PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Psychosocial Work Characteristics, Burnout, Psychological Morbidity Symptoms and Early Retirement Intentions: A Cross-sectional Study of NHS Consultants in the United Kingdom.
AUTHORS	Khan, Atir; Teoh, Kevin; Islam, Saiful; Hassard, Juliet

VERSION 1 – REVIEW

REVIEWER	Charlotte N. L. Chambers Principal analyst, The Association of Salaried Medical Specialists, New Zealand
REVIEW RETURNED	11-Sep-2017

GENERAL COMMENTS	This is an interesting study which adds to the literature on the relationship between burnout, stress and anxiety and job autonomy for doctors working in a public health setting. The article is, on the whole, reasonably well-written but would benefit from revisions which are specified below. The statistics look sound but as I am not a statistician, I am unable to comment specifically on this aspect of the study. To assist with the readability of the article, I would recommend a more careful explication of the relationship between burnout, work-related stress and depressive and anxiety symptoms (collectively referred to as psychological morbidities) throughout the article (see further comments below). Also, the implications of these findings could be made a little more nuanced and specific eg. there could be a reflection on what the demographic differences in results may suggest in terms of further work and/or research. These demographic differences probably warrant greater consideration in the discussion section given the detail provided in the results. Specific points as follows: Use of term 'burnout' in title differs in spelling to 'burn-out' in key words. Would be helpful to check for consistency throughout. Spelling mistake 'strengths of this study' (page 3 line 3) it would be helpful for non UK readers to have NHS defined at first use. It might also be helpful to explain what a 'consultant' is in the context of the UK health system. In the introduction, the aims of the study are noted as 'testing the role of job autonomy and burnout' (page 4 line 22) and 'attempts to test' (page 4 line 54). The objectives and discussion refer to 'examine the associations' (page 2 line 6) and 'examine the link' (page 14 line 48). It might be more accurate use the term examine

	rather than test in framing the aims of the study? In the introduction (page 4 line 7) the term psychiatric morbidity is used and then the remainder of the article refers to psychological morbidities (starting line 15). It would be helpful to have a definition of the latter and, if necessary, how this differs from the former. Furthermore, it would be helpful for the non-expert to understand how the authors understand the relationship between psychological morbidities and burnout. I found the terms were at times used interchangeably, particularly in explaining the results (pg 9) where the dimensions of the maslach burnout inventory are discussed under the subheading of psychological morbidity (albeit alongside the other measures of stress and anxiety). However, in the discussion section (pg 15 line 47 onwards) burnout was discussed as a separate phenomenon albeit one which is significantly associated. My view is that the article would benefit from a clearer explication of these terms and how they inter-relate and that this would strengthen the understanding of the significance of the findings overall. page 7 'measures': is it necessary to have the section in bullet points? page 7 line 41; 'cut-off scores from norm scores' what are norm scores? Can the cut-points be specified please? Page 9 line 30 should it be the 'smallest proportion' of specialists rather than fewest? Capitalisation of specialities is inconsistent eg. page 9 line 35 pathology/micro is in lower case, but caps elsewhere. page 15 line 8 'and is convenience sample'; should this be ''and is 'a' convenience sample'"? page 15 line 13 and 14; this sentence is a bit repetitive; could it be re-worked? In the results, the significant differences by location (eg. Wales) medical speciality and gender are interesting. There is a considerable literature on the rates of burnout in women and why these may be higher than their male counterparts as well as some speciality specific burnout studies. While this is ancillary to the core aims of the article, the demographic differences should warrant some mention in the discussion section. These differences may also provide pointers for what targeted interventions may be required to attend to these findings in the implications and conclusions section. There are some inconsistencies in how the references are presented in the reference list eg. references 17 and 30. would recommend checking against and correcting as per the BMJ open reference style guide. Not all have doi's specified. Also an error in page 21 line 35 (Kivim??ki). There are some additional more recent studies on burnout which could be cited. eg Dewa CS, Loong D, Bonato S, et al The relationship between physician burnout and quality of healthcare in terms of safety and acceptability: a systematic review BMJ Open 2017;7:e015141. doi: 10.1136/bmjopen-2016-015141

REVIEWER	Scott McCain Queen's University Belfast and Belfast City Hospital, Northern

	Ireland
REVIEW RETURNED	22-Sep-2017

GENERAL COMMENTS	1. The objective could be more clearly stated, the terminology is somewhat confusing for the reader. I understand what the authors are examining but it could be described more clearly. 2. The abstract is for the most part, clear but I question the need for such detail in the results section to the extent of reporting all regression co-efficients, CIs and p-values. In the results section it would be perhaps better to describe demographic data prior to discussion of the regression models. 4. Methods- under the study design heading it is unclear what the second sentence "a simple random sample of 500...." is in regard to. the comment on blinding at the end of this section is irrelevant to the study design. There is no comment on the denominator of the study and in my opinion this is the major limitation of this study, it is unclear if this is a true representation of UK doctors- it is unlikely if this survey was sent to all UK consultants. This is not clearly reported. 6. The outcomes and how they are measure have been described. There has been some selection of components of the outcome measurement tools instead of using the complete tools. The justification for this is unclear. The authors have made a somewhat arbitrary decision about cut-offs for categorising patients as "high", I do not think this is useful and the outcome would be better reported as a continuous variable with mean/medians etc as necessary. 7. I am unclear about the multivariable regression models. I have several queries that would be best addressed by a formal statistical review. The regression models appear to be presented as a multivariable model with adjusted and standardised co-efficients. It would be better presented as univariate and multivariate co-efficients to aid reader understanding. I do not see a great deal of benefit in including speciality and country of employment within the models, and am uncertain about the methodology of including physicians as the reference group and scotland as the reference group. In table 1 percentages are reported- it is unclear what these percentages refer to. I think they refer to the proportion of doctors exceeding their arbitrary values for "high" levels of the outcome in question. it would be better to report the continuous results for each variable and a more full explanation of what exactly is being described. 14. my only concern would be their plans for a further paper reporting retirement plans of consultants, I would suggest that it may be better to consolidate both these papers as it is likely to involve significant duplication of reporting. Overall the topic is interesting and relevant with the current interest in burnout among doctors. My biggest concern would be with regard to the omission of a denominator in the reporting of the results/methods. Statistical review would be beneficial and alteration of how the results are reported may be necessary.

REVIEWER	Daniel Schwarzkopf Jena University Hospital, Germany
REVIEW RETURNED	26-Sep-2017

