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)	Fracture in the Elderly Multidisciplinary Rehabilitation (FEMuR): a phase II randomised feasibility study of a multidisciplinary rehabilitation package following hip fracture
AUTHORS	Williams, Nefyn; Roberts, Jessica; Din, Nafees; Totton, Nikki; Charles, Joanna; Hawkes, Claire; Morrison, Val; Hoare, Zoe; Williams, Michelle; Pritchard, Aaron; Alexander, Swapna; Lemmey, Andrew; Woods, Robert; Sackley, Catherine; Logan, Phillipa; Edwards, Rhiannon; Wilkinson, Clare

VERSION 1 - REVIEW

REVIEWER	Dr Toby Smith University of East Anglia, Norwich, UK.
REVIEW RETURNED	11-May-2016

GENERAL COMMENTS	Thank you for the opportunity to review this interesting and timely paper on rehabilitation following hip fracture, funded by NIHR. Overall I recommend the authors make major amendments to the paper. These are largely around reporting of the methods (particularly intervention) and results. If there is an issue with word count limit, I recommend the authors consider the use of online supplementary files. I have itemised the recommended changes below in order of their presentation in the manuscript. I hope these are useful to improve the reporting of this important feasibility study. The same comments have also been attached in a Word document for your records. Abstract Abstract: Setting: I would delete the term “one secondary care site encompassing” as this is redundant. Abstract: Participants: The authors need to provide an indication as to the threshold for ‘mental capacity’. Was this AMTS score? If so, state. Abstract: Intervention: Please provide more information if possible. At least indicate that this was home-based, and who delivered it and what the predominant component was (? Functional Exercise ?) Results: Do you need to present the QALY finding in the abstract? There is no reference to health economic outcomes anywhere else in the Abstract so it doesn’t correspond. Either include it and make reference to this analysis in the Methods, or remove completely. Strengths and limitations of this study

	Bullet 2 and 3: Bullet points 2 and 3 should be the first two bullet points. Bullet 4: The exclusion of people with cognitive impairment due to ethical reasons is not a reason. Please change this. It is possible to recruit people with dementia into trials. The recruitment of this population needs to be considered in great depth and is more complex but is certainly not impossible. The fact that the authors decided not to recruit this population in this feasibility study means that they have not tested the feasibility of recruiting people with cognitive impairment. This is a very different point to that stated in Bullet 4 and therefore should be modified to reflect this. Background Background: Line 10 regarding projected increase in hip fractures. This needs a citation to support this statement. Background: Lines 18 and 23 regarding mortality estimate and return to function level estimate. These are somewhat conservative. There is a body of literature which would suggest that this is much higher. I would recommend providing a range of values for these estimates to improve the readers awareness of the impact hip fracture can have (and therefore the potential importance of rehabilitation) for this population. Background: Line 49: should the text be revised to "…insufficient evidence of overall effectiveness or cost-effectiveness of an overall care pathway"? Background: Line 58 the authors make reference to the first phase of this study was the development of the pathway and the second phase is then the feasibility testing of this. The second phase is reported in this paper but have the authors previously reported the first phase. The reference [9] is attributed to Craig's publication on MRC complex intervention development whereas I was expecting this to relate to the development paper for this intervention? The text may need amended to clarify this. Study Objectives: Study Objective 4: This is regarding exploration of methods for economic evaluation. Based on this, I did not expect to see an analysis of economic data. The abstract (and later in the paper) presents QALY. This therefore left me a little concerned that the formal assessment of QALY and cost-consequence was not originally planned. I would therefore recommend that the authors and research team consider if the study objective presented here are accurate or whether they have over analysed/over reported in this case in respect to this objective. Study Design: Study Design: Line 44: Whilst it later becomes clear, I was initially confused as to the "anonymous cohort study" terminology. Basically you tested the feasibility (and representativeness) of your recruitment process for the feasibility study. The current text makes this seem far more complex than it is in reality. I think the text needs revising and a more clear and succinct reporting of this process is required. This is also evident in the next page (lines 20-25).

	There is no reference at all to the sample size. Whilst this is a feasibility study and a formal power calculation would be inappropriate, it is important that you recruited sufficient numbers of participants to answer all your feasibility study uncertainties and questions. There are numerous methodological papers which can inform this, and I recommend that you report how you estimated the sample size in this paper, supported by the literature. Inclusion Criteria Inclusion Criteria: Bullet 5: How did you determine capacity to give consent? How was this measured? Who did this? When and where was this done? This is really important as cognitive impairment (either delirium or dementia) is incredibly common in this population. Randomisation Randomisation: Line 21: the authors have stratified by hospital. However could they please provide further information for why this was important? There is no information to contextualise this reason based on size of hospital, socioeconomic factors for the locality, geographical proximity to other health and social care provision, surgeon volume and expertise, standard care and co-interventions? This would be valuable to better understand why this was important. Study Intervention There is insufficient information to ascertain what actually happened to these participants. It is not clear what the experimental intervention was, who delivered this (profession and grade), where this was delivered, the adherence to this, the frequency and timings of the intervention. It is not clear what the workbook and goal-setting diaries were, who discussed these to the participant and whether they were reinforced or monitored during the intervention contacts? Could these have been included as Supplementary Files? There is no reference to the care pathways which participants in both groups received i.e. what was the surgical pathway, what was the initial rehabilitation and recovery pathway? Similarly, it is not clear what the standard care/control intervention arm was. What was the intervention? Who delivered this? Was there variability across your three hospital sites? What measures were made to ensure some form of standardisation (or not)? If this information has not been presented due to word count/limit issues, then I strongly recommend the authors use Supplementary Files or online repositories to ensure that the reader can appreciate what was actually provided to participants. Intervention: Inclusion criteria 3 and intervention section: Based on the current reporting of eligibility and cohort characteristics, it is possible that you recruited both dynamic hip screw and cannualted screw fixation patients into the same cohort, reporting these as internal fixation. Please provide information on the weight-bearing status for these patients as some surgeons will recommend their patients who undergo cannualated screws are either non- partial- or full-weight bearing dependent on their recommendation and/or the quality of the fixation. Weight-bearing status can have a significant impact on the early rehabilitation capability of this population. Therefore this may have been a source for between group variability and may have been a source of adaptation of the rehabilitation

