PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	The Surveillance AFter Extremity Tumour surgerY (SAFETY) Trial: Protocol for a pilot study to determine the feasibility of a multi-centre randomized controlled trial
AUTHORS	Ghert, Michelle

VERSION 1 – REVIEW

REVIEWER	Professor Bernadette Brennan Royal Manchester Children's Hospital Oxford Road Manchester M13 9WL TUK
REVIEW RETURNED	08-Mar-2019

GENERAL COMMENTS	Page 11-Definitive Study We hypothesize that more frequent post-operative surveillance (compared to less frequent postoperative surveillance) and the use of post-operative CT scans (compared to CXR) in the first two years following the surgical excision of a STS will improve survival over five years. reviewer- I was surprised that after the introduction and presentation of the literature the authors where suggesting that increased follow up with CT may improve OS- I was expecting they were suggesting less follow up was as good - what is the accepted follow up - more or less ? Page 12 -Undergone surgical resection of the tumour with curative intent and grossly negative margins; reviewer - what is meant by this statement - R1 or R0 resection- please use international defined standards page 19-Participants who refuse to return for a study assessment will be asked if they are willing to provide follow-up data via telephone;- reviewer- how does this help or indeed subsitute for imaging follow up - what are they expected to report on the telephone ? If they fail to attend for imaging is this not lost to follow up Reviewer- should there be a discussion ? So whta is the estimate in year 1 as a minimum number of patients so as to carry on with recruitment - what measures could they discuss at that point to rescue the study While i cannot provide a statistical review - why 2/2 factorial design - what other designs could they consider espescially if the numbers are less that expected to recruit in this rare tuimour

	poulation - e.g. Baysian probabilty acceptioning various clinically meaningful senarios They should discuss other similar studies in other cancers which have been successfully completed on follow up - what were there differences or indeed similarities to justrify this study and design

REVIEWER	Ty Subhawong University of Miami, USA
REVIEW RETURNED	27-Apr-2019

GENERAL COMMENTS	The authors propose a prospective randomized trial to evaluate the efficacy of different distant surveillance strategies. Using a 2 x 2 factorial design, 4 different strategies (CT q3 or q6 months, Xray q3 or q6 months) will be studied over 2 years. The pilot phase of the study aims to show feasibility of performing a larger definitive study. Success of the pilot phase is defined as recruitment of 195 subjects within 2 years; protocol adherence and subject retention of at least 85%, and at least 95% completeness of participant follow-up data for definitive primary outcome. The primary endpoint for the definitive phase of the study will be 5-year overall survival, with a number of appropriate secondary endpoints (patient reported outcomes, local recurrence-free and metastasis-free survival times, etc.); the definitive study will aim to enroll 830 patients, although this number may be adjusted based on pilot study results. My primary concerns regard incidental or off-protocol CT scans (e.g. performed for cough or concern for pneumonia/PE), or equivocal chest radiographs that may then prompt further evaluation with CT. Would these patients then cross-over or would it be marked as a protocol deviation Overall, the trial is well-written and answer a critical question in musculoskeletal oncology.

REVIEWER	Mohamed Fahmy Al Azher University
REVIEW RETURNED	09-May-2019

GENERAL COMMENTS	Authors have to consider the duplication of results if they used both CXY and chest CT for surveillance, I think one radiological investigation in the form of chest CT scan may be enough, and they can consider another parameter. Authors didn't refer to a closely related study published at 2005 in the European Journal of Surgical Oncology (EJSO): https://doi.org/10.1016/j.ejso.2005.07.015,

REVIEWER	Lorenz Uhlmann Germany
REVIEW RETURNED	26-May-2019

GENERAL COMMENTS	General concerns: The authors present a protocol of a pilot study to determine the feasibilty of the SAFETY trial. The protocol is well written and outlines clearly the purpose and design of the trial. However, I have some comments that I would like to ask the authors to clarify. In general, I was sometimes struggling to clearly understand which phase/study you are talking about (pilot study or definitive trial). I