GENERAL COMMENTS	In this cross sectional survey study the relationships between job autonomy, burnout, and indicators of mental health are investigated

	among a sample of NHS consultants. A strength of the study seems to be that the full population of consultants in the NHS was invited to participate (this is not really clear from the methods section) and a substantial number of questionnaires were completed. The major limitation of the study is that it provides little new insights by replicating well established relations between concepts of work-stress. I think the results as presented in the current manuscript are of too little interest to an international audience like the audience of the BMJ open. Further comments: Abstract Participants: It should be stated who was invited to participate. Information on actual participants should be placed in the results (at least that is the convenience in medical journals). Methods: some information on conducted analyses should be given. Introduction Literature: the order of citations in the text is not correct (e.g. 10 follows on 7). I don't think that there exists a knowledge gap on the relation between job autonomy, work stress burnout and mental health problems caused by job stress. I guess that hundreds of studies have studies these relations, dozens among medical staff. To study this relation in one specific population of physicians does not seem to be a research question that is of greater interest to an international audience. To thoroughly study aspects of the work environment using several scales (workload, social support, quality of work relationships et.c) to identify predictors of burnout among NHS physicians would have been of greater interest for UK policy makers. Relationships between burnout and mental health problems have been studied for a long time (compare Schaufeli & Enzman, 1998, who give a summary of the literature to that point). There is no Model or Framework presented for the study (e.g. a Framework of Work stress among physicians that would be the theoretical justification for the investigated relations). Based on existing knowledge, the study aims lack novelty. Methods Study design: "A simple random sample of 500"- Has a power calculation been conducted before the study? How is the n of 500 justified? It is not clear how the sample was obtained- was the full population of NHS consultants invited to participate? Than the sample is not actually a random sample of 500 but is based on non-participation of the initial population. This should be described in more detail. Data on participants (number and demographics) should be presented in the results. Also in the results, the participation rate should be given (what was the initial sample? How many consultants were invited?). Measures: Job autonomy: why was only one single scale studied? There exist numerous aspects of the work environment that have been studied among physicians previously. There are other important topics like workload, leadership, collaboration etc. It would have been interesting to identify important aspects of the work environment that predict negative outcomes. State Trait Personality: Why was this measure chosen? Is it a clinical screening tool? Which scales were used? (state vs. trait- only

	anxiety and depression or also other scales?) Occupational Stress Indicator: This is not a simple measure of work stress, but involves several scales both on stressors, personality factors, coping strategies and stress-outcomes. Based on the “22 items” the authors seem to have used the “Job Satisfaction” subscale. Were no other scales of the OSI used? Why not? The authors should name the concept clearly as “job satisfaction” and not work stress. It is not clear, if this scale should be used as one of the outcomes. E.g. Ramirez et al. 1996 used items on job satisfaction to predict mental health. This is another reason why the authors should present a clear framework for their study. Results Participation rates should be given here- have there been differences in participation (e.g. between England, Wales, and Scotland?). The authors interpret subgroup differences in outcomes without checking for significance- thereby mere random differences might be interpreted. Why are nonparametric correlations used? Since the variables are then used in regression analyses they seem to be nearly normally distributed- use Pearson correlations instead to allow comparability to regression analyses. Regression models: for categorical variables an overall test of significance (F-Test) should be given. It might be more appropriate to present standardized solutions (z-transformed continuous variables) since for analyses of questionnaire data these provide results that could be interpreted in terms of standard deviations. Discussion “convenience sample” (p. 15 line 8)- if you invited all consultants you don’t have a convenience sample but the full population. Non-participation is no sampling strategy but part of the limitations! The authors should provide examples of how job autonomy could be increased for consultants. In theory, consultants already have a high level of job autonomy compared to other jobs. What might explain differences in job autonomy between consultants?

REVIEWER	Naveh, Eitan Technion Israel Inst Technol, IE&M
REVIEW RETURNED	01-Oct-2017

GENERAL COMMENTS	The study aimed to examine the associations between job autonomy and burnout in relation to self-reported work-related stress and depression and anxiety symptoms; and then, to examine the current level of psychological morbidity among a sample of hospital consultants in the NHS. Theoretical contribution. There are many works about burnout and about autonomy. The authors argue that this is the first study to put forward the relationship they suggest between autonomy and burnout. The bottom line conclusion that autonomy is in general an influential factor is not surprising. However, some studies also suggest the existence of negative aspects of autonomy (Stern, Katz-Navon, & Naveh, 2008). The authors seems to ignore this line of work. In this respect, my most important reaction is that I would expect a more complex and comprehensive approach that includes the positive as well as negative aspects of autonomy. Clearly, the effects of autonomy are more positive under certain conditions, while others may emphasize its negative results. It is a little naïve to claim

	that autonomy has a significant main effect in all conditions (either positive or negative). It also does not reflect an updated knowledge about autonomy. I think that a major limitation of the paper derives from a lack of supporting theoretical base. I suggest providing the readers with a short description of the state-of-the-art knowledge in the field. This would allow the authors to see the difficulty the reader has in understanding their unique contribution to the advancement of said knowledge. A conclusion such as "improving consultants' job autonomy, and providing support and resources to prevent burnout" is not surprising and represents a narrow approach to the problem. Investing in additional resources is not a very helpful suggestion. Does the suggested investment come at the expense of other possible types of investment? What are the alternatives? Why should policymakers adapt this approach? I do not think this study provides them with a solid base that allows them to conclude that a higher investment in certain resources is conducive to solving the specific problem at hand. Methods 1. I do not find a clear statement of the response rate. 2. It is not clear to what extent the authors used accepted measurement scales, for example in the case of autonomy. They used a three-item factor. A clear explanation of their selection of measurement tools is required. Analyses From the description provided by the authors it is not clear whether they took into consideration the hierarchal nature of their data, i.e., individuals within departments within hospitals. The hierarchical structure of the data must be taken in to account and explained if the findings are to be based on a solid foundation and accepted. This is not only an issue of analytical method, but also represents a correct theoretical approach to the manner in which individuals operate within their environment.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Reviewer Name: Charlotte N. L. Chambers
Institution and Country: Principal analyst, The Association of Salaried Medical Specialists, New Zealand Please state any competing interests: None declared

Please leave your comments for the authors below
This is an interesting study which adds to the literature on the relationship between burnout, stress and anxiety and job autonomy for doctors working in a public health setting. The article is, on the whole, reasonably well-written but would benefit from revisions which are specified below. The statistics look sound but as I am not a statistician, I am unable to comment specifically on this aspect of the study.

To assist with the readability of the article, I would recommend a more careful explication of the relationship between burnout, work-related stress and depressive and anxiety symptoms (collectively referred to as psychological morbidities) throughout the article (see further comments below). Also,

the implications of these findings could be made a little more nuanced and specific eg. there could be a reflection on what the demographic differences in results may suggest in terms of further work and/or research. These demographic differences probably warrant greater consideration in the discussion section given the detail provided in the results.

Authors' response: Thank you for your comments and feedback. We have revised our manuscript and taken your points on board. We address your individual comments below and hope you find them acceptable.

Specific points as follows:
Use of term 'burnout' in title differs in spelling to 'burn-out' in key words. Would be helpful to check for consistency throughout.
Authors' response: We have reviewed the article for consistency of terms and have used 'burnout' instead.

Spelling mistake 'strengths of this study' (page 3 line 3)
Authors' response: We have corrected this.

it would be helpful for non UK readers to have NHS defined at first use. It might also be helpful to explain what a 'consultant' is in the context of the UK health system.
Authors' response: In the abstract and in the opening paragraph of the introduction we have now introduced the term "National Health Service" before using its acronym. We have also provided some explanation as to who consultants are and why they are important.

In the introduction, the aims of the study are noted as 'testing the role of job autonomy and burnout' (page 4 line 22) and 'attempts to test' (page 4 line 54). The objectives and discussion refer to 'examine the associations' (page 2 line 6) and 'examine the link' (page 14 line 48). It might be more accurate use the term examine rather than test in framing the aims of the study?
Authors' response: We have revised our aims and believe that is also enhances its clarity. As per your suggestion we use the term 'examine' instead of 'test'.