	pathway for people in either group. Reporting the weight-bearing status and between-group incidence of these procedures would therefore be valuable. Outcome measure There is limited information on where and who collected the baseline data. Please document the proportion of your sample where outcome measures were collected in the participant's place of residence compared to the physiotherapy department. This is very important when interpretation the feasibility of your data collection methods. Outcome measures: Was the ATMS a screening score or a follow-up score? It could be both, but I wanted to check that this was not a mistake to ensure that you did not just use it as part of your screening for eligibility, given that you only recruited non-cognitively impaired participants. Trial Analysis Please provide justification for conducting an ANCOVA and Student T-Test. This is a feasibility study. Do you therefore need to test your outcomes with this test. Are you really powered to do so? Was this really one of your study objectives? I agree with the calculation of an effect size, but the ANCOVA and Student T-Test seem less appropriate. Please reconsider this or provide a justification in response. Similarly, as suggested below, I am concerned that it was not the initial intention to assess QALY and this is a later addition. If this is the case, please acknowledged this and be consistent with the message in your reporting of the whole study throughout the paper. Focus Group The reporting of your qualitative component to the study is insufficient. Please provide information on sampling strategy and frame, what the topic guide was (and append as a supplementary file if possible), and what analysis approach was taken. In respect to the MDT focus groups, what did this consist of, what professions, what grades and experiences and geographically where? Similarly the patient and carer group, please provide more information on who these were in respect to their characteristics, their experiences and recovery pathways. This information is critical to provide the reader with a greater appreciation that this sample were (or were not) representative of the study. This is important as the Results and Discussion make considerable reference to these findings, and currently they are not substantiated because of this level of reporting in the Methods section. Results Cohort and Feasibility Study: Lines 17 and 31: Please provide data somewhere (either in the text or table) for the reasons for readmission for the cohort and the feasibility study. This is important. Feasibility Study: Baseline: "the main reason for ineligibility was lack of mental capacity" – it is not clear how this was determined, what

	the threshold was for this, and what the ATMS scores ranged from and to. Please consider providing more information on this. The frequencies of mild, moderate and severe cognitive impairment for instance would be useful to consider what the distribution of cognitive impairment was, particularly given that 49% of the cohort were excluded for this reason. Feasibility Study: Three month follow-up: What and in what group were the nine patient withdraws and six carer withdraws? Please provide this in the text. Feasibility Study: Three month follow-up: There is not assessment on the frequency of interventions, fidelity of the experimental intervention, adherence to either of the interventions. This is important so the reader understands that you really have tested the feasibility of this experimental intervention and the proposed standard care/treatment as usual intervention. This is very important as it is currently underreported. Feasibility Study: Three month follow-up: the authors reported that there were two readmissions and one death but do not provide information about these or in what groups they were. Please provide this. Feasibility Study: Three month follow-up: there is reference to 'minimal differences' and 'trend for a greater improvement' – I would avoid this terminology and just reported if the data reached a clinically significant difference or not and reported the effect size. For a feasibility study, this would answer all our important questions on intervention effect regarding feasibility. Feasibility Study: Three month follow-up: There is also a tendency not to interpret confidence intervals. For example, for 50 foot walk test, the interpretation is for a medium effect size, but the confidence intervals ranges from -0.41 or 1.20. This is considerable and should therefore be acknowledged. This is the case throughout the Results section. Economic analysis Please see my earlier comment regarding the value of assessing an economic analysis with 43 complete cases? Was this the a priori plan? Focus Group There are a number of points raised in the results of the focus group which will be supported and contextualised once there is more information on the sampling strategy for this part of the study, and from better reporting the intervention. It is difficult to detangle the MDT, patient and carer views. Was there variability in responses between respondents in your focus groups? The reporting suggests that this was uniform. Was this really the case, or were there any differing opinions and views which were accountable or non-accountable based on group allocation, respondent characteristic, geographical location (for example)? Discussion