	would suggest to be more clear about that and use subsections more often (for example, in the study design section). Another general concern I have is about the transition of patients between the pilot study and the definitive trial. Is it actually an internal pilot study design or is it a separate pilot study? My last general concern is about the hypothesis tested in the definitive trial. Is it only one or is it more than one? These questions are mentioned below again to be more specific about my concerns. Specific concerns: Abstract: "Definite" protocol/phase. I am used to the term "phase III" trial or main trial. Is this related to fact that there is a transition between the two trials/phases? (see my comments below) Page 9, line 52: This section needs some clarification. In my opinion, the two study designs (pilot study vs main study) get mixed up a little. It would help if you separate both trials more clearly, for example, using subsections. Page 10, line 20-29: You state that you anticipate the duration of the pilot phase to be three years. Compared to the main study, this is short which might be absolutely fine. However, please clarify how you plan to assess part B) of the primary feasibility objectives, especially the "post-intervention phase visits". page 11, line 47 you state the following: "We hypothesize that more frequent post-operative surveillance (compared to less frequent postoperative surveillance) and the use of post-operative CT scans (compared to CXR) in the first two years following the surgical excision of a STS will improve survival over five years." How many hypotheses will be tested here? Please be more specific (see also my comment below). Page 15, line 40 to page 17, line 8: You list the primary outcome criteria in separate paragraphes which is very helpful. However, it would be easier for the reader if you would follow the same structure as provided on page 10, line 46 ff. Furthermore, in the abstract, you state that a composite primary endpoint is used. I assume that it is the composite of all these endpoints listed here. Please be more specific in this section about this issue and explicitely mention that a composite endpoint consisting of these endpoints is used. Page 17, line 12: My understanding was that this pilot study is planed to last for three years. How will a five-year survival be assessed? My apologies, if I misunderstood the total time of the pilot study. However, some clarification would help the reader (see also my comment above). Page 19, line 19: This exclusion criteria is very unspecific and might lead to some bias. What kind of criteria are you planning to apply? What do you think about the generalizability of the results? Please clarify.

	Page 20-21 (Section "Definitive Sample Size"): On page 11, line 47 you state the following: "We hypothesize that more frequent post-operative surveillance (compared to less frequent postoperative surveillance) and the use of post-operative CT scans (compared to CXR) in the first two years following the surgical excision of a STS will improve survival over five years." How many hypotheses will be tested here? (see comment above) If there is more than one hypothesis you should consider some adjustment for multiplicity. Please clarify (see also comment below). Again, page 20-21 (Section "Definitive Sample Size"): The sample size calculation presented is not clear. It is not clear to me how you obtained your results as there are different approaches for sample size calculation in survival time analysis and I am not sure if the assumptions you made are sufficient. Please provide more details about your approach (the method as well as the software used). Page 21, line 17: This sentence has to be revised. There is a verb missing. Page 21, line 26: You mention that the sample size may be adjusted based on the results during the pilot study. This makes totally sense. However, please provide more details about this idea. Which assumption will exactly be adjusted? Is it the mortality in the control groups? Will you also consider the rate observed in the non-control groups in this feasibility study? What about the percentage of loss to follow up? There are three more issues here: The term "pilot study" is used here for the first time (my apologies if I missed it before). My preference would be to keep the terms as consistent as possible and to use only one term throughout the manuscript (but this is only my personal taste). Further, I am struggling with the meaning of "transition from the feasibility to the definitive phase". Will the same patients actually be included in both, the feasibility and the definitive phase? If so, this is rather an internal pilot trial design. Please clarify. My last concern is that the sample size is quite high in the feasibility trial compared to the definitive trial (assuming that there is no actual transfer between the two trials). Please consider to lower the sample size of the feasibility trial if there is no actual transition. Page 21, Table 1: Why do you highlight and use the sample size of the two rates 45% vs. 35%. In the text (line 5) you write that the best estimate of the control group is 55%. My apologies is I misunterstood something here. Please clarify. Page 22, line 8: In this paragraph, you describe the analysis used in the definitive phase (see comment below). You state that two independent treatment comparisons between treatment groups will be made. Please be more precise and state which groups will be compared. (I know that this might be somewhat repetitive but I strongly assume this helps the reader to understand your approach.)