In the introduction (page 4 line 7) the term psychiatric morbidity is used and then the remainder of the article refers to psychological morbidities (starting line 15). It would be helpful to have a definition of the latter and, if necessary, how this differs from the former. Furthermore, it would be helpful for the non-expert to understand how the authors understand the relationship between psychological morbidities and burnout. I found the terms were at times used interchangeably, particularly in explaining the results (pg 9) where the dimensions of the maslach burnout inventory are discussed under the subheading of psychological morbidity (albeit alongside the other measures of stress and anxiety). However, in the discussion section (pg 15 line 47 onwards) burnout was discussed as a separate phenomenon albeit one which is significantly associated. My view is that the article would benefit from a clearer explication of these terms and how they inter-relate and that this would strengthen the understanding of the significance of the findings overall.
Authors' response: For consistency our revised manuscript uses the term 'psychological morbidities' to collectively refer to depressive and anxiety symptoms. Our revised manuscript focuses on the role of burnout as a mediator and we hope that our introduction (together with Figure 1) makes the relationship between burnout and psychological morbidities clearer.

page 7 'measures': is it necessary to have the section in bullet points?
Authors' response: We have revised and done away with bullet points.

page 7 line 41; 'cut-off scores from norm scores' what are norm scores? Can the cut-points be specified please?

Authors' response: We have specified what the exact cut –off scores are.

Page 9 line 30 should it be the 'smallest proportion' of specialists rather than fewest?
Authors' response: This has been corrected.

Capitalisation of specialities is inconsistent eg. page 9 line 35 pathology/micro is in lower case, but caps elsewhere.
Authors' response: We have corrected this.

page 15 line 8 'and is convenience sample'; should this be ''and is 'a' convenience sample'"?
Authors' response: Our revisions mean that this phrase is no longer in the revised manuscript.

page 15 line 13 and 14; this sentence is a bit repetitive; could it be re-worked?
Authors' response: This section has been revised.

In the results, the significant differences by location (eg. Wales) medical speciality and gender are interesting. There is a considerable literature on the rates of burnout in women and why these may be higher than their male counterparts as well as some speciality specific burnout studies. While this is ancillary to the core aims of the article, the demographic differences should warrant some mention in the discussion section. These differences may also provide pointers for what targeted interventions may be required to attend to these findings in the implications and conclusions section.
Authors' response: We have made amendments to the Practical Implications section to highlight the individual issues.

There are some inconsistencies in how the references are presented in the reference list eg. references 17 and 30. would recommend checking against and correcting as per the BMJ open reference style guide. Not all have doi's specified. Also an error in page 21 line 35 (Kivim??ki).
Authors' response: We have reviewed and revised our references. Where papers have DOIs these have been included.

There are some additional more recent studies on burnout which could be cited. eg Dewa CS, Loong D, Bonato S, et al. The relationship between physician burnout and quality of healthcare in terms of safety and acceptability: a systematic review. BMJ Open 2017;7:e015141. doi: 10.1136/bmjopen-2016-015141
Authors' response: In revising this manuscript we have updated our reference to include more contemporary references, including the one by Dewa et al. (2017). However, much of the research involving UK consultants is dated, highlighting the relevance of this study.

Reviewer: 2
Reviewer Name: Scott McCain
Institution and Country: Queen's University Belfast and Belfast City Hospital, Northern Ireland Please state any competing interests: None declared

Please leave your comments for the authors below 1. The objective could be more clearly stated, the terminology is somewhat confusing for the reader. I understand what the authors are examining but it could be described more clearly.
Authors' response: We have re-written the introduction and discussion and with your comment in mind believe it is much clearer and consistent now.

2. The abstract is for the most part, clear but I question the need for such detail in the results section to the extent of reporting all regression co-efficients, CIs and p-values. In the results section it would be perhaps better to describe demographic data prior to discussion of the regression models.

Authors' response: We have revised the results section in the abstract, including removing the coefficients. The results section begins with a brief description of the demographic data, with reference to Table 1.

4. Methods- under the study design heading it is unclear what the second sentence "a simple random sample of 500...." is in regard to the comment on blinding at the end of this section is irrelevant to the study design.
Authors' response: We have updated the sample size calculation sentences, and have removed the term simple random sample. Unfortunately, we did not have the scope to know the total number of consultants approached in this survey. This is because as we adopted an electronic survey tool which was forwarded onto consultants through their respective health board and/or trusts. Therefore, we are unable to determine how many consultants actually received invitations to participate. Consequently, our calculation is based on an assumption on the total number of consultants approached and assumed response rate, with margin of error and employing the principle of finite population correction. As this is not an intervention study and we are not evaluating or comparing any intervention between groups, a proper statistical power analysis doesn't fit here nor is it useful.

There is no comment on the denominator of the study and in my opinion this is the major limitation of this study, it is unclear if this is a true representation of UK doctors- it is unlikely if this survey was sent to all UK consultants. This is not clearly reported.
Authors' response: Our revision now makes it clear that we approached consultants though HR department of 7 health boards in Wales, 10 health board in Scotland and 12 trusts in England. However, as we were not able to directly invite consultants to participate we are not able to determine the total number of consultants approached in this survey. Therefore, the our calculation of suitable sample size is based on some assumption on the total number of consultants approached and assumed response rate, with margin of error and employing the principle of finite population correction. We acknowledge this is a limitation and have made this clear in the manuscript (please note the section "strengths and limitations of this study" – after abstract).

6. The outcomes and how they are measure have been described. There has been some selection of components of the outcome measurement tools instead of using the complete tools. The justification for this is unclear. The authors have made a somewhat arbitrary decision about cut-offs for categorising patients as "high", I do not think this is useful and the outcome would be better reported as a continuous variable with mean/medians etc as necessary.
Authors' response: We have now elaborated in our measures section the questionnaires where the measures were obtained from. Here we have also explained how we separated consultants who scored "high" on emotional exhaustion, depersonalisation, and the number of depressive and anxiety symptoms. This includes stating what the exact cut-off score is. Our analysis is based on continuous variables, with means and standard deviation reported in Table 1.

7. I am unclear about the multivariable regression models. I have several queries that would be best addressed by a formal statistical review. The regression models appear to be presented as a multivariable model with adjusted and standardised co-efficients. It would be better presented as univariate and multivariate co-efficients to aid reader understanding. I do not see a great deal of benefit in including speciality and country of employment within the models, and am uncertain about the methodology of including physicians as the reference group and scotland as the reference group.
Authors' response: We have now provided the standardised and unstandardized coefficients of the Multiple Linear regression models along with the statistical significance and confidence interval. These are multivariate coefficients. The univariate coefficients can be calculated if we separately check the relationship of the response variable with each of the independent variable using any statistical test of hypothesis. However, these will not reveal any useful information as we are presenting the results of each of the independent variable adjusting the other covariates. Rather it will

create more confusion if we present those results under MLR and separately presenting them will beyond the scope of the limited space allowed for the BMJ.

It is standard in any multivariate analysis that not all the independent variable (either continuous or categorical by nature) will significantly influence the response variable. This is what happened in this case. Please note that speciality and country of employment are categorical variables in the MLR model and hence the introduction of dummy variable and respective reference category. These two categorical variables are important to answer the questions whether the outcome measures vary with respect to the speciality of the consultant and/or their place of work adjusting for other included factors in the model.

We use the standard methodology of categorisation of the dummy variable in Multiple Linear Regression where each of category needs to be compared with the largest group which is considered as the reference category. Following the initial exploratory results presented under the Table 1, it is observed that the Physicians and Scotland are the largest group respectively in speciality and country, hence considered as the reference category.

In table 1 percentages are reported- it is unclear what these percentages refer to. I think they refer to the proportion of doctors exceeding their arbitrary values for "high" levels of the outcome in question. it would be better to report the continuous results for each variable and a more full explanation of what exactly is being described.
Authors' response: As indicated above, we have elaborated in our measures section the questionnaires where the measures were obtained from. Here we have also explained how we separated consultants who scored "high" on emotional exhaustion, depersonalisation, and the number of depressive and anxiety symptoms. This includes stating that the exact cut-off score is and the source for doing so. Our analysis is based on continuous variables, with means and standard deviation reported in Table 1.

