PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Strategies to Promote ResiliencY (SPRY): A Randomized Embedded Multifactorial Adaptative Platform (REMAP) Clinical Trial Protocol to Study Interventions to Improve Recovery after Surgery in High Risk Patients
AUTHORS	Reitz, Katherine; Seymour, Christopher W.; Vates, Jennifer; Quintana, Melanie; Viele, Kert; Detry, Michelle; Morowitz, Michael; Morris, Alison; Methe, Barbara; Kennedy, Jason; Zuckerbraun, Brian; girard, Timothy; Marroquin, Oscar; Esper, Stephen; Holder-Murray, Jennifer; Newman, Anne; Berry, Scott; Angus, Derek; Neal, Matthew

VERSION 1 – REVIEW

REVIEWER	Iain K Robertson University of Tasmania, Australia
REVIEW RETURNED	06-Mar-2020

GENERAL COMMENTS	**Overall comments:** This paper describes the protocol of a clinical trial currently being conducted (status: patients being recruited). The trial appears to be a substantial body of work being conducted by competent professionals in an appropriate setting. The trial protocol will have undergone extensive scrutiny and peer review. In this review, I am not intending to suggest any study redesign at this stage in its conduct: that is the prerogative of the trial team, and the institutional process that brought it into being. The target condition, loss of resilience, is relevant to all who survive to old age. Thus, the authors are addressing a common problem. They are using surgery as a methodologically convenient stressor to test the effects of treatments for promotion of resilience. The trial does not identify at this point those people who are at highest or soonest risk of suffering from loss of resilience. The trial uses two methodological processes that are relatively novel, which may have advantages over previous and current methodologies: 1) a computer algorithm process that guides and notifies the clinicians managing the patient within the routine electronic health record system that partially replaces a traditional

	RCT management centre; 2) statistical analysis using “Bayesian” principles and methods, as opposed to the more usual “probabilistic” principles and methods more familiar to practicing clinicians. The amount and complexity of the information required to be useful in terms of reproducing the RCT in another hospital system seems to have overwhelmed the recommended word count for a publication (or at least, the usual understanding of the appropriate word count). Some required information is missing, and some of information required is rendered in language that is both imprecise and opaque. This may have been the result of trying to fit everything in, but without the necessary editorial effort. I would suggest that the SPIRIT guidelines be used in a rigid manner (answer all the questions in the order in which they are numbered in the guidelines checklist), and place this in the separate additional materials document. The text of the main paper would then be a description of the innovations, along with a discussion of each component of the protocol (e.g. advantages, disadvantages, whatever else seems to the authors to be important). “Our paper follows the standard SPIRIT guidelines (see additional materials), and these are the particular features of the trial that are of interest and why they are interesting.” Of particular note is the use of multi-adaptive platform methodology (why it is used, advantages and disadvantages), how the trial methodology responds to the accumulated results following the interim analyses. Early in the paper, they might have an expanded discussion of resilience, how this trial is being used as an example of resilience in other contexts, how future selection of who should receive treatments for resilience promotion might be determined (for example). This paper would seem to me to be of much more interest than the paper in its current form. The section of the paper currently in the additional materials document (i.e. the statistical methodology and sample size estimation) needs to be improved. A justification for use of Bayesian methodology needs to be made beyond the current apparent explanation of “that’s the way we like to do this sort of analysis”. This needs to be written in language that is understandable by the clinicians who may wish to establish trials using this methodology of the same or different treatments. It is the responsibility of the statisticians to write their descriptions in the appropriate language understandable by competent health practitioners, and that competence does not include knowledge of arcane details of