	Furthermore, I was wondering if you might want to consider the interaction between frequency and CXR vs. CT (maybe in a sensitivity analysis). My last point here that on page 10, line 10, you state the following: "In extremity STS patients who undergo surgical resection with curative intent, what is the impact of surveillance frequency (every three vs. every six months) and surveillance imaging modality (CXR vs. CT scan) on overall survival at five years?" However, here, to me it sounds more like you focus only on the difference between frequencies within each modality. Please further clarify your approach. Page 22, line 10: Which primary analysis do you refer to? I would assume that it should be the primary analysis of the feasibility analysis. However, you compare the overall 5-year survival which is the primary analysis of the definitive trial. I suggest to stay consistent and refer here to the primary objective of the feasibility trial. Page 24, line 24: See my comment above, I am still confused about the actual overall time of the feasibility trial. There is another one year of recruitment time. I think you did not take this into account on page Also, consider to use only either "feasibility" or "pilot" trial (see comment above).

VERSION 1 – AUTHOR RESPONSE

reviewer- I was surprised that after the introduction and presentation of the literature the authors where suggesting that increased follow up with CT may improve OS- I was expecting they were suggesting less follow up was as good - what is the accepted follow up - more or less ?

Thank you for this comment. We have added a statement in the section “best evidence for surveillance strategies” in which we state that earlier detection of metastatic disease may improve long-term survival but there is no definitive evidence to support this assumption.

The accepted follow up, based on the JNCCN 2018 guidelines on soft tissue sarcomas, suggest that for stage I tumors chest imaging with CT or radiographs be performed every 6 to 12 months for the first 2-3 years then annually, while stage II or III tumors should have chest imaging every 2 to 6 months for the first 2-3 years then annually. The guidelines submit that there is very little data available to support one surveillance strategy over another and that there has never been a study which proved that using CT scans during surveillance improved outcomes. This information has been added to the end of the section “magnitude of the problem”.

Page 12 -Undergone surgical resection of the tumour with curative intent and grossly negative margins;
reviewer - what is meant by this statement - R1 or R0 resection- please use international defined standards

Thank you. We have edited the statement to clarify that they are R1 or R0 resections.

Page 19-Participants who refuse to return for a study assessment will be asked if they are willing to provide follow-up data via telephone;-

reviewer- how does this help or indeed subsitute for imaging follow up - what are they expected to report on the telephone ? If they fail to attend for imaging is this not lost to follow up

This is an important point for clarification. The participants will be contacted by phone to deter-mine the primary endpoint of the study, which is survival. They will also be asked to complete study questionnaires over the phone. This clarification has been added to the text.

Reviewer- should there be a discussion ?

The final section, “potential impact of the study”, serves as a short discussion of the importance of the study for the purposes of specifically a protocol publication.

So whta is the estimate in year 1 as a minimum number of patients so as to carry on with recruit-ment - what measures could they discuss at that point to rescue the study

Thank you. This is an important point. Since we have considered the enrollment of 195 patients over 2 years to be a criteria for success of the pilot study, then we would aim to recruit approxi-mately 100 patients during the 1st year. If we fail to achieve 90% of that goal (90 patients) during the first year, then our plan to rescue the study will be to increase the number of participating sites. This has been added to the section “pilot study primary outcome”.

While i cannot provide a statistical review - why 2/2 factorial design - what other designs could they consider especially if the numbers are less that expected to recruit in this rare tumour popula-tion - e.g. Baysian probability acceptioning various clinically meaningful scenarios
They should discuss other similar studies in other cancers which have been successfully complet-ed on follow up - what were there differences or indeed similarities to justify this study and de-sign

The 2x2 factorial design is ideal and the most efficient approach to study two treatment interven-tions in a single randomized controlled trial, particularly when there is no biologic plausibility that the two interventions interact. This is unlike a scenario in which the two interventions are medica-tions that may have a synergistic or negative effect when combined. A Bayesian design would be useful do avoid the question of whether or not an interaction exists, however for the purposes of the present trial it is clear that no interaction exists between the frequency and intensity of surveil-lance. As Freidlin and Korn (JNCI 2017) discuss in their commentary, the 2x2 factorial design is an efficient design to evaluate two interventions in a cancer clinical trial when there are no inter-actions between treatments. This information has been added to the section “study design”.