My only concern would be their plans for a further paper reporting retirement plans of consultants, I would suggest that it may be better to consolidate both these papers as it is likely to involve significant duplication of reporting.
Authors' response: In revising our paper we have taken your comment on board and have included early retirement intention as an outcome variable in this study.

Overall the topic is interesting and relevant with the current interest in burnout among doctors. My biggest concern would be with regard to the omission of a denominator in the reporting of the results/methods. Statistical review would be beneficial and alteration of how the results are reported may be necessary.
Authors' response: We thank you for your comments and hope that the amendments that we have made are satisfactory to you.

Reviewer: 3
Reviewer Name: Daniel Schwarzkopf
Institution and Country: Jena University Hospital, Germany Please state any competing interests: None declared

Please leave your comments for the authors below In this cross sectional survey study the relationships between job autonomy, burnout, and indicators of mental health are investigated among a sample of NHS consultants. A strength of the study seems to be that the full population of consultants in the NHS was invited to participate (this is not really clear from the methods section) and a substantial number of questionnaires were completed. The major limitation of the study is that it provides little new insights by replicating well established relations between concepts of work-stress. I

think the results as presented in the current manuscript are of too little interest to an international audience like the audience of the BMJ open.
Authors' response: Thank you for your feedback. We hope that our revised manuscript addresses your concerns. As this revised manuscript indicates we did not survey the entire NHS consultant workforce and instead surveyed a number of health boards. We have responded to your specific comments below and hope that you find them satisfactory.

Further comments:
Abstract
Participants: It should be stated who was invited to participate. Information on actual participants should be placed in the results (at least that is the convenience in medical journals).
Authors' response: We have clarified our recruitment process. The information on actual participants is presented at the start of the results section, and in Table 1 (also in the results).

Methods: some information on conducted analyses should be given.
Authors' response: We have included a sentence at the start of the results to clarify this.

Introduction
Literature: the order of citations in the text is not correct (e.g. 10 follows on 7).
Authors' response: We have revised and updated our references. All references are listed in numerical order in-text.

I don't think that there exists a knowledge gap on the relation between job autonomy, work stress burnout and mental health problems caused by job stress. I guess that hundreds of studies have studies these relations, dozens among medical staff. To study this relation in one specific population of physicians does not seem to be a research question that is of greater interest to an international audience. To thoroughly study aspects of the work environment using several scales (workload, social support, quality of work relationships et.c) to identify predictors of burnout among NHS physicians would have been of greater interest for UK policy makers.
Authors' response: We have broadened and made our model clearer. This is to look at the role of burnout as a mediator between job autonomy and working pressure on one hand, with outcomes on depressive symptoms, anxiety symptoms and early retirement. This extends the strain-health relationship to help inform the pathways between working conditions and mental health symptoms. We hope that our revised introduction captures the existing gap in research.

Relationships between burnout and mental health problems have been studied for a long time (compare Schaufeli & Enzman, 1998, who give a summary of the literature to that point).
Authors' response: We were not able to obtain a copy of the book you have referenced. We do not dispute the evidence base between burnout and mental health issues. However, although many postulate that burnout is a mediator between working conditions and mental health issues, there have been few attempts to test this – particularly in the health services and among consultants/ senior doctors.

There is no Model or Framework presented for the study (e.g. a Framework of Work stress among physicians that would be the theoretical justification for the investigated relations).
Authors' response: We hope our revised introduction, including Figure 1, makes our model clearer.

Based on existing knowledge, the study aims lack novelty.
Authors' response: We have revised our study aims and introduction based on this comment.

Methods
Study design:

“A simple random sample of 500”- Has a power calculation been conducted before the study? How is the n of 500 justified? It is not clear how the sample was obtained- was the full population of NHS consultants invited to participate? Than the sample is not actually a random sample of 500 but is based on non-participation of the initial population. This should be described in more detail.
Authors’ response: We have now updated the sample size calculation sentences and removed the reference to simple random sample. We do not have the scope to know the total number of consultants approached in this survey as we adopted an electronic survey tool and tried to reach all the consultants through respective health board and/or trusts. Therefore, the calculation is based on some assumption on the total number of consultants approached and assumed response with margin of error and employing the principle of finite population correction. We believe the reviewer understands that this is not an interventions study and we are not evaluating or comparing any intervention between groups. Therefore, a proper statistical power analysis doesn’t fit here and it therefore not useful.

Data on participants (number and demographics) should be presented in the results. Also in the results, the participation rate should be given (what was the initial sample? How many consultants were invited?).
Authors’ response: We did not collect any additional demographics other than age, gender and country. The results of these demographics are presented in Table 1 and are described in the first paragraph under the result section.
The results presented here are based on the initial responded sample of the consultants. We approached the consultants though HR department of 7 health boards in Wales, 10 health boards in Scotland and 12 trusts in England via an electronic survey tool. Therefore, we are not able to determine the total number of consultants approached or invited in this survey. The calculation is based on some assumption of the total number of consultants approached and assumed response rate, with margin of error and employing the principle of finite population correction. We hope that our revisions in the manuscript make this clearer.

Measures: Job autonomy: why was only one single scale studied? There exist numerous aspects of the work environment that have been studied among physicians previously. There are other important topics like workload, leadership, collaboration etc. It would have been interesting to identify important aspects of the work environment that predict negative outcomes.
Authors’ response: We have included a second measures that represents job demands – working pressure. For brevity we did not include a wide range of measures of the work environment into our data collection process. However, as in this revised manuscript we are focusing on burnout as a mediator we believe that too many measures of working conditions would increase the complexity of the study.

State Trait Personality: Why was this measure chosen? Is it a clinical screening tool? Which scales were used? (state vs. trait- only anxiety and depression or also other scales?) Occupational Stress Indicator: This is not a simple measure of work stress, but involves several scales both on stressors, personality factors, coping strategies and stress-outcomes. Based on the “22 items” the authors seem to have used the “Job Satisfaction” subscale. Were no other scales of the OSI used? Why not? The authors should name the concept clearly as “job satisfaction” and not work stress. It is not clear, if this scale should be used as one of the outcomes. E.g. Ramirez et al. 1996 used items on job satisfaction to predict mental health. This is another reason why the authors should present a clear framework for their study.
Authors’ response: We have revised our section on measures to clarify these points. We accept that the initial work stress measure was the “job satisfaction” subscale on the OSI. However, based on the revised model in this paper we have removed this measure altogether.

Results

Participation rates should be given here- have there been differences in participation (e.g. between England, Wales, and Scotland?).
Authors' response: This is unfortunately not possible as we were not able to determine response rate as described above.

The authors interpret subgroup differences in outcomes without checking for significance- thereby mere random differences might be interpreted.
Authors' response: We accept this point and have included a sentence making it clear that these are descriptive comparisons.

Why are nonparametric correlations used? Since the variables are then used in regression analyses they seem to be nearly normally distributed- use Pearson correlations instead to allow comparability to regression analyses.
Authors' response: We have now employed the Pearson correlation coefficient. There is a controversy in social statistics whether Likert scale should be treated as nonparametric in bivariate relationships. However, we accept that it is usual practice in the social sciences to use Pearson correlation coefficients for Likert scales.

