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)	A Study Protocol for a Prospective, Longitudinal Cohort Study Investigating the Medical and Psychosocial Outcomes of United Kingdom Combat Casualties from the Afghanistan War: the ADVANCE Study
AUTHORS	Bennett, Alexander; Dyball, Daniel; Boos, Christopher J.; Fear, Nicola; Schofield, Susie; Bull, Anthony; Cullinan, Paul

VERSION 1 – REVIEW

REVIEWER	Dr Sougat Ray Asvini Hospital, India
REVIEW RETURNED	16-Mar-2020

GENERAL COMMENTS	Cohort Study is carried out to determine and analyse 'harm', i.e, to analyse the risk factors of the desired outcome, and in this case incidence of adverse medical, psychosocial and vocational long-term outcomes compared to an equivalent but non-injured service personnel. The authors have also hypothesised that adverse outcomes will be found in the study group compared to the control, and it is known. The study should have analysed other risk factors (other than combat trauma) in the study group which might contribute to the desired outcome. Think this is a basic flaw in hypothesis generation. While analysing the cohort study, RR along with AR and ARR are required to be analysed.

REVIEWER	Joost Op 't Eynde Duke University Durham, North Carolina United States
REVIEW RETURNED	19-Mar-2020

GENERAL COMMENTS	I think this is a well designed study I look forward to seeing the results of. Comments: Page 12 section on HRV switches tenses between present and future tense. Make sure all text is written in the same tense, for this and other sections. Check line spacing between titles and paragraphs to make sure it is consistent.

REVIEWER	Helena Chmura Kraemer Department of Psychiatry and Behavioral Sciences Stanford University (Emerita) US

REVIEW RETURNED	02-Jun-2020

GENERAL COMMENTS

There are, in my opinion, many problems with the proposal, but the study has apparently been underway since 2016, and the design cannot now be changed. However, it may be that anticipating questions that might be raised, might help with the analysis and presentation of results?

#1. There is no way that inferences of causality can be based on this type of observational study. You are exploring possible outcomes, perhaps to establish that combat casualty in the Afghanistan War is a risk factor for certain outcomes, but there are always factors that you do not know preceding the event and following the event that might explain away any causal connection between a risk factor and outcome. Such a study may still be important, but not because of causality inferences that might seriously mislead.

#2. There is here a major problem with time. The Afghanistan War occurred in 2003-2014, and to be included in this study, someone who was relatively healthy before their service, and who served during that war, had to have survived and be available in 2016 (some 2-13 years later), which is when they are to be recruited into this study.

It may well be that many who sustained war injuries may have died between their time of service and 2016, or become unavailable to be recruited in 2016 (become homeless or emigrated etc.), perhaps having suffered more serious outcomes of injury than those you are seeking within your study. Thus your cases are those that were combat casualties and were yet available in 2016 and willing to participate. You may be dealing with those with the less serious long-term outcomes.

Moreover, the time span between the precipitating event (the injury) and entry to the study may vary from 2 to 13 years, and the last planned follow up time (20 years after entry) may occur from 22 to 33 years after the precipitating event. Many of your outcomes (e.g., heart problems, arthritis) are age-related. The issue of time between the precipitating event and the timing of outcomes measures, as well as the age of the person at the time of the injury and at each follow-up point deserves a great deal of consideration.

#3. Sampling: Figure 1 appears to say that you will sample all 257 casualties with amputations, all 304 with VSI/SI, and a subsample of 839 of the 1038 others, the three groups to total 1400. Thus this is a stratified sample, undersampling the “others”.

Since the long-term outcomes (both before 2016 and after) may be different in the first two groups, and since some of your outcomes are specific to amputations, all your analyses, and your selection of appropriate controls, must be in the 3 strata you here define. I don’t see consideration of stratification anywhere in your analysis plan.

The selection of controls is also not clear. You are “frequency matching” the selection of controls to the cases using service, age, rank, regiment, roulement and role-in-theater. I understand what age, rank, and regiment are, but what are “roulement” and “role-in-theater”? These may be specific to the British military and should be explained.

If you simply randomly selected controls from the uninjured Afghanistan veterans, I would suspect the cases and controls would not be matched on age, and that would impact outcomes, but why are you using all of the remaining four? Would it not suffice, for example, to match on battle-involved versus not or some such?