	Summary of main findings: “the majority of patients required multiple visits prior to consent” – this is not reported in the Results. Please report data to support this earlier. Summary of main findings: “the new rehabilitation intervention was acceptable to patient and clinicians” – this is not explicitly stated in the Results – please state and provide evidence for this earlier. Strengths and Weakness: I think the authors have missed two major weaknesses. Firstly, there is no process evaluation. This is recommended by the MRC 2015 framework on complex interventions (Moore 2015 - http://www.bmj.com/content/350/bmj.h1258) and should be acknowledged. Secondly, there was no assessment of recruiting and consenting people with cognitive impairment. This questions the generalizability of the trial’s results as this morbidity is prevalent in this population (and was predictable when designing the study). Consideration should be made on the implication of this on the definitive trials but also on the external validity of your findings. Strength and Weaknesses: “to test the effectiveness or cost-effectiveness of the…” this is not a weakness, it was never the intervention of this study design to do this. I would delete that as a weakness. Strengths and Weaknesses: “struggled to understand the exercise self-efficacy scale” – where is this in the Results. Please report data to support this. Strength and Weaknesses: “The lack of statistical significant effect for the main outcome measure…” – the study was never powered to do this. I am therefore concerned that you are proposing to change you original primary outcome measure for this reason. Please reconsider the reasoning for this and I have considerable methodological issue with this suggestion. Strength and Weakness: No reference is made to the threat of intervention contamination and between-group contamination. Was this assessed? Was a cluster randomised controlled trial warranted? Was this considered during the analysis particularly in the focus groups? Comparison with previous literature: This whole section is very difficult to interpret whilst there is limited information on the experimental and control interventions used in this feasibility study. Once the reporting of these has been improved, it will be possible to make a more informed assessment of this section. Implications for future practice and research: “…include those who lack mental capacity…” can you really recruit these people for a definite trial when you have not tested the feasibility of recruiting and consenting these participants in a feasibility study. I think this needs more consideration as it is not as simple and the current text suggests. Implications for future practice and research: “..of the intervention to those lacking capacity” – how did you know this as you did not provide the experimental intervention to those who were not recruited and you only recruited those who were not cognitive impaired. I am confused.

	Implications for future practice and research: Changing the primary to the NEADL for the reasons you have provided is worrying. Please provide your reasoning for this. Please also provide a reference for the clinically significant change as this is not supported with a citation. Implications for future practice and research: the estimation for the sample size is based on the current data. This does not include people with cognitive impairment. It is not inconceivable that your standardised difference may be different for people with compared to without cognitive impairment. Therefore the number of 322 may not be correct for your definitive trial. Please reflect on this and consider whether the current text is adequate and accurate. Conclusions "Screening methods successfully identified most patients with a hip fracture" – I am not keen on 'most' as a term. Please quantify this with a percentage. "The intervention was acceptable to patients, carers and healthcare professionals" – I don't think you can actually state this based on the data reported in this paper. Please reconsider this and/or provide more data explicitly supporting this conclusion. "…primary outcome measure for a future definitive RCT" – as stated before, please clarify your rationale and clearly argue why the change in primary outcome is warranted to ensure that the most important outcome for patients, health professionals and stakeholders is selected, rather than the one which is going to provide the largest difference between the intervention arms. Tables Table 1: Please provide data on ASA grade and/or medical morbidities; mean AMTS or cognitive status score; length of hospital stay; whether participants lived alone or not pre- and post-discharge; reason for readmission; cause of death; and could mortality be presented as 30 and 90 day mortality. Table 2: Please provide data on ASA grade and/or medical morbidities; length of hospital stay; reason for readmission; cause of death; and could mortality be presented as 30 and 90 day mortality. Could the data on internal fixation for 'type of surgery' also be subdivided to cannulated screws and dynamic hip screw as this will have a potential impact on early rehabilitation capability. Table 2: I presume 'Direct Discharge' referred to home. Can you please clarify this?

REVIEWER	Donald R Hoover Rutgers The State University of New Jersey
REVIEW RETURNED	28-May-2016

GENERAL COMMENTS	While this reviewer does not have expertise in the specific topic area, from a methodological standpoint I find the study to overall be thorough, well thought out and informative. The Statistical Methods

	seem to be for the most part appropriate, but are confusing and in reading the article come across as randomly flying in from all over the place. • For example, why is bootstrapping used in some places for CIs and not in others.? While outcomes such as medical expenditures are probably skewed which could invoke need for the bootstrap, it is not so clear this is true for QALY and ICECAP that are also bootstrapped • Why are partial eta squares used to express Tx Arm differences in some Tables while “Effect Size” is used to do this in others. (I do believe that the differences seen in ANCOVA models can be expressed a “Effect Size” where ES_ANCOVA = ES_T-TEST * (1 – R2) where R2 is the square of the correlation between the variable at baseline and at follow up.) But maybe partial eta squared is standard for this setting. However, I find the squared scale and skewed CIs about the partial etas to be somewhat non-informative While the paper is already somewhat long, perhaps adding a separate section and maybe even a Table (perhaps expanding “Box 1”) explaining what statistical methods and comparators were used for what outcomes and why it was done would help. Other Comments A. In Table 4, a t-test does not seem appropriate for “50 foot walk” where there appears to be an outlier(s) for the intervention group that makes the standard deviation larger than the mean. Similarly for Table 3 the VAS, CSI and HADS scores also have means the same size as the standard deviation raising the question as to whether normal distribution based methods such as ANCOVA are appropriate B. Pages 31-32 Is there a Table number and Title? In that Table for XX(YY) the fact that YY corresponds to standard deviation was not specified. Or does (YY) mean something else?

VERSION 1 – AUTHOR RESPONSE

Reviewer 1:

Abstract
Abstract: Setting: I would delete the term “one secondary care site encompassing” as this is redundant.
Reply - This has been done.