	statistical methods or language. The SPIRIT guidelines protocol in the additional materials (as I am suggesting) should also include how the primary report of the results of the trial will be usable by those needing to perform meta-analysis that would include this trial. What are the authors' intentions about cooperation with meta-analyses. If the authors are able to comply with these suggestions, then my opinion is that this paper would be suitable for publication. **More specific comments:** **Aims:** The immediate aim of the study is to determine the optimum duration and dose of post-op Metformin in older patients following surgery of different types; the trial is part of a multi-modality trial (patients randomised to multiple possibly synergistic therapies), and the protocol is reporting a single component of the protocol of that trial. It is unclear whether the effects of the treatments are to be examined in an RCT manner simultaneously or sequentially. The broader aim is to evaluate and demonstrate a more elaborate, and possibly more effective, method of conduct of randomised controlled trials; using information technology to include in a more routine way the conduct of RCTs within normal hospital practice. Comment on Aims: 1. The medical condition being targeted by the therapies is age-related loss of reserve capacity within various systems of the body that would have previously allowed adequate response to a variety of stressors (e.g. illnesses or injuries). a. Surgery is being used in this trial as an example of such a stressor whose timing will be known precisely, and whose stressor strength can also be judged with reasonable precision. b. The target population implied by the inclusion criteria is older people, but it is likely that only a proportion of those initially eligible will have sufficient deficits in reserve capacity to justify the costs and potential adverse effects of the interventions. The trial has mechanisms for selecting sub-populations who have greater or lesser benefits. c. There may be interactions between the proposed multiple modalities being tested that go beyond simple additive or similar interactions.

2. The broader aim of demonstrating a possibly more effective way of conducting randomised controlled trials within highly developed health systems is laudable.
   a. This particular pathway of “embedding” the trial within routine health information systems within a hospital system requires a functioning electronic health record permitting machine-reading of the health record of individual patients, with connectivity between the different components of the system. This would seem the goal of any electronic health record system, although it may not have been achieved everywhere to date.
   b. The aim of achieving “cultural embedding” is also appropriate, although no plans for auditing the degree of achievement of this are described, and this will be relevant to the interpretation of the trial results.
   c. If the researchers are able to achieve their proposed aims and can demonstrate to others how this methodology can be performed, a more intense identification of the treatments that work and do not work can be undertaken due to increased RCT volumes facilitated by the computer technologies. All health systems are recipients of the benefits of past RCT results, and it is arguable that all should contribute within their capacity to future RCT evaluations of treatments. There is an ethical obligation to offer to patients only treatments that have proven benefits that have been measured and not guessed. This development of methodology has the potential to contribute to this, as the authors suggest.

Study Design:

Once commended, the core SPRY trial will use the same protocol processes in what looks like a continuous sequence of recruitments measuring the effects of a number of different treatment modalities. The treatment(s) being offered to each patient will depend on randomisation to whichever treatments are being studied at the time of the recruitment. The trial design involved recurrent analysis of the results, with closure of the trials for certain patients and certain treatments as the success or futility of those treatments are demonstrated by the interim analyses. Metformin appears to be the first treatment being commenced.

Comments:

1. The rules, if these exist, governing the commencement of the different treatments in the core trial are not described. The

	apparent (possibly default) assumption behind this decision might be that there are minimal and simple interactions between the different treatment modalities, and that the low-cost modalities (e.g. a metformin tablet) would be a reasonable place to start. 2. It would seem appropriate for the authors to describe how and why the different components (Domains) of the SPRY-Core trial will interact, since the treatments are likely to be used together for the same purpose once the Spry trials reach a successful conclusion. **Digital platform:** 1. The Multi-Adaptive Platform is a major innovation in this study. The laudable intention of this is to reduce the resource barriers to the conduct of randomised controlled trials. 2. However, one of the potential costs of this strategy might be to limit the degree of participant selection to only those participant characteristics that can be retrieved from the electronic health record (EHR). There may be participant characteristics that may require specific-measure screening that are not reliably available in the HER: if they are recorded in free-text data-fields, the absence of recording may be due to factor not present, or factor present by examination not done. It might be helpful if the extent of this limitation, is discussed in the paper. 3. The details of the digital platform used need to be described in more detail. The authors may wish to think about the nature of their platform, and how the generic specifications can be described as part of a protocol for this type of study. The SPIRIT guidelines do not specify what details should be described, presumably because few people knew of the possibility of such a trial innovation. 4. Also, to what extent does the digital platform represent proprietary software requiring purchase and/or building of the platform, and to what extent is the platform intended to be open source software. This question goes to the cost and difficulty of reproducing the results of the trial in different hospital systems, and different countries. **Trial participants:** Participants will be included in SPRY-Metformin who are over 60 years or have 3 or more Charlson Comorbidities (plus other detailed exclusions). Stratification will occur by surgical type (376 surgical