I assume that corresponding to the precipitating event, you are using the LAST time of service in Afghanistan as the event time for uninjured controls? All who were injured are only eligible to be cases, and presumably the injury, if serious, was incurred in the last deployment?
You say you are “frequency matching”, which sounds like you are selecting cases only after the controls have been selected? Are you “frequency matching” separately to the amputees, the VSI/SI and the other injured? Since these may have different distributions of the matching factors, that should be done.

#4. Analysis: How is the “frequency matching” handled in the analysis? If you had randomly sampled controls, the usual statistical analyses, t-test, ChiSquare test etc., would suffice. If, at the other extreme, you had created matched pairs, one control matched to one case, you would have to use matched pairs t-tests instead of t-test, McNemar test instead of ChiSquare, etc. Not only is the analysis different, but the research question changes. You appear to be considering use of the Proportional Hazards Model (Hazard Ratio), but to what event, with timing relative to what initial time point (entry to the study? Age at injury/last deployment?, Chronological Age?). I am not aware of any version of this model for use with paired or matched data.
While the sample size is based on time to events, most of the measures you propose are ordinal. On page17, you propose repeated measures ANOVA and Friedman tests, neither of which is appropriate for use with time-series data. It is known that use of these approaches result in exaggerated p-values (false positives). You might consider Random Regression Linear Models with autoregressive covariance structure between multiple time points within a subject.
What you are doing is somewhere in between, not as close as pairwise matching, not as independent as random selection. It’s not exactly clear that using the usual tests is correct, and clearly the matched tests are not appropriate. This issue should be discussed and dealt with.
On page 7, you also say “whilst adjusting for confounders”. Do you really mean “confounders”, or are you just throwing in covariates? A “confounder” is defined as a variable that is associated both with the independent variable of interest, here case/control, and with the outcome of interest. That automatically raises collinearity issues in regression analyses, and then there is the problem of possible interaction between the independent variable and the “confounder”. If you are thinking of multiple confounders, there is further concern about the correlations among the multiple confounders and their interactions as well. This may seem a casual remark, but I would strongly urge you to think carefully about this.

#5. The lack of specific ‘a priori’ hypotheses, and the multiplicity of outcome measures, suggests that this is an exploratory study, designed to generate hypotheses to be tested in future independent studies. But then the analytic plans seem to suggest that you are testing unspecified hypotheses without stated rationale and justification. Please consider before you start exactly what the goals of this study are. Once you explore, all tests are ‘post hoc’ and p-values and conclusions likely invalid.

#6. Some of these measures appear to be only appropriate for amputees, thus only for one stratum of cases, and for none of controls. These are useless for comparisons between cases and

	controls. All the results, except for two measures related to brain injury are self-reported. Since all know whether they are cases and controls, there is a problem here of possible bias, i.e., that the answer reflects not the physical or psychological state, as much as a different mind set in answering questionnaires. It would be very useful to do a pilot study or two to generate information comparing these self-reports with objective measures to validate. Could you not have access to Health Service records to validate some of these? Table 2: The inclusion/exclusion criteria should apply both to cases and controls. Controls were not, e.g., “Injured between 2003 and the end of 2014”. You might put the general inclusion/exclusion criteria that apply to both groups first, and then add the criteria that differentiate cases from controls at the bottom of this table, starting with what you label “Comparison group only”. When you say “Past…” you mean before or immedidately after the last deployment? If “Past” includes, e.g., diabetes diagnosed 5 years after service in Afghanistan, you will be excluding the long-term outcomes this study was meant to examine. However, the “infection” exclusion seems to apply to the initial assessment in 2016 or later, and probably to all subsequent assessments? Please clarify. Table 3: Define the three strata here with sample sizes 257, 304 and 839. Alternatively since you take all amputations and VSI/SI, you could have two strata, with sample sizes 561 and 839. I would think three strata would make more sense, because some outcome measures are specific to amputations. Table 4: “Duplicate records….” Please specify that it was the last Afghanistan deployment that is used to be comparable to the cases. I do not understand how you are creating a “frequency matched” sample. . I wonder whether other readers would. You do need to do this separately for the 2 or 3 strata defined by your sampling The last table (listing all outcomes measured) is what suggests that this is an exploratory study. Some effort should be made to label those measures only appropriate for amputees, and those measures only appropriate for non-amputees in the text of the table. This table should be shortened simply to list Baseline Only measures, and Longitudinal Measures (at baseline 3,5,10,15,20 years). There is no need to have separate columns for the multiple time points. The table should be much shorter. Before analysis of outcomes, there should be some effort to investigate the inter-correlations of these measures, and to reduce the long list to relatively independent measures. As it is, if you test every measure separately, and perhaps each measure at each time point (which would undermine the longitudinal aspect of the study),such multiple testing guarantees many false positive results. At the same time, with the stratified sample, with the “frequency matching”, and the likely missing data, there will also be a lot of false negative results. At that point, all your conclusions will be questionable.