Abstract: Participants: The authors need to provide an indication as to the threshold for ‘mental capacity’. Was this AMTS score? If so, state.
Reply - Mental capacity was assessed by the clinical team. The AMTS was not used to exclude participants, but rather to describe the cognitive functioning of participants at baseline. Following text added; “(assessed by their clinical team)”.

Abstract: Intervention: Please provide more information if possible. At least indicate that this was home-based, and who delivered it and what the predominant component was (? Functional Exercise ?)
Reply - Text amended; “including six additional home-based physiotherapy sessions delivered by a therapist or technical instructor”.

Results: Do you need to present the QALY finding in the abstract? There is no reference to health economic outcomes anywhere else in the Abstract so it doesn't correspond. Either include it and make reference to this analysis in the Methods, or remove completely.
Reply - Text amended; "economic evaluation" added to objective. EQ-5D and costs added to outcome measures.

Strengths and limitations of this study
Bullet 2 and 3: Bullet points 2 and 3 should be the first two bullet points.
Reply - Bullet points moved

Bullet 4: The exclusion of people with cognitive impairment due to ethical reasons is not a reason. Please change this. It is possible to recruit people with dementia into trials. The recruitment of this population needs to be considered in great depth and is more complex but is certainly not impossible. The fact that the authors decided not to recruit this population in this feasibility study means that they have not tested the feasibility of recruiting people with cognitive impairment. This is a very different point to that stated in Bullet 4 and therefore should be modified to reflect this.
Reply – We wanted to recruit patients lacking mental capacity, but were forbidden to do so by the NHS REC (letter attached). The bullet point has been amended to state that this was on the basis of the ethics committee decision.

Background
Background: Line 10 regarding projected increase in hip fractures. This needs a citation to support this statement.
Reply – Citation added

Background: Lines 18 and 23 regarding mortality estimate and return to function level estimate. These are somewhat conservative. There is a body of literature which would suggest that this is much higher. I would recommend providing a range of values for these estimates to improve the readers awareness of the impact hip fracture can have (and therefore the potential importance of rehabilitation) for this population.
Reply – Further citations have been added to represent the range of mortality.

Background: Line 49: should the text be revised to "…insufficient evidence of overall effectiveness or cost-effectiveness of an overall care pathway"?
Reply – Thank you; 'overall care pathway' added

Background: Line 58 the authors make reference to the first phase of this study was the development of the pathway and the second phase is then the feasibility testing of this. The second phase is reported in this paper but have the authors previously reported the first phase. The reference [9] is attributed to Craig's publication on MRC complex intervention development whereas I was expecting this to relate to the development paper for this intervention? The text may need amended to clarify this.
Reply – A paper describing the development of the intervention has been submitted but has not yet been accepted. However, more detail of the intervention is given in our published protocol, which is also referenced here. A copy of the previously published protocol paper is attached.

Study Objectives:
Study Objective 4: This is regarding exploration of methods for economic evaluation. Based on this, I did not expect to see an analysis of economic data. The abstract (and later in the paper) presents QALY. This therefore left me a little concerned that the formal assessment of QALY and cost-consequence was not originally planned. I would therefore recommend that the authors and research

team consider if the study objective presented here are accurate or whether they have over analysed/over reported in this case in respect to this objective.
Reply –
The objective of the economic evaluation was two-fold; we set out to explore methods by using both a generic health–related quality of life measure (EQ-5D) and a capability approach measure (ICECAP-O). We also planned from the outset to undertake further exploratory economic analyses. Objective 4 has been expanded to clarify this.

Study Design:
Study Design: Line 44: Whilst it later becomes clear, I was initially confused as to the "anonymous cohort study" terminology. Basically you tested the feasibility (and representativeness) of your recruitment process for the feasibility study. The current text makes this seem far more complex than it is in reality. I think the text needs revising and a more clear and succinct reporting of this process is required. This is also evident in the next page (lines 20-25).
Reply – This paragraph has been re-worded to clarify the objective of the cohort study.

There is no reference at all to the sample size. Whilst this is a feasibility study and a formal power calculation would be inappropriate, it is important that you recruited sufficient numbers of participants to answer all your feasibility study uncertainties and questions. There are numerous methodological papers which can inform this, and I recommend that you report how you estimated the sample size in this paper, supported by the literature.
Reply – Thank you, the following sentence with reference has been added; "In order to estimate the standard deviation of the primary outcome measure to be used in a power calculation for a future definitive RCT, we aimed for a sample size of 50 participants completing the study."

Inclusion Criteria
Inclusion Criteria: Bullet 5: How did you determine capacity to give consent? How was this measured? Who did this? When and where was this done? This is really important as cognitive impairment (either delirium or dementia) is incredibly common in this population.
Reply – This is described in more detail in the published protocol paper (attached). Capacity to give informed consent was assessed by a nurse or physiotherapist in the clinical team. Patients with post-operative delirium were approached if this was resolved prior to discharge from the acute hospital. The bullet point in the inclusion criteria has been amended to include this.