Reviewer: 2
Reviewer Name: Ty Subhawong
Institution and Country: University of Miami, USA
Please state any competing interests or state ‘None declared’: consultant for Agios and Arog Pharmaceuticals

Please leave your comments for the authors below
The authors propose a prospective randomized trial to evaluate the efficacy of different distant surveillance strategies. Using a 2 x 2 factorial design, 4 different strategies (CT q3 or q6 months, Xray q3 or q6 months) will be studied over 2 years. The pilot phase of the study aims to show feasibility of performing a larger definitive study. Success of the pilot phase is defined as re-cruitment of 195 subjects within 2 years; protocol adherence and subject retention of at least 85%, and at least 95% completeness of participant follow-up data for definitive primary out-come. The primary endpoint for the definitive phase of the study will be 5-year overall survival, with a number of appropriate secondary endpoints (patient reported outcomes, local recurrence-free and metastasis-free survival

times, etc.); the definitive study will aim to enroll 830 patients, although this number may be adjusted based on pilot study results.

My primary concerns regard incidental or off-protocol CT scans (e.g. performed for cough or concern for pneumonia/PE), or equivocal chest radiographs that may then prompt further evalua-tion with CT. Would these patients then cross-over or would it be marked as a protocol devia-tion
Thank you for raising this concern. We have amended the section “minimization of crossover of surveillance interventions”. Patients that have incidental or off-protocol imaging done would not crossover. We would document as a protocol deviation. In the case of chest radiographs that warrant further investigation with a CT scan, we will document. If the patient is found to have disease recurrence, we will document how the disease recurrence was A) first identified; and B) confirmed. If after a CT scan the patient is found to not have disease recurrence, the patient will resume surveillance as per the arm to which they were randomized.

Overall, the trial is well-written and answer a critical question in musculoskeletal oncology.

Thank you.

Reviewer: 3
Reviewer Name: Mohamed Fahmy
Institution and Country: Al Azher University
Please state any competing interests or state ‘None declared’: ‘None declared’

Please leave your comments for the authors below
Authors have to consider the duplication of results if they used both CXY and chest CT for sur-veillance, I think one radiological investigation in the form of chest CT scan may be enough, and they can consider another parameter.

Thank you for your comment. Since it is a 2x2 factorial study, patients will be assigned to either CXR or CT, not both. This is clarified in the “Study interventions” section, whereby the four treatment groups are described.

Authors didn't refer to a closely related study published at 2005 in the European Journal of Sur-gical Oncology (EJSO):https://doi.org/10.1016/j.ejso.2005.07.015,

Thank you. The study referred to was a retrospective review of a single surveillance strategy with no comparison arm, and which included more bone sarcomas than soft tissue sarcomas, and therefore does not have specific relevance to this randomized prospective comparative study of specifically patients with soft-tissue sarcoma.

Reviewer: 4
Reviewer Name: Lorenz Uhlmann
Institution and Country: Germany
Please state any competing interests or state ‘None declared’: None declared

Please leave your comments for the authors below
General concerns:
The authors present a protocol of a pilot study to determine the feasibilty of the SAFETY trial. The protocol is well written and outlines clearly the purpose and design of the trial. However, I have some comments that I would like to ask the authors to clarify.

In general, I was sometimes struggling to clearly understand which phase/study you are talking about (pilot study or definitive trial). I would suggest to be more clear about that and use subsec-tions more often (for example, in the study design section).

Thank you for this comment. We apologise for the lack of clarity. The subsections have been re-named to reflect either the pilot study or definitive study.

Another general concern I have is about the transition of patients between the pilot study and the definitive trial. Is it actually an internal pilot study design or is it a separate pilot study?

Thank you for raising this important concern. The sentences pertaining to the transition of patients under the heading “Study design” were rephrased as follows: “Following the two-year interven-tion phase, study participants will continue to be assessed at regular intervals for an additional three years. As such, all pilot study patients will be transitioned into the definitive study and be in-cluded in it.”

My last general concern is about the hypothesis tested in the definitive trial. Is it only one or is it more than one?

Thank you. Given the 2X2 factorial design of the trial, there are two hypotheses: one for imaging modality used and one for followup frequency. The hypothesis section has been edited to reflect this.

These questions are mentioned below again to be more specific about my concerns.

Specific concerns:
Abstract: "Definite" protocol/phase. I am used to the term "phase III" trial or main trial. Is this related to fact that there is a transition between the two trials/phases? (see my comments below)

Thank you. As mentioned above, we have revised the language to be consistent for either“pilot study” or “definitive study” .

