reviewer 4's comments for previously stated reasons.

Once again, we thank you, the editor and all those of have contributed their time to this manuscript.