REVIEWER	Deepak Nag Ayyala Medical College of Georgia, Augusta University Georgia, USA
REVIEW RETURNED	18-Jun-2020

GENERAL COMMENTS	The study investigates the medical and psychosocial outcomes of adults who sustained physical combat trauma while on deployment in Afghanistan. The sample size analysis indicates that the study is sufficiently powered for the hypothesis of interest. The authors indicate in their "Statistical methods" section that missing data will be imputed. Statistical significance will be determined using a significance level of 5%. Minor comments: 1. Would help to include what method of imputation (missing at random or missing completely at random) will be used during the analysis. 2. Address how the p-values will be adjusted for multiple comparisons (Bonferroni's correction or false discovery rate). 3. Fix the reference error on Page 11, line 5.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1: Dr Sougat Ray

Point 1: “Cohort Study is carried out to determine and analyse 'harm', i.e, to analyse the risk factors of the desired outcome, and in this case incidence of adverse medical, psychosocial and vocational long-term outcomes compared to an equivalent but non-injured service personnel. The authors have also hypothesised that adverse outcomes will be found in the study group compared to the control, and it is known”.

The reviewer states that it is “known” that “adverse outcomes will be found in the study group”. There is no debate that there are several retrospective or cross-sectional studies published on combat casualties from previous conflicts which indicate that combat casualties have worse outcomes. However, as discussed in the manuscript, the findings are from studies that are less methodologically robust than a cohort study, in populations with less severe or different injuries with different or less comprehensive outcomes.

Point 2: “The study should have analysed other risk factors (other than combat trauma) in the study group which might contribute to the desired outcome. Think this is a basic flaw in hypothesis generation.”

The ADVANCE study’s main hypothesis is: “…combat trauma casualties will have an increased incidence of adverse medical, psychosocial and vocational long-term outcomes compared to equivalent but non-injured service personnel”. We recognised that there may be risk factors other than combat trauma which will contribute to the long term outcomes and have therefore collected information on a wide range of factors; a full list of the data we collect can be found in supplementary materials 1.

Point 3: “While analysing the cohort study, RR along with AR and ARR are required to be analysed”

We agree that RR is best even if the outcome is rare and therefore, have adjusted the methods to reflect this. See below.

Page 14: “Generalized Linear Models with a binomial distribution will be used to assess the relative risk of……..”

Reviewer 2: Dr Joost Op 't Eynde

Point 1: "Page 12 section on HRV switches tenses between present and future tense." and "Check line spacing between titles and paragraphs to make sure it is consistent".

We have reread the manuscript and made appropriate changes to tense and formatting.

Reviewer 3: Helena Chmura Kraemer

Point 1: "There is no way that inferences of causality can be based on this type of observational study. "

The reviewer quite rightly points out that the ADVANCE study has, necessarily, an observational design. It is a prospective cohort study investigating the adverse medical, psychosocial and vocational outcomes of those with a battlefield injury. As such it is comparing an "exposed" group of battlefield injured UK ex/-serving military personnel with a similar but unexposed (uninjured) group. We agree that this design cannot infer causality and accordingly our hypotheses do not reference causality. We do however acknowledge that causality was incorrectly mentioned in the article summary 'strengths and limitations' and this has now been removed (see below and in main manuscript). Causality was also mentioned in the introduction as a criticism of other previous papers, inferring perhaps that ADVANCE may prove causality; this is also incorrect and has been removed (see below and in main manuscript).