Page 9, line 52: This section needs some clarification. In my opinion, the two study designs (pilot study vs main study) get mixed up a little. It would help if you separate both trials more clearly, for example, using subsections.

Thank you. As mentioned above, the subsections have been renamed to reflect either the pilot study or definitive study.

Page 10, line 20-29: You state that you anticipate the duration of the pilot phase to be three years. Compared to the main study, this is short which might be absolutely fine. However, please clarify how you plan to assess part B) of the primary feasibility objectives, especially the "post-intervention phase visits".

Thank you for this important comment. We agree that we will not be able to assess feasibility for “post-intervention phase visits” on all patients in the pilot study. This was removed from the text.

page 11, line 47 you state the following:
"We hypothesize that more frequent post-operative surveillance (compared to less frequent post-operative surveillance) and the use of post-operative CT scans (compared to CXR) in the first two years following the surgical excision of a STS will improve survival over five years."
How many hypotheses will be tested here? Please be more specific (see also my comment below).

Thank you. This section has been clarified to demonstrate that there are two hypotheses, as is standard for the 2X2 factorial design.

Page 15, line 40 to page 17, line 8: You list the primary outcome criteria in separate paragraphes which is very helpful. However, it would be easier for the reader if you would follow the same structure as provided on page 10, line 46 ff. Furthermore, in the abstract, you state that a compo-site primary endpoint is used. I assume that it is the composite of all these endpoints listed here. Please be more specific in this section
about this issue and explicitely mention that a composite endpoint consisting
of these endpoints is used.

Thank you for this suggestion. The structure has been adjusted to that used on page 10.

The word “composite” was replaced with “combination” since we will be looking at the combi-nation of feasibility endpoints.

Page 17, line 12: My understanding was that this pilot study is planed to last for three years. How will a five-year survival be assessed? My apologies, if I misunderstood the total time of the pilot study. However, some clarification would help the reader (see also my comment above).

Thank you. Indeed, five-year survival cannot be assessed in the pilot study. We have rephrased as follows: “The main secondary outcome for the pilot study will be death from any cause.”

Page 19, line 19: This exclusion criteria is very unspecific and might lead to some bias. What kind of criteria are you planning to apply? What do you think about the generalizability of the results? Please clarify.

Thank you for this comment. This is indeed a subjective criteria; however, lack of compliance would affect the validity of the study. As a pragmatic study, individual site investigators will have the latitude to exclude a patient who is unlikely to followup in clinic or agree to investigations.

Page 20-21 (Section "Definitive Sample Size"): On page 11, line 47 you state the following:
"We hypothesize that more frequent post-operative surveillance (compared to less frequent post-operative surveillance) and the use of post-operative CT scans (compared to CXR) in the first two years following the surgical excision of a STS will improve survival over five years."
How many hypotheses will be tested here? (see comment above) If there is more than one hy-pothesis you should consider some adjustment for multiplicity. Please clarify (see also comment below).

These are two different hypotheses, as noted above, with patients serving as intervention and con-trols simultaneously in two different hypothesis tests. This is the reason for choosing a 2x2 factorial study design.

Again, page 20-21 (Section "Definitive Sample Size"): The sample size calculation presented is not clear. It is not clear to me how you obtained your results as there are different approaches for sample size calculation in survival time analysis and I am not sure if the assumptions you made are sufficient. Please provide more details about your approach (the method as well as the soft-ware used).

We have added the following sentence in the section “definitive study sample size” to clarify the design used: “Given that intensive surveillance will detect metastatic disease at an earlier stage, we will use a superiority design to compare survival between more versus less intensive surveil-lance.”

We have also added the software used (SPSS).

Page 21, line 17: This sentence has to be revised. There is a verb missing.

Thank you. This sentence was revised as follows: "With a desired power of 0.80, we calculated a sample size of 396 participants per study arm."

Page 21, line 26: You mention that the sample size may be adjusted based on the results during the pilot study. This makes totally sense. However, please provide more details about this idea. Which assumption will exactly be adjusted? Is it the mortality in the control groups? Will you also consider the rate observed in the non-control groups
in this feasibility study? What about the percentage of loss to follow up?