	strata), co-morbidities and age>75 (possibly others not stated). Comments: 1. The magnitude of the benefits of the SPRY interventions will accrue in absolute terms to those participants with the more severe levels of loss of “resilience” (possibly with loss of effect in those with lowest resilience). 2. The magnitude of the benefits of the SPRY interventions will accrue in absolute terms to those participants with the more severe levels of surgical “stress”. 3. This implies a great deal of heterogeneity in the benefits of the intervention in patients with different characteristics. 4. There is no explanation of the choice of the age criteria, or how sensitivity analysis might identify an appropriate age threshold for the decision to advise metformin use. Such sensitivity analysis might be considered for addition to the protocol (assuming that analysis has not commenced in this trial). 5. The protocol may allow identification of surgical groups which may receive more or less benefit, although the sample size simulation has assumed equal benefit in all surgical groups (an assumption that may not be correct). **Outcome measures**: Primary outcome: 90-day hospital free days (HFD) Secondary outcomes: 1) Index hospital course: Incidence and duration of post-op ICU admission; Length of stay; Discharge location; in-hospital mortality; 2) 30-days post-op: Surgical site infection<="" span="" style="font-family: Calibri; font-size: 11pt;">Surgical Site events; Organ failure free days; 3) 12-month post-op: Re-admission, Re-operation, DVT, PE, Mortality. Comments: 1. The primary outcome measure is being used as a proxy measure for resilience, which presumably (resilience) cannot be measured directly. a. The question is whether the measure of hospital free days is an appropriate model for testing the different treatments that may be used for enhancing resilience. [I will also elaborate this concern later in the comments concerning the sample size simulations.] b. The authors do not describe extensively the medical model of “resilience” and the effects of the treatments, although they do

	suggest the following may be occurring. It may be argued that many diseases involve diminution of “resilience” in a focal component of bodily function, and that the Charlson Co-morbidity Index is counting the numbers of such components that have crossed some threshold of loss of “resilience” relevant to that component. For many, the process of life approaching death will involve all the body systems progressing at different rates towards “resilience” being overwhelmed by “stressors”, and death occurring when an acute or subacute event occurs where the reserve capacity to repair the damage is no longer present. c. The authors claim that a long follow-up period is required to measure these effects, but that rests on the assumption that it is not possible to identify those at highest risk of early death, or delayed or inadequate recovery. However, it may be perfectly possible to select such a high-risk group, but that may require more elaborate screening of pre-operative surgical patients than is envisaged by the REMAP procedure chosen to address this particular clinical problem. The result of the trial design decision is that relatively large numbers of surgical patients will be recruited as participants of the trial who are at very-low risk of being affected by the lack of “resilience” being targeted by the SPRY treatments. d. The 90-day HFD measure will be affected by a number of effects, not all of which are affected in the same way by loss of “resilience”: i. The time from surgery to readiness for discharge; ii. Whether any post-operative rehabilitation services are provided, and whether those services are provided within the hospital environment or at some location outside the definition of in-hospital care; iii. Death within 90 days; iv. Whether readmission to hospital occurs as a result of failure to repair the surgical injury; v. Whether readmission to hospital occurs as a result of a new disease process made more likely by the lack of sufficient “resilience”; vi. Whether readmission to hospital occurs as a result of an entirely new disease process unrelated to the surgical “stressors” or lack of “resilience”. The effects of any treatments to reduce the effects of lack of “resilience” may affect those different processes to a different degree. With the large numbers of different