Page 3: "ADVANCE will provide a wide range of longitudinal data across sociodemographic, physical health and mental health outcomes, providing evidence for incidence and risk of disease and other non-disease related outcomes."

Page 4: "Many studies investigating veterans' long-term outcomes are either not specifically related to combat trauma [22-37] or are of cross sectional or retrospective design making it difficult to draw robust conclusions from them. [10-13, 15-25, 27, 29, 30, 32, 34, 38, 39]."
Point 2: "There is here a major problem with time. The Afghanistan War occurred in 2003-2014, and to be included in this study, someone who was relatively healthy before their service, and who served during that war, had to have survived and be available in 2016 (some 2-13 years later), which is when they are to be recruited into this study.
It may well be that many who sustained war injuries may have died between their time of service and 2016, or become unavailable to be recruited in 2016 (become homeless or emigrated etc.), perhaps having suffered more serious outcomes of injury than those you are seeking within your study….You may be dealing with those with the less serious long-term outcomes."

The authors agree that in an ideal scenario all participants would have been recruited into ADVANCE at the time of, or very soon after, injury. This unfortunately was not possible. We can give reassurance, however, that if injured UK military personnel survived initial trauma management in operational theatre, very few subsequently died once back in hospital care in the UK. All moderately to severely injured patients then went on to several months of multidisciplinary in-patient rehabilitation before being medically discharged with a social and care package in place if required. The short to medium term outcomes of UK military personnel are good and have been published (see below1-5) and the incidence of death in the first few years after injury, although acknowledged as a possibility, is low. Any study of long-term outcomes will, of course, be limited to those who survived in the short term.

The ADVANCE study has made every effort to recruit eligible participants in all circumstances and locations from all over the UK and Europe. Multiple sources have been used to contact eligible participants including charities, patient and veteran groups and the electoral roll. All travel has been funded. Participants who are unable to travel on public transport have been offered disabled friendly taxis. Participants in EEA countries have been flown in, leaving only participants who have moved further afield unable to take part, though even in those cases if they make regular trips back to the UK we have invited them to take part. It is also important to highlight that the recruitment to the ADVANCE study is from a Ministry of Defence generated list of UK service personnel with known severe injuries from the Afghanistan war and it is not open to any volunteer who meet the inclusion criteria. Thus far, the response rate has been good at circa 60% and the mean time from injury/appropriate deployment is 7 yrs. However even with all these avenues of recruitment explored, we acknowledge there may be a response bias when we complete the study. We will be weighting analyses by any significant sampling or non-responder characteristic differences to make the study as reflective as possible of the whole population.

1. Outcomes for UK service personnel indicate high quality trauma care and rehabilitation. Etherington J, Bennett AN, Phillip R, Mistlin A.BMJ. 2016 Sep 6;354:i4741. doi: 10.1136/bmj.i4741.
2. Bahadur, S., McGilloway, E., & Etherington, J. (2016). Injury severity at presentation is not associated with long-term vocational outcome in British Military brain injury. Journal of the Royal Army Medical Corps, 162(2), 120–124. http://doi.org/10.1136/jramc-2014-000393
3. Dharm-Datta, S., Gough, M. R. C., Porter, P. J., Duncan-Anderson, J., Olivier, E., McGilloway, E., & Etherington, J. (2015). Successful outcomes following neurorehabilitation in military traumatic brain injury patients in the United Kingdom. The Journal of Trauma and Acute Care Surgery, 79(4 Suppl 2), S197–203. http://doi.org/10.1097/TA.0000000000000721
4. Jarvis H, Baker R, Bennett A, Twiste M, Phillip R. Kinematics, kinetics and gait profile score in highly functional amputees. Prosthet Orthot Int2015. p. 258
5. Ladlow, P., Phillip, R., Etherington, J., Coppack, R., Bilzon, J., McGuigan, M. P., & Bennett, A. N. (2015). Functional and Mental Health Status of United Kingdom Military Amputees Postrehabilitation. Arch Phys Med Rehabil, 96(11), 2048–2054. http://doi.org/10.1016/j.apmr.2015.07.016

Point 3: “Moreover, the time span between the precipitating event (the injury) and entry to the study may vary from 2 to 13 years, and the last planned follow up time (20 years after entry) may occur from 22 to 33 years after the precipitating event. Many of your outcomes (e.g., heart problems, arthritis) are age-related. The issue of time between the precipitating event and the timing of outcomes measures, as well as the age of the person at the time of the injury and at each follow-up point deserves a great deal of consideration.”