Thank you for raising this point. We plan on potentially adjusting the percent lost to followup as this may be apparent towards the end of the pilot study. Other factors such as the estimated con-trol group overall five-year survival, the clinically meaningful outcome, and power cannot be amended. We have added this information to the text in the end of the section “definitive study sample size”

There are three more issues here: The term "pilot study" is used here for the first time (my apolo-gies if I missed it before). My preference would be to keep the terms as consistent as possible and to use only one term throughout the manuscript (but this is only my personal taste).

Thank you for that comment. We agree that the inconsistency was confusing. We have edited the text in multiple sections to be consistent in using the term “pilot study”.

Further, I am struggling with the meaning of "transition from the feasibility to the definitive phase". Will the same patients actually be included in both, the feasibility and the definitive phase? If so, this is rather an internal pilot trial design. Please clarify.

Thank you. We have clarified in the text that all pilot study patients transitioned into the definitive study would be included in both.

My last concern is that the sample size is quite high in the feasibility trial compared to the defini-tive trial (assuming that there is no actual transfer between the two trials). Please consider to low-er the sample size of the feasibility trial if there is no actual transition.

Thank you. Yes, the sample size is quite high, however since patients in the pilot study will be tran-sitioned into the definitive study, we would not be spending time and effort to recruit a new cohort of patients to replace them in the definitive study. In order for the definitive study to be completed within a reasonable amount of time, we will need to be sure that a sufficient number of patients can be recruited for the pilot study. We have amended the pilot study primary outcome “recruit-ment measure” to include a plan to increase the number of participating sites as a rescue measure if needed.

Page 21, Table 1: Why do you highlight and use the sample size of the two rates 45% vs. 35%. In the text (line 5) you write that the best estimate of the control group is 55%. My apologies is I misunterstood something here. Please clarify.

Thank you. 45% and 35% are actually death rates, not survival. In the table it is clarified that “event rate = death”.

Page 22, line 8: In this paragraph, you describe the analysis used in the definitive phase (see comment below). You state that two independent treatment comparisons between treatment groups will be made. Please be more precise and state which groups will be compared. (I know that this might be somewhat repetitive but I strongly assume this helps the reader to understand your approach.)

Thank you. This has been clarified in the paragraph.

Furthermore, I was wondering if you might want to consider the interaction between frequency and CXR vs. CT (maybe in a sensitivity analysis).

Thank you. The premise of using the 2x2 factorial design was that there is no biologi-cal plausibility for interaction between the two treatment arms; therefore sensitivity analyses would not be relevant..

My last point here that on page 10, line 10, you state the following:
"In extremity STS patients who undergo surgical resection with curative intent, what is the im-pact of surveillance frequency (every three vs. every six months) and surveillance imaging modal-ity (CXR vs. CT scan) on overall survival at five years?"
However, here, to me it sounds more like you focus only on the difference between frequencies within each modality. Please further clarify your approach.

Thank you. We have edited that section by dividing the two questions to say the following: “We plan to assess the feasibility of conducting the pragmatic, international, multi-centre, 2X2 factorial Surveillance AFter Extremity Tumour surgerY (SAFETY) RCT that answers the following ques-tions: In extremity STS patients who undergo surgical resection with curative intent, what is (1) the impact of surveillance frequency (every three vs. every six months) on overall survival at five years, and what is (2) the impact of surveillance imaging modality (CXR vs. CT scan) on overall survival at five years?”

Page 22, line 10: Which primary analysis do you refer to? I would assume that it should be the primary analysis of the feasibility analysis. However, you compare the overall
5-year survival which is the primary analysis of the definitive trial. I suggest to stay consistent and refer here to the primary objective of the feasibility trial.

Thank you. We have moved this section to follow the section titled “Analysis of feasibility out-comes” and have renamed it to “Analysis of definitive study primary outcome”

Page 24, line 24: See my comment above, I am still confused about the actual overall time of the feasibility trial. There is another one year of recruitment time. I think you did not take this into account on page Also, consider to use only either "feasibility" or "pilot" trial (see comment above).

Thank you. We have edited the language to be more consistent throughout the manuscript and clarified the expected recruitment time for the pilot study.