Regression models: for categorical variables an overall test of significance (F-Test) should be given. It might be more appropriate to present standardized solutions (z-transformed continuous variables) since for analyses of questionnaire data these provide results that could be interpreted in terms of standard deviations.
Authors' response: The categorical variables in Multiple Linear regression (MLR) are used by considering the dummy variable approach where the largest of the categories are used as the reference category. There is no scope overall test of significance for all the categorical variables under MLR, rather each of the category gets a test of significance (t-test) which checks the difference are significant or not with respect to the reference category. We have provided the standard reporting style of MLR along with all the necessary outcomes that are needed to interpret the results from MLR.

Discussion
"convenience sample" (p. 15 line 8)- if you invited all consultants you don't have a convenience sample but the full population. Non-participation is no sampling strategy but part of the limitations!
Authors' response: We have revised the design and sample size section along with the discussion. We admit that it is not a convenience sample and non-participation is part of the limitation.

The authors should provide examples of how job autonomy could be increased for consultants. In theory, consultants already have a high level of job autonomy compared to other jobs. What might explain differences in job autonomy between consultants?
Authors' response: We have revised our conclusion to provide a more tangible suggestion, while acknowledging that that the NHS itself is under pressure.

Reviewer: 4
Reviewer Name: Eitan Naveh
Institution and Country: Please state any competing interests: None

Please leave your comments for the authors below The study aimed to examine the associations between job autonomy and burnout in relation to self-reported work-related stress and depression and anxiety symptoms; and then, to examine the current level of psychological morbidity among a sample of hospital consultants in the NHS.
Theoretical contribution. There are many works about burnout and about autonomy. The authors argue that this is the first study to put forward the relationship they suggest between autonomy and

burnout. The bottom line conclusion that autonomy is in general an influential factor is not surprising. However, some studies also suggest the existence of negative aspects of autonomy (Stern, Katz-Navon, & Naveh, 2008). The authors seems to ignore this line of work. In this respect, my most important reaction is that I would expect a more complex and comprehensive approach that includes the positive as well as negative aspects of autonomy. Clearly, the effects of autonomy are more positive under certain conditions, while others may emphasize its negative results. It is a little naïve to claim that autonomy has a significant main effect in all conditions (either positive or negative). It also does not reflect an updated knowledge about autonomy.
Authors' response: Thank you for your comments. Based on your comments we have revised our manuscript (and model), emphasising instead the role of burnout as a mediator. The main focus of the manuscript is no longer on autonomy.

I think that a major limitation of the paper derives from a lack of supporting theoretical base. I suggest providing the readers with a short description of the state-of-the-art knowledge in the field. This would allow the authors to see the difficulty the reader has in understanding their unique contribution to the advancement of said knowledge.
Authors' response: In our revised introduction we have focused on the role of burnout as a mediator between working conditions and the outcomes (depressive symptoms, anxiety symptoms, early retirement intention). We hope this makes the identified research gaps clearer.

A conclusion such as "improving consultants' job autonomy, and providing support and resources to prevent burnout" is not surprising and represents a narrow approach to the problem. Investing in additional resources is not a very helpful suggestion. Does the suggested investment come at the expense of other possible types of investment? What are the alternatives? Why should policymakers adapt this approach? I do not think this study provides them with a solid base that allows them to conclude that a higher investment in certain resources is conducive to solving the specific problem at hand.
Authors' response: We have revised our conclusion and included more specific recommendations, including the need for open dialogue and comprehensive interventions, as well as the possible role of job crafting.

Methods
1. I do not find a clear statement of the response rate.
Authors' response: Unfortunately our study design means it is not possible to determine the response rate. The results presented here are based on a sample of the consultants that responded to a call for participants. We approached the consultants though HR department of 7 health boards in Wales, 10 health board in Scotland and 12 trusts in England. We were not able to determine the total number of consultants approached or invited to our online survey. Therefore, the calculation based on some assumption on the total number of consultants approached and the assumed response rate, with margin of error and employing the principle of finite population correction. We hope that our revisions in the manuscript make this clearer.

2. It is not clear to what extent the authors used accepted measurement scales, for example in the case of autonomy. They used a three-item factor. A clear explanation of their selection of measurement tools is required.
Authors' response: We have elaborated our measure sections for all the included measures, including providing more details on the measure from which autonomy was used.

Analyses
From the description provided by the authors it is not clear whether they took into consideration the hierarchal nature of their data, i.e., individuals within departments within hospitals. The hierarchical structure of the data must be taken in to account and explained if the findings are to be based on a

solid foundation and accepted. This is not only an issue of analytical method, but also represents a correct theoretical approach to the manner in which individuals operate within their environment.
Authors' response: We did not perform any Multilevel modelling (i.e., hierarchical linear models or nested models). Because of the nature of the design of the study, the structure of the data was not collected in such levels. The participants of the study (i.e. the consultants) are fairly on the same level and but categorised into their respective specialities. We admit this is an interesting suggestion that we can employ in designing similar studies in the future.

VERSION 2 – REVIEW

REVIEWER	Charlotte N. L. Chambers Association of Salaried Medical Specialists, New Zealand
REVIEW RETURNED	19-Feb-2018

GENERAL COMMENTS	This appears a much better written and thought through paper than the first iteration. It is a shame that the response rate can't be calculated and this remains my lingering concern. The small numbers of people who participated in this study make it of limited value but it is still fulfilling an important role in understanding the experiences of consultants in the NHS.

REVIEWER	Daniel Schwarzkopf Jena University Hospital, Germany
REVIEW RETURNED	29-Mar-2018

GENERAL COMMENTS	Dear Editor, Thank you for giving me the possibility to review the revise of this paper. The authors invested a lot of effort to improve the paper, which I highly appreciate. The paper has improved a lot. I have some more suggestions and comments. Abstract: a) Methods- please do not refer to Hayes macro but name the concrete method to test the indirect effects (also Hayes macro provides several different methods)- don't use References in the Abstract. b) the description or results is not fully clear to me: Does job autonomy have a significant total effect on the outcomes or a significant direct effect? Perhaps it would be most clear if you describe if there are significant total effects for both predictors on the outcomes (not controlling for burnout dimensions) and then explain if these effects are "fully mediated" (no direct effect) or "partly mediated" by burnout dimensions. Readers from the medical profession oftentimes expect a description of all significant predictors for outcomes, perhaps you could include information which covariates were significant in predicting the three outcomes? I think it would be more convenient to describe the "incidence" of symptoms first and then refer to the test of relationships. Introduction The new introduction is well written and shows a gap in knowledge, since mediational analyses regarding burnout are not often done among physicians. Some literature, where the mediating role of burnout was discussed and tested could be added (e.g. Leiter MP, Maslach C. Nurse turnover: the mediating role of burnout. J Nurs Manag 2009;17(3):331-339., Schwarzkopf D, Rüddel H, Thomas-Rüddel DO, et al. Perceived nonbeneficial treatment of patients, burnout, and intention to leave the job among ICU nurses and junior and senior physicians. Crit Care Med 2017;45(3):e265-e273.)