Randomisation
Randomisation: Line 21: the authors have stratified by hospital. However could they please provide further information for why this was important? There is no information to contextualise this reason based on size of hospital, socioeconomic factors for the locality, geographical proximity to other health and social care provision, surgeon volume and expertise, standard care and co-interventions? This would be valuable to better understand why this was important.
Reply – Each hospital was thought to have differing usual care and practices,
as well as different staff that were to complete the intervention. It was thought these unknown differences could have an effect on the study and so stratification by hospital was included. In addition, logistically we wanted an equal allocations as possible over the centres to reduce staff burden. This has been clarified in the text.

Study Intervention
There is insufficient information to ascertain what actually happened to these participants. It is not clear what the experimental intervention was, who delivered this (profession and grade), where this was delivered, the adherence to this, the frequency and timings of the intervention. It is not clear what the workbook and goal-setting diaries were, who discussed these to the participant and whether they

were reinforced or monitored during the intervention contacts? Could these have been included as Supplementary Files? There is no reference to the care pathways which participants in both groups received i.e. what was the surgical pathway, what was the initial rehabilitation and recovery pathway? Similarly, it is not clear what the standard care/control intervention arm was. What was the intervention? Who delivered this? Was there variability across your three hospital sites? What measures were made to ensure some form of standardisation (or not)? If this information has not been presented due to word count/limit issues, then I strongly recommend the authors use Supplementary Files or online repositories to ensure that the reader can appreciate what was actually provided to participants.
Reply – This section has been amended to clarify what happened to both control and intervention participants. The intervention is described in detail in the published protocol paper (attached) and a reference added to this. The workbook and diary have also been added as supplementary files.

Intervention: Inclusion criteria 3 and intervention section: Based on the current reporting of eligibility and cohort characteristics, it is possible that you recruited both dynamic hip screw and cannualted screw fixation patients into the same cohort, reporting these as internal fixation. Please provide information on the weight-bearing status for these patients as some surgeons will recommend their patients who undergo cannualated screws are either non- partial- or full-weight bearing dependent on their recommendation and/or the quality of the fixation. Weight-bearing status can have a significant impact on the early rehabilitation capability of this population. Therefore this may have been a source for between group variability and may have been a source of adaptation of the rehabilitation pathway for people in either group. Reporting the weight-bearing status and between-group incidence of these procedures would therefore be valuable.
Reply – Unfortunately we do not have the information on what type of internal fixation was used, as this was not collected at baseline. The data collection forms for a definitive trial will be refined accordingly to collect this data. Similar numbers of patients received internal fixation (7 in control and 10 in the intervention) and therapists used their professional judgement as to when was the best time to administer intervention sessions so that these could be used appropriately according to weight-bearing status.

Outcome measure
There is limited information on where and who collected the baseline data.
Reply – This is already described; “Outcome measures (Box 1) were completed by participants, assisted by a member of the research team who was blinded to treatment allocation, at baseline and at three month follow-up. Baseline measures were completed as soon as possible after surgery on the acute orthopaedic ward, inpatient rehabilitation ward or in the patient’s home following discharge.” More detail is available in the published protocol paper.

Please document the proportion of your sample where outcome measures were collected in the participant’s place of residence compared to the physiotherapy department. This is very important when interpretation the feasibility of your data collection methods.
Reply – Location of outcome measurement was according to patient preference, which has been added to the text.

Outcome measures: Was the ATMS a screening score or a follow-up score? It could be both, but I wanted to check that this was not a mistake to ensure that you did not just use it as part of your screening for eligibility, given that you only recruited non-cognitively impaired participants.
Reply – This was used to measure cognitive functioning at the baseline assessment. It was not used to screen patients for mental capacity, which was assessed by clinical staff according to whether the participants could give informed consent. We did recruit cognitively impaired patients if they retained mental capacity. It was not used as an outcome measure at follow-up, because participants were asked to complete a large number of outcome measures at follow-up and this measure was not

prioritised.

Trial Analysis
Please provide justification for conducting an ANCOVA and Student T-Test. This is a feasibility study. Do you therefore need to test your outcomes with this test. Are you really powered to do so? Was this really one of your study objectives? I agree with the calculation of an effect size, but the ANCOVA and Student T-Test seem less appropriate. Please reconsider this or provide a justification in response.
Reply – More detail on the analysis plan has been reported in our published protocol paper. Reference to protocol paper added. The analysis completed was intended as exploratory and used to calculate a representative effect size for the outcome measures. The study was not powered to detect a difference between the groups and this was never intended from the analysis, the wording in the paper has been amended to make this clearer.

Similarly, as suggested below, I am concerned that it was not the initial intention to assess QALY and this is a later addition. If this is the case, please acknowledged this and be consistent with the message in your reporting of the whole study throughout the paper.
Reply – We intended to explore and report outcomes from the economic measures administered from the outset. The economic objective has been expanded to state this, and further detail of our planned analysis is described in our protocol paper.

Focus Group
The reporting of your qualitative component to the study is insufficient. Please provide information on sampling strategy and frame, what the topic guide was (and append as a supplementary file if possible), and what analysis approach was taken. In respect to the MDT focus groups, what did this consist of, what professions, what grades and experiences and geographically where? Similarly the patient and carer group, please provide more information on who these were in respect to their characteristics, their experiences and recovery pathways. This information is critical to provide the reader with a greater appreciation that this sample were (or were not) representative of the study. This is important as the Results and Discussion make considerable reference to these findings, and currently they are not substantiated because of this level of reporting in the Methods section.
Reply – More detail on the focus groups' method has been reported in our published protocol paper. Reference to protocol paper added and description of the methods and analysis revised for clarity. The topic guides used have been added as a supplementary file. A table detailing the participant demographics, clinical role (where applicable) and geographic location of groups has been added to the results.