Since the study is investigating the long-term effects of battlefield injury, it is unavoidable that many years will have passed since the deployment to Afghanistan and injury (for the injured group). Years since injury and age will be added as covariates to analysis where appropriate. However, it is also of note that the comparison group will have a similar length of follow-up since their deployment, making comparisons between the two groups valid.

Point 4: “Figure 1 appears to say that you will sample all 257 casualties with amputations, all 304 with VSI/SI, and a subsample of 839 of the 1038 others, the three groups to total 1400. Thus this is a stratified sample, undersampling the “others”. Since the long-term outcomes (both before 2016 and after) may be different in the first two groups, and since some of your outcomes are specific to amputations, all your analyses, and your selection of appropriate controls, must be in the 3 strata you here define. I don’t see consideration of stratification anywhere in your analysis plan.”

The ADVANCE study is investigating the long-term effect of battlefield injury on numerous outcomes. Within the whole battlefield injury group, we wanted to ensure good representation of the most severely injured/those with amputations. The figure shows that first, amputees were selected, then Very Seriously Injured (VSI) and Seriously Injured (SI) men. By doing this we ensure that the most severely injured are represented in our pool of potential participants. We will apply weights to take into account the disproportionate sampling of exposed and unexposed. This has been made clearer in our methods section.

Point 5: “The selection of controls is also not clear. You are “frequency matching” the selection of controls to the cases using service, age, rank, regiment, roulement and role-in-theater. I understand what age, rank, and regiment are, but what are "roulement" and "role-in-theater"? These may be specific to the British military and should be explained."

Thank you for this comment. 'Roulemont' has been replaced with "specific deployment period". Role in theatre has also been described in the text and refers to their job role during their deployment.

Point 6: "If you simply randomly selected controls from the uninjured Afghanistan veterans, I would suspect the cases and controls would not be matched on age, and that would impact outcomes, but why are you using all of the remaining four? Would it not suffice, for example, to match on battle-involved versus not or some such? "
"You say you are "frequency matching", which sounds like you are selecting cases only after the controls have been selected? Are you "frequency matching" separately to the amputees, the VSI/SI and the other injured? Since these may have different distributions of the matching factors, that should be done. "

The ADVANCE study chose a comparison group that was frequency matched based on qualities in the exposed/injured group: namely age, rank, deployment, role in theatre and regiment. These characterise 'battle-involvement' as suggested. Younger age, lower ranks, those who deployed on more combat-involved deployments, or worked in specific roles (e.g. bomb disposal, infantry) or for certain regiments (e.g. The Royal Engineers, Marine Commandos) were more represented in the injured group and were thus frequency matched to the comparison group. They were not matched separately to the amputees, VSI/SI or other groups, but as one group.

Point 7: "I assume that corresponding to the precipitating event, you are using the LAST time of service in Afghanistan as the event time for uninjured controls? All who were injured are only eligible to be cases, and presumably the injury, if serious, was incurred in the last deployment?

Thank you for the question. You are correct in that essentially in all those who were injured, the injury occurred during that individual's last deployment; few injured service personnel deployed again later. For the comparison, uninjured group they were frequency matched for deployment as well as the other factors discussed in point 6. Therefore, for example, the uninjured group were selected so that the proportion of personnel deployed on Herrick 8 (a 6 month deployment to Afghanistan) was the same in both the injured and uninjured groups and also frequency matched for the other factors. For the comparison group this was not necessarily their last deployment to Afghanistan. It is therefore not the last time of deployment, but rather a specific deployment of interest we are using to frequency match between groups. We have added details in our manuscript, in the methods: recruitment section, to highlight this. (see below)

Page 7: "Deployment refers to a specific deployment period of interest. For the exposed (injured) group, this is the deployment period in which they sustained their injury. The unexposed (comparison) group were frequency matched based on deploying within the same period without sustaining a physical combat related injury."