	The provision of the research model is very helpful. I think it would also be of interest to test the mediating role of Depression and Anxiety on early retirement intentions- do depression and anxiety mediate the effects of emotional exhaustion and depersonalization on retirement intentions? If there is no theoretical reasoning to assume this, than this reasoning should be added to the paper. Methods I have concerns regarding the power calculation "This exceeded our power calculations assuming a total consultant population of 10,000 across these Health Boards and Trusts, a sample proportion of 50% and allowing 5% margin of error considering the finite population correction." I have some experience with power calculations but I don't understand this sentence…You need to justify the N of 500 and this has nothing to do with the possibly available N of 10.000. Why is 500 an appropriate N for this study? Why not 250 or 1000? If you did not have a proper reasoning for the N of 500 before the study was conducted- just delete the mentioning of the power calculation. 500 is an appropriate sample size for the calculations you did and you could refer to this in the strengths of the study. Name the socio demographics assessed in the survey. "internal reliability": correct would be "internal consistency". Do you refer to the literature? Or did you calculate by yourself? If yes give the method (Cronbach's alpha?) and the range of it. You have a repetition concerning this (measures and statistical analyses). I would name the method of calculation in the statistical analysis and give the numbers in the results. Regarding the measures on work characteristics "based on what the contemporary literature repeatedly highlights as particularly salient work characteristics to NHS consultants 34–36." I did not find how the references justify the selection of only the two characteristics? Perhaps you can comment. You could also use the classical Karasek demands/control model for justification. Statistical analyses: Please name the covariates/predictors used in the multiple regression analyses. I would think it is interesting also to know the predictors of the burnout scales, so I recommend additional analyses predicting them. Also a test of the mediational role of depression and anxiety on retirement intention would be of interest. Results: You repeat the sample size that was already given in the methods. I would only present this in the results. Table 1: please also present the "low risk" groups and mark statistical significant differences. Add explanation to the table so it can be read without having to look up information in the text (e.g. give the definition of high risk beneath the table). You provide Tenure in Table 2 but I don't see it in Table 1, please add this. You give the indication of "high level" I have some doubt that the manuscripts of the surveys speak of "high anxieity" and "high depression" (the numbers would not be reasonable compared to psychiatric patients!) please refere to the manuscripts of the surveys- I believe regarding burnout it is more appropriate to speak of a "burnout risk" and not of "high burnout". You report the reliabilities of the measures in the results- pleas again check for redundancy between methods and results. Direct effects: If you assume that emotional exhaustion mediates the effect of work pressure it is not a surprise that work pressure has no effect anymore. Please correct this. It is not appropriate to compare the effect sizes (standardized beta) in multiple logistic regression. I suggest adding the multiple regression results for predicting emotional exhaustion and depersonalization to Table 3. Also an

	analysis to test for the mediating role of anxiety and depressive symptoms could be added to this table, where anxiety and depression are added as additional variables. Please add overall tests for categorical variables (F-tests for specialty, age, country). Perhaps Table 3 could also be transferred to the online supplement and you concentrate on the total, direct and indirect effects of the studies primary concepts. Mediation analyses: To help the reader understand the concept of mediation I suggest to include total and direct effects in Table 4 (direct and indirect should sum up to the total effect (compare the online supplement of Schwarzkopf et al.). In Mediation analysis you only control for the significant predictors- you should control for all predictors. If you only control for significant predictors the indirect effect presented here and the direct effect presented in Table 3 are not estimated with the same model- which is not appropriate. Always use the same set of predictors for total, direct and indirect effects. Also it is wrong, to take only the significant "indicators" from a categorical variable ("consultants from wales is not the appropriate variable to choose- if the categorical variable country is significant, you must control for the full categorical variable- which indicator is significant strongly depends on the reference you choose- that is why you need the overall F-tests for the categorical variables see above). Add explanation to Table 4 to help the reader understand it without reading the manuscript. You should not discuss congruency with the multiple linear regression and the mediation analysis- if done correctly these have to be congruent! Give a thorough discussion of total, direct and indirect effects and present these as one analysis- The Hayes macro just takes the information of several linear regressions and calculates standard errors of the indirect effect- it is no analysis that is separate from the multiple regressions you put in. You could add a second figure resembling your theoretical framework but including the results of the analyses (standardized beta added to the arrows. Discussion Please check if the reported studies of NHS consultants that showed lower levels of burnout etc. used the same cut-offs. If yes- give possible explanations why the incidences were higher in your study. Include some theoretical literature on the mediational role of burnout to the third pargraph. Pragraph 4 ("in contrast to the existing literature)- to discuss the absence of a direct effect compared to existing literature would only make sense if this existing literature would have tested the mediation through burnout- I guess this is not the case- so omit this. You found an effect of work pressure on the outcomes- this corresponds to the existing literature ("total effect"), your finding is, that this effect is fully mediated by burnout. Refere to some literature that has shown mediational effects of burnout (e.g. Leiter et al. Schwarzkopf et al. or what else you find) Paragraph 5: You give an explanation for finding no indirect effect through depersonalization on retirement. You state that physicians might depersonalize and thereby be able to stay longer in the job. This would imply a negative relationship between depersonalization and retirement intention (Table 2 shows a significant positive effect). The effects of predictors on depersonalization and the effect of depersonalization on retirement are smaller than those for emotional exhaustion- and that is the explanation- you just lack power to find a significant indirect effect. Retirement might mostly be driven by the

	feeling of being exhausted. Practical implications: You could give some hints regarding which interventions could reduce work pressure and increase job autonomy. Individual interventions might still be effective since they help cope with work stressors and prevent burnout (e.g. mindfulness based interventions West CP, Dyrbye LN, Rabatin JT, et al. Intervention to Promote Physician Well-being, Job Satisfaction, and Professionalism A Randomized Clinical Trial. JAMA Intern Med 2014;174(4):527-533.) Strengths and limitations of the study should be discusses in an extra paragraph. There are further limitations that are not discussed by the authors: - It is unknown how many consultants were the population, the response rate is low given 10.000 possible consultants - It is unknown if the participating consultants are different from nonparticipants (e.g. participants might show higher strain) - External validity of the findings for the population of NHS consultants might be questionable - Burnout as well as the measures of anxiety and depression don't give clinical diagnosis! Scoring high on these scales does not imply clinical relevant conditions. So it should be discussed what meaning "high" values have. In a dutch study only very high levels (above the 95% percentile of the norm population) of burnout correspond to levels shown by patients seeking treatment for neurasthenia (Schaufeli & Enzmann, 1998, The burnout companion, p. 58, Schaufeli WB, Bakker AB, Hoogduin K, et al. On the clinical validity of the Maslach Burnout Inventory and the Burnout Measure. Psychology & Health 2001;16(5):565-582.) - The problem of causality and need for longitudinal studies has already been discussed by the authors. Some suggestions on a possible longitudinal study design could be included.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 1
Reviewer Name: Charlotte N. L. Chambers
Institution and Country: Association of Salaried Medical Specialists, New Zealand Please state any competing interests: None Declared
Please leave your comments for the authors below

This appears a much better written and thought through paper than the first iteration. It is a shame that the response rate can't be calculated and this remains my lingering concern. The small numbers of people who participated in this study make it of limited value but it is still fulfilling an important role in understanding the experiences of consultants in the NHS.
Authors' response: We thank the Reviewer for their comment and the time taken review our manuscript. We acknowledge the limitation highlighted by the Reviewer and hope that by highlighting this more clearly in the limitations we are able to make this point more salient to readers.

Reviewer: 3
Reviewer Name: Daniel Schwarzkopf
Institution and Country: Jena University Hospital, Germany Please state any competing interests: no competing interests

Dear Editor,
Thank you for giving me the possibility to review the revise of this paper.
The authors invested a lot of effort to improve the paper, which I highly appreciate. The paper has

improved a lot. I have some more suggestions and comments.
Authors' response: We thank the Reviewer for their thorough review of our manuscript, and for the considerate and constructive comments provided. We have reflected and made changes were possible. These are detailed in response to the specific comments below, and we hope they are acceptable to the Reviewer.