Results
Cohort and Feasibility Study: Lines 17 and 31: Please provide data somewhere (either in the text or table) for the reasons for readmission for the cohort and the feasibility study. This is important.
Reply – We have added a table containing this data as a supplementary file.

Feasibility Study: Baseline: "the main reason for ineligibility was lack of mental capacity" – it is not clear how this was determined, what the threshold was for this, and what the ATMS scores ranged from and to. Please consider providing more information on this. The frequencies of mild, moderate and severe cognitive impairment for instance would be useful to consider what the distribution of cognitive impairment was, particularly given that 49% of the cohort were excluded for this reason.
Reply – Mental capacity was assessed by clinical staff according to whether the participants could give informed consent, the AMTS was not used as a screening tool. The AMTS has not been used in the literature to categorise the severity of cognitive impairment but a cut off of 7 is generally used as a threshold to indicate the presence of cognitive impairment. The average baseline score of 9.1 supports our conclusion that we have recruited a healthier sub-population of our cohort with good

cognitive functioning.

Feasibility Study: Three month follow-up: What and in what group were the nine patient withdraws and six carer withdraws? Please provide this in the text.
Reply – There were nine withdrawals, one before baseline and eight during the intervention (four from each group). We have added the reasons for the withdrawals in a supplementary table.

Feasibility Study: Three month follow-up: There is not assessment on the frequency of interventions, fidelity of the experimental intervention, adherence to either of the interventions. This is important so the reader understands that you really have tested the feasibility of this experimental intervention and the proposed standard care/treatment as usual intervention. This is very important as it is currently underreported.
Reply – A paper reporting a process evaluation of the intervention is in preparation. The word count limit does not permit us to publish this here. We believe that the process evaluation would be more accessible as a separate publication rather than as a supplementary file.

Feasibility Study: Three month follow-up: the authors reported that there were two readmissions and one death but do not provide information about these or in what groups they were. Please provide this.
Reply – The group allocation of the readmissions and deaths has been added to the text. The reasons for re-admission have been added as a Supplementary Table along with information on cohort readmissions.

Feasibility Study: Three month follow-up: there is reference to 'minimal differences' and 'trend for a greater improvement' – I would avoid this terminology and just reported if the data reached a clinically significant difference or not and reported the effect size. For a feasibility study, this would answer all our important questions on intervention effect regarding feasibility.
Reply – We have amended the text accordingly, and on the advice of yourself and reviewer 2 have presented all differences as Cohen's d for consistency.

Feasibility Study: Three month follow-up: There is also a tendency not to interpret confidence intervals. For example, for 50 foot walk test, the interpretation is for a medium effect size, but the confidence intervals ranges from -0.41 or 1.20. This is considerable and should therefore be acknowledged. This is the case throughout the Results section.
Reply – Acknowledgement of wide 95% CI has been added.

Economic analysis
Please see my earlier comment regarding the value of assessing an economic analysis with 43 complete cases? Was this the a priori plan?
Reply – Text amended to an exploratory economic analysis. Published protocol referenced. As this was a feasibility trial, our a priori statistical analysis stated that we would not input missing data. The use of complete cases would be required for full economic analyses, for example cost-effectiveness analysis. Therefore, we felt it important to use the data in this way, which could inform future economic evaluations.

Focus Group
There are a number of points raised in the results of the focus group which will be supported and contextualised once there is more information on the sampling strategy for this part of the study, and from better reporting the intervention.
Reply –The methods sections has been amended which we hope will add clarity and context to the points raised in the results section. Reference to the previously published protocol paper has also been added to signpost to the detailed methodology which unfortunately cannot be presented here in

full.

It is difficult to detangle the MDT, patient and carer views. Was there variability in responses between respondents in your focus groups? The reporting suggests that this was uniform. Was this really the case, or were there any differing opinions and views which were accountable or non-accountable based on group allocation, respondent characteristic, geographical location (for example)?
Reply – We obtained rich data from the focus groups and have attempted to present the pertinent and recurring themes here within the confines of the word limit, though we have attempted to expand and clarify where possible. There was consensus around the results presented here and we aim to discuss these in further depth in our mixed methods process evaluation paper which is in preparation.

Discussion
Summary of main findings: "the majority of patients required multiple visits prior to consent" – this is not reported in the Results. Please report data to support this earlier.
Reply – The following text was added to the results section; "The majority of patients had two visits from researchers before they were recruited, and many requested a second visit to discuss the study after they had been discharged."

Summary of main findings: "the new rehabilitation intervention was acceptable to patient and clinicians" – this is not explicitly stated in the Results – please state and provide evidence for this earlier.
Reply – The following text was added to the focus group section; "The study was acceptable to patients, carers and therapists and the intervention was viewed positively. The most useful component was the extra sessions that participants received from the therapy teams. The goal-setting diary and workbook were seen as useful supporting documents by the majority of intervention group participants."

Strengths and Weakness: I think the authors have missed two major weaknesses. Firstly, there is no process evaluation. This is recommended by the MRC 2015 framework on complex interventions (Moore 2015 - http://www.bmj.com/content/350/bmj.h1258) and should be acknowledged.
Reply – A process evaluation was conducted but due to the word count will be reported separately. The following text was added to the discussion section; "A process evaluation was performed as part of the feasibility study and will be reported separately."