Point 8: "Analysis: How is the "frequency matching" handled in the analysis? If you had randomly sampled controls, the usual statistical analyses, t-test, ChiSquare test etc., would suffice. If, at the other extreme, you had created matched pairs, one control matched to one case, you would have to use matched pairs t-tests instead of t-test, McNemar test instead of ChiSquare, etc. Not only is the analysis different, but the research question changes. You appear to be considering use of the Proportional Hazards Model (Hazard Ratio), but to what event, with timing relative to what initial time point (entry to the study? Age at injury/last deployment?, Chronological Age?). I am not aware of any version of this model for use with paired or matched data. While the sample size is based on time to events, most of the measures you propose are ordinal. On page17, you propose repeated measures ANOVA and Friedman tests, neither of which is appropriate for use with time-series data. It is known that use of these approaches result in exaggerated p-values (false positives). You might consider Random Regression Linear Models with autoregressive covariance structure between multiple time points within a subject. What you are doing is somewhere in between, not as close as pairwise matching, not as independent as random selection. It's not exactly clear that using the usual tests is correct, and clearly the matched tests are not appropriate. This issue should be discussed and dealt with."

We apologise that these points were not clear in the original manuscript and have revised the manuscript accordingly. The matching is frequency matching and not 1:1 individual matching; matched pairs t-tests are not appropriate in this case. Thank you for your point on repeated measures; we have amended the methods as we will be considering the use of Mixed Effects Models.

Point 9: "On page 7, you also say "whilst adjusting for confounders". Do you really mean "confounders", or are you just throwing in covariates? A "confounder" is defined as a variable that is associated both with the independent variable of interest, here case/control, and with the outcome of interest. That automatically raises collinearity issues in regression analyses, and then there is the problem of possible interaction between the independent variable and the "confounder". If you are thinking of multiple confounders, there is further concern about the correlations among the multiple confounders and their interactions as well. This may seem a casual remark, but I would strongly urge you to think carefully about this."

All potential confounders will be carefully considered; some will be selected a priori based on previous literature/knowledge, others will be identified during statistical analysis based on their confounding effects. We are aware of problems with multicollinearity and mention in our protocol that we will use variance inflation factor postestimation tools to review models and ensure that this is not an issue.

Point 10: "The lack of specific 'a priori' hypotheses, and the multiplicity of outcome measures, suggests that this is an exploratory study, designed to generate hypotheses to be tested in future independent studies. But then the analytic plans seem to suggest that you are testing unspecified hypotheses without stated rationale and justification. Please consider before you start exactly what the goals of this study are. Once you explore, all tests are 'post hoc' and p-values and conclusions likely invalid."

ADVANCE is powered to address three specified 'a priori' hypotheses (see manuscript) which are based on the available - but inconclusive - literature. It would, however, be a wasteful cohort that did not also test other, 'secondary' hypotheses that are both plausible and relevant; we will treat these not as 'invalid' but as hypothesis-generating.

Point 11: "Some of these measures appear to be only appropriate for amputees, thus only for one stratum of cases, and for none of controls. These are useless for comparisons between cases and controls."

It is acknowledged that some of the outcome measures used are specific to amputees; the "exposed group" of the cohort will have a significant number of amputees in it. The amputee-specific outcomes are important and clinically relevant data to collect especially in relation to data/publications in other amputee groups. Thus, some analyses, of outcomes such as "amputee stump pain" (which is only collected in amputees) will be undertaken only in this group.

Point 12: "All the results, except for two measures related to brain injury are self-reported. Since all know whether they are cases and controls, there is a problem here of possible bias, i.e., that the answer reflects not the physical or psychological state, as much as a different mind set in answering questionnaires. It would be very useful to do a pilot study or two to generate information comparing these self-reports with objective measures to validate. Could you not have access to Health Service records to validate some of these? "

There are potential biases when using questionnaire data, as the reviewer describes; this is of course evident with all studies that use questionnaires. The ADVANCE study however also relies on medical records and clinical databases to collect data about original injury, treatment, and military demographic data. The study includes many objective health measures, such as those derived from the Vicorder, X-rays, blood tests and DEXA. All of our self-reported questionnaires are validated and the vast majority have previously been used in military studies (relevant citations are provided in the manuscript).