Abstract: a) Methods- please do not refer to Hayes macro but name the concrete method to test the indirect effects (also Hayes macro provides several different methods)- don't use References in the Abstract. b) the description or results is not fully clear to me: Does job autonomy have a significant total effect on the outcomes or a significant direct effect? Perhaps it would be most clear if you describe if there are significant total effects for both predictors on the outcomes (not controlling for burnout dimensions) and then explain if these effects are "fully mediated" (no direct effect) or "partly mediated" by burnout dimensions. Readers from the medical profession oftentimes expect a description of all significant predictors for outcomes, perhaps you could include information which covariates were significant in predicting the three outcomes? I think it would be more convenient to describe the "incidence" of symptoms first and then refer to the test of relationships.
Authors' response: Based on the Reviewer's comments and considering the restriction we have by the abstract word count, we have: removed Hayes Macro and the reference from the abstract; we have made minor edits to the description of the direct and indirect effects; reference made to socio-demographic details; and we have also moved the incidence statistics to the start of the results section.

Introduction
The new introduction is well written and shows a gap in knowledge, since mediational analyses regarding burnout are not often done among physicians. Some literature, where the mediating role of burnout was discussed and tested could be added (e.g. Leiter MP, Maslach C. Nurse turnover: the mediating role of burnout. J Nurs Manag 2009;17(3):331-339., Schwarzkopf D, Rüddel H, Thomas-Rüddel DO, et al. Perceived nonbeneficial treatment of patients, burnout, and intention to leave the job among ICU nurses and junior and senior physicians. Crit Care Med 2017;45(3):e265-e273.) The provision of the research model is very helpful. I think it would also be of interest to test the mediating role of Depression and Anxiety on early retirement intentions- do depression and anxiety mediate the effects of emotional exhaustion and depersonalization on retirement intentions? If there is no theoretical reasoning to assume this, than this reasoning should be added to the paper.
Authors' response: We have added in the two references provided. We also used the latter study to indicate how different healthcare groups report different levels of burnout, burnout and perception of work environment, which lends to support to focus on this group of consultants where not as much known in comparison to other occupational groups.
We also accept that the Reviewer has a valid point about depressive and anxiety symptoms functioning as a mediator. This could be either: (i) in parallel with burnout, or (ii) where job autonomy/ perceived work pressure influence burnout, which in turn influences depressive/ anxiety symptoms, which in turn influences retirement intention. Although the analysis is possible, this needs to be accompanied by a suitable review of the literature where depression/anxiety could function as a mediator. More crucially, this would also have to consider and compare depressive/anxiety symptoms as a mediator in relation to burnout. To do this justice, we feel would substantially increase the introduction and the complexity of the paper. More importantly, we feel it distracts from the two stated aims of this paper, which in turn may create an unclear narrative for this study. We hope that the Reviewer is able to accept our rationale for this decision.

Methods
I have concerns regarding the power calculation "This exceeded our power calculations assuming a total consultant population of 10,000 across these Health Boards and Trusts, a sample proportion of 50% and allowing 5% margin of error considering the finite population correction." I have some

experience with power calculations but I don't understand this sentence…You need to justify the N of 500 and this has nothing to do with the possibly available N of 10.000. Why is 500 an appropriate N for this study? Why not 250 or 1000? If you did not have a proper reasoning for the N of 500 before the study was conducted- just delete the mentioning of the power calculation. 500 is an appropriate sample size for the calculations you did and you could refer to this in the strengths of the study.
Authors' response: We agree with the concern in relation of mentioning 'power calculation' in the sample size calculation and acknowledge the confusion it presents. The setting of the study didn't allow us to perform a proper and standard power analysis as we didn't employ any intervention targeting a primary outcome measure likewise any traditional randomised control trial. We thank the Reviewer for correcting us on this.
The method of sampling was a survey sample method where we assumed the total number of consultants as 10,000 and population proportion as 50% (the assumed proportion of the consultants having 'job autonomy and work-related') and considering 5% margin of error. This yielded about a required sample size of 370. Since we have much higher complete response of (593), we considered that 500 would show a round sample size number for us to target. In addition, this slightly larger number better represents the overall population (as a larger sample size been considered than the figure came out of the calculation). A similar approach was used and published in BMJ Open by Lorini et al. (2017). Lorini C, Santomauro F, Grazzini M, et al Health literacy in Italy: a cross-sectional study protocol to assess the health literacy level in a population-based sample, and to validate health literacy measures in the Italian language BMJ Open 2017;7:e017812. doi: 10.1136/bmjopen-2017-017812

Name the socio demographics assessed in the survey.
Authors' response: We have added this in.

"internal reliability": correct would be "internal consistency". Do you refer to the literature? Or did you calculate by yourself? If yes give the method (Cronbach's alpha?) and the range of it. You have a repetition concerning this (measures and statistical analyses). I would name the method of calculation in the statistical analysis and give the numbers in the results.
Authors' response: We have changed this to internal consistency and specified that this refers to Cronbach's alpha. Reference to internal consistency has been removed from the measures section, with the Cronbach's alpha specified in the statistical analysis section, and a sentence briefly outlining the range of scores in the results.

Regarding the measures on work characteristics "based on what the contemporary literature repeatedly highlights as particularly salient work characteristics to NHS consultants 34–36." I did not find how the references justify the selection of only the two characteristics? Perhaps you can comment. You could also use the classical Karasek demands/control model for justification.
Authors' response: We have added in a sentence referencing Karasek's model. For the references included, the report by Martin et al. (2015) focuses on the erosion of autonomy among NHS consultants in Scotland; Walker et al. (2016) found that professional/personal demands was the biggest predictor of surgeons leaving the NHS; while Smith et al. (2017) asked NHS doctors what would keep them in medicine for longer, with the top two responses pertaining to demands ('reduced impact of work-related bureaucracy' and 'workload reduction/shorter hours'.

Statistical analyses: Please name the covariates/predictors used in the multiple regression analyses. I would think it is interesting also to know the predictors of the burnout scales, so I recommend additional analyses predicting them. Also a test of the mediational role of depression and anxiety on retirement intention would be of interest.
Authors' response: We have clarified further in the statistical analyses section the predictors and covariates used in the multiple regression analyses. We have made edits to reflect the examination of job autonomy and work-related pressure as predictors of the two burnout dimensions. As discussed

earlier, we did not examine the mediational role of depressive and anxiety symptoms.

Results: You repeat the sample size that was already given in the methods. I would only present this in the results.
Authors' response: We have removed the sample size from the method second.

Table 1: please also present the "low risk" groups and mark statistical significant differences. Add explanation to the table so it can be read without having to look up information in the text (e.g. give the definition of high risk beneath the table). You provide Tenure in Table 2 but I don't see it in Table 1, please add this.
Authors' response: We have added in the 'low' groups for emotional exhaustion and depersonalisation (See Table 1). No corresponding cut-off points are available for depression and anxiety so it was not possible to estimate this for these measures. We do not seek to compare the comparison across the different categories in Table 1. This is beyond the scope of our study aims. We have already included a sentence in the paragraph under Table 1 that "these comparisons are not based on inferential statistics and are only for descriptive purposes".
We have added in the 'Note' section under Table 1 the cut-off scores for emotional exhaustion, depersonalisation, anxiety and depressive symptoms. Tenure has also been included into Table 1.

You give the indication of "high level" I have some doubt that the manuscripts of the surveys speak of "high anxieity" and "high depression" (the numbers would not be reasonable compared to psychiatric patients!) please refere to the manuscripts of the surveys- I believe regarding burnout it is more appropriate to speak of a "burnout risk" and not of "high burnout".
Authors' response: In terms of burnout, Maslach distinguishes between 'high', 'medium' and 'low' burnout which is what the cut of points used in Table 1 refers to. In terms of anxiety and depression, our measures used represents a high level of corresponding symptoms. This does not refer to clinical depression or anxiety, and is explained in our measures section (e.g., high scores representing more frequent experience of anxiety and depressive symptoms). We have been careful to refer to depression and anxiety symptoms and have reviewed the manuscript again to make sure that this is consistent throughout.