Secondly, there was no assessment of recruiting and consenting people with cognitive impairment. This questions the generalizability of the trial's results as this morbidity is prevalent in this population (and was predictable when designing the study). Consideration should be made on the implication of this on the definitive trials but also on the external validity of your findings.
Reply – The discussion has been amended to include; "The NHS REC did not permit us to recruit participants who lacked mental capacity, which has implications for the generalisability of our findings in this important group."

Strength and Weaknesses: "to test the effectiveness or cost-effectiveness of the…" this is not a weakness, it was never the intervention of this study design to do this. I would delete that as a weakness.
Reply – This has been deleted

Strengths and Weaknesses: "struggled to understand the exercise self-efficacy scale" – where is this in the Results. Please report data to support this.
Reply – A sentence has been added to the Results section stating; "General self-efficacy and self-efficacy for exercise had the lowest rate of completion at follow up, with patients expressing confusion to researchers conducting the interviews about how to complete these measures." A supplementary

table containing data on completion rates has also been added.

Strength and Weaknesses: “The lack of statistical significant effect for the main outcome measure…” – the study was never powered to do this. I am therefore concerned that you are proposing to change you original primary outcome measure for this reason. Please reconsider the reasoning for this and I have considerable methodological issue with this suggestion.
Reply – We hope that additional changes in the manuscript have clarified that the change from BADL to NEADL for primary measure is based on the ceiling effect observed in the BADL and that the NEADL was more appropriate for our recruited population. This section related to the justification of the use of a cost-consequence analysis rather than other economic analysis. We agree that the wording of this section may have been confusing and we have amended this to clarify.

Strength and Weakness: No reference is made to the threat of intervention contamination and between-group contamination. Was this assessed? Was a cluster randomised controlled trial warranted? Was this considered during the analysis particularly in the focus groups?
Reply – A section has been added to the strengths and weaknesses stating:
“Whilst the content of intervention and usual care sessions may have contained similar exercise activities, care pathways in this area did not use patient-led goal-setting or provide information on what to expect during recovery. In addition, the provision of usual care sessions was variable, whilst the intervention offered continuity and a definite number of sessions. These were major components of the intervention which were only available to participants randomised to that group. Whilst it is possible that other participants may have viewed information materials if they were used in group sessions, the one-to-one aspect of the delivery minimised the possibility of intervention contamination.” These aspects will be further explored in our process evaluation paper.

Comparison with previous literature: This whole section is very difficult to interpret whilst there is limited information on the experimental and control interventions used in this feasibility study. Once the reporting of these has been improved, it will be possible to make a more informed assessment of this section.
Reply – We hope that our amendments to clarify the intervention will add context to this section. In addition, we have re-worded the content and added additional references.

Implications for future practice and research: “…include those who lack mental capacity…” can you really recruit these people for a definite trial when you have not tested the feasibility of recruiting and consenting these participants in a feasibility study. I think this needs more consideration as it is not as simple and the current text suggests.
Reply – Text amended. The following was added; “Although we were not able to test the feasibility of recruiting these patients,”

Implications for future practice and research: “..of the intervention to those lacking capacity” – how did you know this as you did not provide the experimental intervention to those who were not recruited and you only recruited those who were not cognitive impaired. I am confused.
Reply – This was in the opinion of clinicians whom we interviewed in the focus groups and a sentence has been added to the focus group results section to reflect this.

Implications for future practice and research: Changing the primary to the NEADL for the reasons you have provided is worrying. Please provide your reasoning for this. Please also provide a reference for the clinically significant change as this is not supported with a citation.
Reply – The main reasoning behind changing the primary outcome measure to the NEADL was due to the ceiling effect seen within the Barthel. The baseline scores were 18/20 and it was agreed with our clinical experts that this is a common issue with the Barthel and the NEADL should be a more sensitive measure for assessing ADL. Text has been added to the strengths and weaknesses section

to clarify this and support the change of primary measure: “As a younger, healthier sub-population of the cohort was recruited to the feasibility study, baseline scores were high, causing a ceiling effect in this measure.”

Implications for future practice and research: the estimation for the sample size is based on the current data. This does not include people with cognitive impairment. It is not inconceivable that your standardised difference may be different for people with compared to without cognitive impairment. Therefore the number of 322 may not be correct for your definitive trial. Please reflect on this and consider whether the current text is adequate and accurate.
Reply – The number 322 is powered on the population recruited to this feasibility study. We plan to recruit 322 patients with mental capacity and as many as possible lacking capacity in addition. We have amended the text to say: “a full trial of similar design would need to recruit 322 participants.”

Conclusions

“Screening methods successfully identified most patients with a hip fracture” – I am not keen on ‘most’ as a term. Please quantify this with a percentage.
Reply – This value is 81% and has been added to the manuscript.

“The intervention was acceptable to patients, carers and healthcare professionals” – I don’t think you can actually state this based on the data reported in this paper. Please reconsider this and/or provide more data explicitly supporting this conclusion.
Reply – This conclusion was drawn from feedback given in the focus groups and a sentence has now been added to the results section to reflect this:
“The study was acceptable to patients, carers and therapists and the intervention was viewed very positively.”