Point 13: “Table 2: The inclusion/exclusion criteria should apply both to cases and controls. Controls were not, e.g., “Injured between 2003 and the end of 2014”. You might put the general inclusion/exclusion criteria that apply to both groups first, and then add the criteria that differentiate cases from controls at the bottom of this table, starting with what you label “Comparison group only”. When you say “Past…” you mean before or immedidately after the last deployment? If “Past” includes, e.g., diabetes diagnosed 5 years after service in Afghanistan, you will be excluding the long-term outcomes this study was meant to examine. However, the “infection” exclusion seems to apply to the initial assessment in 2016 or later, and probably to all subsequent assessments?”

We have put all exposed/comparison group specific criteria at the bottom of this table and have covered the fact that deployment refers to a specific deployment upon which they were frequency matched. We note that we did not make clear in this table that ‘past medical history of’ refers to medical history prior to the deployment of interest. We have rectified this in the table.

The “infection” exclusion is temporary to exclude participants with acute infection at the time of data collection as this may affect some of the measurements. Anyone with an acute infection will be invited back to attend once this has resolved; we have clarified this in the manuscript, Table 1.

Point 14: “Table 3: Define the three strata here with sample sizes 257, 304 and 839. Alternatively since you take all amputations and VSI/SI, you could have two strata, with sample sizes 561 and 839. I would think three strata would make more sense, because some outcome measures are specific to amputations. “

We thank you for your notes and presume them to refer to figure 1. We have defined the strata as suggested and retained all three of them.

Point 15: “Table 4: “Duplicate records….” Please specify that it was the last Afghanistan deployment that is used to be comparable to the cases. I do not understand how you are creating a “frequency matched” sample. . I wonder whether other readers would. You do need to do this separately for the 2 or 3 strata defined by your sampling”

Thank you. We presume the reviewer is referring to Figure 2. Please see answer to Point 6 & 7 re: “last deployment” and “frequency matching”. Duplicate records refer specifically to those from multiple deployments during the same operation (e.g. deploying twice within Herrick 4). Since it is the specific deployment period that is of interest, duplicate records within this deployment period were removed from the frequency matching process as described.

Point 16: “The last table (listing all outcomes measured) is what suggests that this is an exploratory study. Some effort should be made to label those measures only appropriate for amputees, and those measures only appropriate for non-amputees in the text of the table. This table should be shortened simply to list Baseline Only measures, and Longitudinal Measures (at baseline 3,5,10,15,20 years). There is no need to have separate columns for the multiple time points. The table should be much shorter. “

We thank you for your comment and have adjusted the table as suggested.

Point 17: “Before analysis of outcomes, there should be some effort to investigate the inter-correlations of these measures, and to reduce the long list to relatively independent measures. As it is, if you test every measure separately, and perhaps each measure at each time point (which would undermine the longitudinal aspect of the study), such multiple testing guarantees many false positive results. At the same time, with the stratified sample, with the "frequency matching", and the likely missing data, there will also be a lot of false negative results. At that point, all your conclusions will be questionable."

We agree that there is always the risk of false positives, however, as we stated in point 11, no outcome measures are included which are entirely exploratory in nature; all are based on previous scientific knowledge. We do not understand why analysis of data at different time points undermines the longitudinal nature of the study, and nor how frequency matching will produce false negative results. We will of course attempt to investigate the inter-correlations of these measures.

Reviewer 4: Dr Deepal Nag Ayyala

Point 1: "Would help to include what method of imputation (missing at random or missing completely at random) will be used during the analysis."

As we mention in the methods, we will consider the use of multiple imputation. As for the exact method, this cannot be established until all the data have been collected and we can assess the nature of those that are missing.

Point 2: Address how the p-values will be adjusted for multiple comparisons (Bonferroni's correction or false discovery rate).

We have added in more detail on multiple comparisons in the statistical methods section.

Point 3: "Fix the reference error on Page 11, line 5."

The authors thank you for your comment. We have addressed the issue in the revised manuscript.

VERSION 2 – REVIEW

REVIEWER	Helena Chmura Kraemer Stanford University (Emerita) USA
REVIEW RETURNED	30-Jul-2020

GENERAL COMMENTS	I look forward to see what insights this study will bring to the problem.

REVIEWER	Deepak Nag Ayyala Medical College of Georgia, Augusta University, USA
REVIEW RETURNED	11-Aug-2020

GENERAL COMMENTS	The reviewer has no further comments.