You report the reliabilities of the measures in the results- pleas again check for redundancy between methods and results.
Authors' response: We now only report the internal consistencies in the results section.

Direct effects: If you assume that emotional exhaustion mediates the effect of work pressure it is not a surprise that work pressure has no effect anymore. Please correct this.
Authors' response: We have corrected this.

It is not appropriate to compare the effect sizes (standardized beta) in multiple logistic regression.
Authors' response: We have corrected this.

I suggest adding the multiple regression results for predicting emotional exhaustion and depersonalization to Table 3. Also an analysis to test for the mediating role of anxiety and depressive symptoms could be added to this table, where anxiety and depression are added as additional variables. Please add overall tests for categorical variables (F-tests for specialty, age, country).
Authors' response: We have added in the regressions for predicting the two burnout dimensions into Table 3. For the reasons described to the suggestion in the introduction, we have elected to not examine the mediating role of depressive/anxiety symptoms. We did not carry out hierarchical regressions, therefore we do not have separate F-tests for only for the categorical variables. Instead, our F-tests are for the total model, we have added these into Table 3.

Perhaps Table 3 could also be transferred to the online supplement and you concentrate on the total, direct and indirect effects of the studies primary concepts.
Authors' response: We have been able to make some of edits suggested by the Reviewer, allowing Table 3 to remain on a single page. We differ this decision to the Editorial Staff if they feel this appropriate.

Mediation analyses:
To help the reader understand the concept of mediation I suggest to include total and direct effects in Table 4 (direct and indirect should sum up to the total effect (compare the online supplement of Schwarzkopf et al.).
Authors' response: We have added in the total and direct effects into Tables 4 and 5, using the suggestion of Schwarzkopf et al. as an initial template. This now has the a, b, a*b, c and c' paths.

In Mediation analysis you only control for the significant predictors- you should control for all predictors. If you only control for significant predictors the indirect effect presented here and the direct effect presented in Table 3 are not estimated with the same model- which is not appropriate. Always use the same set of predictors for total, direct and indirect effects. Also it is wrong, to take only the significant "indicators" from a categorical variable ("consultants from wales is not the appropriate variable to choose- if the categorical variable country is significant, you must control for the full categorical variable- which indicator is significant strongly depends on the reference you choose- that is why you need the overall F-tests for the categorical variables see above).
Authors' response: Our initial plan was to use the principle of parsimony and to try and preserve degrees of freedom. Nevertheless, we accept your point and have re-run this analysis by including all the predictors used in Table 3 into the subsequent mediations.

Add explanation to Table 4 to help the reader understand it without reading the manuscript.
Authors' response: We hope that our revised Tables 4 and 5 are clearer than before.

You should not discuss congruency with the multiple linear regression and the mediation analysis- if done correctly these have to be congruent! Give a thorough discussion of total, direct and indirect effects and present these as one analysis- The Hayes macro just takes the information of several linear regressions and calculates standard errors of the indirect effect- it is no analysis that is separate from the multiple regressions you put in.
Authors' response: We acknowledge this point and as part of our re-write of these sections have removed reference to this. This now focuses solely on the indirect effects.

You could add a second figure resembling your theoretical framework but including the results of the analyses (standardized beta added to the arrows.
Authors' response: We considered and discussed this, but feel that with five tables and a figure we should be more selective in the number an additional files included with this manuscript. This is particularly as we feel this may duplicate rather than complement our existing tables and in-text discussion.

Discussion
Please check if the reported studies of NHS consultants that showed lower levels of burnout etc. used the same cut-offs. If yes- give possible explanations why the incidences were higher in your study.
Authors' response: These studies either used different measures to assess stress and depressive symptoms, or did not publish the cut-off scores burnout. We acknowledge that this needs to be made clear and have added in two sentences to highlight this.

Include some theoretical literature on the mediational role of burnout to the third pargraph.
Authors' response: We have added in an explanation drawing on conservation of resources theory.

As a result of this longer paragraph we have also separated this paragraph into two.

Pragraph 4 (“in contrast to the existing literature)- to discuss the absence of a direct effect compared to existing literature would only make sense if this existing literature would have tested the mediation through burnout- I guess this is not the case- so omit this. You found an effect of work pressure on the outcomes- this corresponds to the existing literature (“total effect”), your finding is, that this effect is fully mediated by burnout. Refere to some literature that has shown mediational effects of burnout (e.g. Leiter et al. Schwarzkopf et al. or what else you find)
Authors’ response: We have removed this sentence and made the edits as suggested by the Reviewer.

Paragraph 5: You give an explanation for finding no indirect effect through depersonalization on retirement. You state that physicians might depersonalize and thereby be able to stay longer in the job. This would imply a negative relationship between depersonalization and retirement intention (Table 2 shows a significant positive effect). The effects of predictors on depersonalization and the effect of depersonalization on retirement are smaller than those for emotional exhaustion- and that is the explanation- you just lack power to find a significant indirect effect. Retirement might mostly be driven by the feeling of being exhausted.
Authors’ response: We have reflected on the Reviewer’s comment and feel that depersonalisation can exist as a potential coping mechanism (Storm & Rothmann, 2003). We acknowledge the correlation coefficient in Table 2, but feel that the lack of an effect size in the regressions (Table 3) can potentially be explained by our initial point. We have, however, edited for clarity purpose. We have also included an additional line to acknowledge the feelings of exhaustion as a driver to retirement. [Storm, K., & Rothmann, S. (2003). The relationship between burnout, personality traits and coping strategies in a corporate pharmaceutical group. Sa Journal of industrial psychology, 29(4), 35-42.]

Practical implications: You could give some hints regarding which interventions could reduce work pressure and increase job autonomy. Individual interventions might still be effective since they help cope with work stressors and prevent burnout (e.g. mindfulness based interventions West CP, Dyrbye LN, Rabatin JT, et al. Intervention to Promote Physician Well-being, Job Satisfaction, and Professionalism A Randomized Clinical Trial. JAMA Intern Med 2014;174(4):527-533.)
Authors’ response: We have included this reference and another (Regehr et al., 2014) to lend some support for individual interventions.

Strengths and limitations of the study should be discusses in an extra paragraph. There are further limitations that are not discussed by the authors:
- It is unknown how many consultants were the population, the response rate is low given 10.000 possible consultants
- It is unknown if the participating consultants are different from nonparticipants (e.g. participants might show higher strain)
- External validity of the findings for the population of NHS consultants might be questionable
- Burnout as well as the measures of anxiety and depression don’t give clinical diagnosis! Scoring high on these scales does not imply clinical relevant conditions. So it should be discussed what meaning “high” values have. In a dutch study only very high levels (above the 95% percentile of the norm population) of burnout correspond to levels shown by patients seeking treatment for neurasthenia (Schaufeli & Enzmann, 1998, The burnout companion, p. 58, Schaufeli WB, Bakker AB, Hoogduin K, et al. On the clinical validity of the Maslach Burnout Inventory and the Burnout Measure. Psychology & Health 2001;16(5):565-582.)
- The problem of causality and need for longitudinal studies has already been discussed by the authors. Some suggestions on a possible longitudinal study design could be included.
Authors’ response: We have included a new paragraph that focuses on limitations. These address the

points made about longitudinal designs; anxiety and depression; response rate and non-participation; and common method variance.