“…primary outcome measure for a future definitive RCT” – as stated before, please clarify your rationale and clearly argue why the change in primary outcome is warranted to ensure that the most important outcome for patients, health professionals and stakeholders is selected, rather than the one which is going to provide the largest difference between the intervention arms.
Reply – We hope that the rationale for choosing NEADL is now clearer throughout the manuscript, we have also expanded this sentence to reiterate the problems observed with the ceiling effect in the BADL.

Tables

Table 1: Please provide data on ASA grade and/or medical morbidities; mean AMTS or cognitive status score; length of hospital stay; whether participants lived alone or not pre- and post-discharge; reason for readmission; cause of death; and could mortality be presented as 30 and 90 day mortality.
Reply – We did not collect data on ASA grade, AMTS, whether the patient lived alone pre or post-discharge for the cohort. We did however have the number who lacked mental capacity and mean hospital stay, which has been added to the text. The reasons for re-admissions has been included as a Supplementary Table.

Table 2: Please provide data on ASA grade and/or medical morbidities; length of hospital stay; reason for readmission; cause of death; and could mortality be presented as 30 and 90 day mortality. Could the data on internal fixation for ‘type of surgery’ also be subdivided to cannulated screws and dynamic hip screw as this will have a potential impact on early rehabilitation capability.
Reply – Unfortunately we do not have data on ASA grade, or type of internal fixation for feasibility study participants. Length of stay is discussed in the economics section. There was only one death in

the feasibility study, details of which have been added to the text. Reasons for readmission have been added to the supplementary files.

Table 2: I presume 'Direct Discharge' referred to home. Can you please clarify this?
Reply – Direct discharge refers to discharge to their usual place of residence from the acute hospital, with no stay in a community/rehabilitation hospital in the interim and has been amended in the table.

Reviewer: 2

While this reviewer does not have expertise in the specific topic area, from a methodological standpoint I find the study to overall be thorough, well thought out and informative. The Statistical Methods seem to be for the most part appropriate, but are confusing and in reading the article come across as randomly flying in from all over the place.

• For example, why is bootstrapping used in some places for CIs and not in others.? While outcomes such as medical expenditures are probably skewed which could invoke need for the bootstrap, it is not so clear this is true for QALY and ICECAP that are also bootstrapped
Reply: We set out in our analysis plan to produce confidence intervals around the economic outcomes of; costs; QALYs; and ICECAP capability indices to account for and report uncertainty. Costs, ICECAP capability indices and EQ-5D indices were all skewed at both baseline and follow-up, furthering the justification of bootstrapping to produce confidence intervals.

• Why are partial eta squares used to express Tx Arm differences in some Tables while "Effect Size" is used to do this in others. (I do believe that the differences seen in ANCOVA models can be expressed a "Effect Size" where ES_ANCOVA = ES_T-TEST * (1 – R2) where R2 is the square of the correlation between the variable at baseline and at follow up.) But maybe partial eta squared is standard for this setting. However, I find the squared scale and skewed CIs about the partial etas to be somewhat non-informative
Reply: Following comments from both reviewers we have amended the text for the analysis section and all effect sizes have now been converted to Cohen's d.

While the paper is already somewhat long, perhaps adding a separate section and maybe even a Table (perhaps expanding "Box 1") explaining what statistical methods and comparators were used for what outcomes and why it was done would help.
Reply: We hope that the amendments to the analysis section will provide clarity and that this point has now been addressed by the changes made.

Other Comments

A. In Table 4, a t-test does not seem appropriate for "50 foot walk" where there appears to be an outlier(s) for the intervention group that makes the standard deviation larger than the mean. Similarly for Table 3 the VAS, CSI and HADS scores also have means the same size as the standard deviation raising the question as to whether normal distribution based methods such as ANCOVA are

appropriate
Reply: Thank you for this comment, it is correct that there is one outlier for the 50 foot walk test which is skewing the data. As a sensitivity analysis, the outlier was removed and the group means and effect size recalculated. This has now been added to the report (Feasibility study – three-month follow-up section). The analysis was used purely to calculate an effect size and not formal hypothesis testing. The ANCOVA was used in order to gain adjusted mean differences to gain a more representative effect size from the population. As the VAS, CSI and HADS are generally evaluated using parametric testing, we believe that the issues with large standard deviations in this study are down to the small sample size.

B. Pages 31-32 Is there a Table number and Title? In that Table for XX(YY) the fact that YY corresponds to standard deviation was not specified. Or does (YY) mean something else?
Reply: The numbers for EQ-5D Utility Score, VAS and ICECAP capability are SDs. However, the brackets for QALY, change in ICECAP and change in cost are the 1,000 bootstrap confidence intervals which is stated in the table – due to the suggested table limit of the journal we collapsed the economic outcomes into one table. We have amended the table to clarify this issue.

VERSION 2 – REVIEW

REVIEWER	Dr Toby Smith University of East Anglia, Norwich, UK
REVIEW RETURNED	20-Jul-2016

GENERAL COMMENTS	Thank you for addressing the points raised following the first peer-review assessment. There are just a couple of recommendations which can be ignored or adopted. Well done.

REVIEWER	Donald R Hoover Rutgers The State University of New Jersey, United States of America
REVIEW RETURNED	26-Jul-2016

GENERAL COMMENTS	The reviewer completed the checklist but made no further comments.