VERSION 2 – REVIEW

REVIEWER	Lorenz Uhlmann Germany
REVIEW RETURNED	02-Aug-2019

GENERAL COMMENTS	I would like to thank the authors for addressing all my comments and for revising their manuscript. I am happy with the replies and the amendments of the manuscript with only few exceptions.

	The first concern I want the authors to further work on is the connection between the pilot study and the definitive study. There is a transition of the patients meaning that statistical analysis of the pilot study and the definitive study are based on (partially) the same patients. This is a multiple testing problem. As far as I understood, you assess in both studies the survival of patients. In the definitive study it is the primary endpoint. Therefore, I would suggest that you need to adjust your alpha level. Please clarify how you plan to deal with this problem. The second concern is another multiplicity issue in the definitive study itself. You state that you will test two hypotheses. I agree that a 2x2 factorial design is a good choice. However, you still might need to adjust your alpha level due to the multiple testing issue. This might also have an impact on the sample size calculation. A work-around solution would be to use a hierarchical testing procedure but I do not see or assume that you are planning to apply such a procedure. Please clarify. One additional minor comment: Please write “2x2” (or “2X2”) consistently throughout the manuscript.

VERSION 2 – AUTHOR RESPONSE

Reviewer 4:

The first concern I want the authors to further work on is the connection between the pilot study and the definitive study. There is a transition of the patients meaning that statistical analysis of the pilot study and the definitive study are based on (partially) the same patients. This is a multiple testing problem. As far as I understood, you assess in both studies the survival of patients. In the definitive study it is the primary endpoint. Therefore, I would suggest that you need to adjust your alpha level. Please clarify how you plan to deal with this problem.

Thank you for your comment. We will not be analyzing outcomes statistically in the pilot study so multiple testing does not apply in this scenario. As discussed by Moore et al. (2011), the purpose of the proposed pilot study is to investigate the process rather than the outcome. The pilot study sample size was determined based on adherence, rather than the outcome of survival. As such, statistical analysis with regards to the outcome of survival will only be performed for the definitive trial, though this data will be collected during the pilot phase in order to ensure continuity and smooth transition of these patients into the definitive trial. We have added a sentence in the ‘Pilot study primary outcome’ section to reflect this.

Also, in order to avoid confusion, the following sentence has been removed from the section ‘Pilot study secondary outcomes’: “The main secondary outcome for the pilot study will be death from any cause.” It has been replaced with: “Death from any cause will be recorded during the pilot study”.

Leon et al. (2011) suggest that when the research tools are standardized and the methods are not adjusted, subject data used in the pilot study could be pooled with those of the definitive study. We have added a sentence at the end of the section 'definitive study sample size’ discussing this point as the rationale for transitioning the same subjects from the pilot study to the definitive study.

The second concern is another multiplicity issue in the definitive study itself. You state that you will test two hypotheses. I agree that a 2x2 factorial design is a good choice. However, you still might need to adjust your alpha level due to the multiple testing issue. This might also have an impact on

the sample size calculation. A work-around solution would be to use a hierarchical testing procedure but I do not see or assume that you are planning to apply such a procedure. Please clarify.

Thank you. Statistical adjustments, particularly related to sample size calculation, may need to be considered in 2x2 factorial studies only if there is a possible interaction between the two interventions. According to Montgomery et al.*, logistic regression analysis is required for binary outcomes in a 2x2 factorial design, however no statistical adjustment or change in sample size is needed. We have added this reference as well as the reference by Moher et al. 2001 regarding the CONSORT guidelines to the section 'analysis of definitive study primary outcome'.

One additional minor comment: Please write "2x2" (or "2X2") consistently throughout the manuscript.

Thank you. We have edited the text to say "2x2" instead of "2X2".

VERSION 3 – REVIEW

REVIEWER	Lorenz Uhlmann Germany
REVIEW RETURNED	16-Aug-2019

GENERAL COMMENTS	Thank you for your reply and your amendments of the manuscript. I am satisfied with your replies and have no further comments. On a side note: I apologize for asking to clarify the multiplicity issue in your 2x2 factorial design once again in my last review. As you could see, I was (and am) not a big fan of non-adjustment for multiplicity. However, after going through several publications on this topic, I see that there is an ongoing discussion and that many authors rather suggest that no adjustment is necessary in this specific case. Therefore, I agree with your approach.