	surgeries in a broad range of patients, the heterogeneity in the components of the 90-day HFD measure have the possibility of creating significant confounding effects. The randomisation processes of randomised controlled trial are an attempt to create evenly matched populations in the different treatment groups. However, a heterogenous population that is intended to occur in this trial could result in treatment groups mismatched for confounders, even if there has been no defect in the randomisation process, simply due to the large number of subgroups that will be created. It might have been possible to quantify this effect by comparing the degree of confounder mismatching in each of the many simulated participant sample generated: however, this has not been reported (see later comments of sample size estimation). 2. The assignment of a value of -1 HFD for any death is intended as a way of getting around the problem of scoring of mortality, it assumes that any death occurred as a result of failure to recover from surgery (i.e. would include road trauma death). This arbitrary work-about will not allow different gradations of failure of resilience between different patients who die (e.g. after 30 days compared to after 75 days). A time-to-event analysis might get around this problem, but this would add further complexity (in terms of understanding what is going on). 3. It should be noted that the paper does not describe the method of determination of death (with date and cause of death), which would be expected in the protocol. In Tasmania, the State government regularly searches the Australian National Death Register and transfers this information to the local public hospital patient medical records (in order to avoid sending clinic appointments to people who have died): this may not include the causes of death in the hospital records. A similar system may operate in their local hospital system, but the authors need to describe how their local system works. It is a standard procedure in medical research that the full list of participants is sent to the national death register to determine whether they have died, and if so, the causes. Is this what is intended to happen? 4. Monitoring of participant status during their time in the clinical trial is to be conducted by examination of the electronic health record for encounters within the hospital system, and by phone interviews at 30- and 90-days following the date of surgery for encounters outside the trial hospital system. The paper does not

	describe how loss to follow-up will be handled (unless it is intended that the missing data will be filled by statistical multiple imputation), or how large differences in actual follow-up times occur compared with intended times will be handled. This lack of detail in the measurement of the primary outcome needs to be corrected, since presumably this information has already been specified in trial documentation. It should be noted that multiple imputation is usually conducted under the assumption that missing data has occurred at random, whereas missing follow-up data may be missing preferentially amongst those showing the greatest effects of lack of resilience. Is there pilot data available to demonstrate the size of this problem? **Sample size estimation:** To avoid the problem of false negative measurements in RCTs of clinical treatments, adequate numbers of participants must be observed to completion. This process involves estimation of the minimum numbers of participants require to detect a minimum relevant effect size (based on judgments that can be understood when a doctor and patient discuss consent for treatment) with predetermined levels of certainty (rate of Type 1 and 2 errors that are acceptable). There is no place for wishful thinking ("what benefits might be hoped for or expected") in therapeutic agent RCTs that may be acceptable in other scientific contexts (such as testing the concepts of scientific theories). Prospective power estimation is only appropriate where there is no control over the numbers of cases that can be observed. There are at least two ways of performing sample size estimation for the detection of treatment effects: 1. Calculations using the statistical tests (mathematical model) that best match the clinical problem (medical model). There are limited numbers of sample size calculation equations for different analysis methods, and there all assume that the population effects of treatment are uniform and known. At least, if the calculations are possible, they can be done quickly; 2. Simulation of large numbers of datasets, based on the best understanding of what the proposed participant population might look like, and how they are affected by treatment. This is a much more laborious process, but it does allow a much wider range of populations, effects and associations to be examined. There are two types of simulation:

	a. One using a hypothesised distribution of a study population, effects and associations, with the population distribution being derived from a known population (such as an audit of recent patients in the hospital who might be enrolled as participants in the proposed trial). This would be based primarily on the numerical model of the trial; b. One modelling individual patients, including what is known about the clinical problem and the hypothesised effects of treatment, again with the simulated datasets derived from audit results. This would be a combination of the numerical and medical models of the trial; c. The numerical model would be simpler to perform but would not necessarily give much information about tailoring treatment to specific patients, whilst the combined numerical/medical model would require much more data and knowledge to reliably make additional predictions for the benefit of specific patients. Comments: 1. The use of simulations in this trial is an appropriate response to the uncertainties that exist about the distribution of the participant population. The question arises as to what was done, and whether it has been adequately described. 2. Samples were taken from the audit sample of observed patients. 3. The numbers of participants required to achieve a pre-specified degree of certainty for a pre-specified effect size benefit were not calculated. 4. Instead, it appears that a pre-specified sample size was determined by some process, and then an analysis was undertaken to determine whether the power was adequate. Apparently, the power of the trial using the pre-specified numbers was adequate, but this result seems to have occurred by chance. If some other process occurred, the authors need to describe what has happened more precisely. 5. No clear justification was given about the choice of effect size being sought in the trial. It is likely to be difficult to give an appropriate clear justification, since the primary outcome is a proxy measure made up of component effects, some of which would be responsive to the treatment and some would not. Given that larger benefits will be confined to the biologically older participants, and that the stressors from different surgeries will vary in this respect, there is a level of uncertainty around the measurement of effect size that does not appear to have been

	acknowledged in the sample size determination process. 6. What I would expect to see from the description of the sample size simulation would be that a series of simulations would be conducted in the manner described that involve a series of different inputted effect sizes, and a series of different sample sizes. Each of the effect size and sample size condition combinations would then be repeated multiple times (possibly 100 to 10,000 times depending on the judgment of those conducting the analysis) with the planned statistical analysis applied to each of the generated samples. The proportion (the power) of simulations in each of the conditions that produced a positive result (the effect size measurement was greater than the pre-determined threshold, usually defined by the alpha value in "probabilistic" statistical analysis) would then be counted for each of the samples sizes. Those counts would then be analysed by non-linear regression (power as outcome and sample size as predictor variable) to determine the exact number of participants that would be required that correspond to the pre-determined level of power being sought for their RCT. The result would be an exact number (say 2,347 total participants). It would be very unlikely that the answer would be a significant number (2,000 or 2,500). The authors may have decided that they would round their numbers up to a significant number for reasons of tidiness, but if this is the case, they should have said so. a. My point is that it is impossible to be certain from the description whether the uncertainties that had been pre-defined determined the numbers of participants (which is the appropriate way of doing this sample size estimation), or that a choice of the numbers of participants drove the determination of the levels of uncertainties that were to be tolerated (which is not the appropriate way of going about this exercise). b. If it was the former that happened, a simple improvement in the description in the protocol paper is all that would be required. c. If it was the latter, either a re-analysis of the simulation data would be required if this could be done without repeating the simulations, or the whole sample size estimation may need to be repeated if the final trial result is to be taken as a definitive answer to the question of the utility of the trial treatment as applied to the trial population. d. Please note that I think that this is less of a problem than my comments above might appear to imply. RCTs are only

	going to be definitive as a single study when the effect size is much larger than the minimum effect size that is judged before the trial to sufficient to justify applying the treatment to patients. In most circumstances, multiple RCTs are provide such a definitive answer, and those multiple trials need to be combined by meta-analysis (see comments below concerning how this RCT is compatible with meta-analysis). So long as the authors are committed to being cautious in their interpretation of their final results, I do not regard this issue as being the major problem that is usually thought. 7. A further issue with the sample size estimation process is the impact of the use of Bayesian analysis versus the more usual "probabilistic" statistical analysis methods. No justification is made of the use of the Bayesian methods, although they may be justifiable. One of the possible justification might be a reduced sample size requirement. It is a legitimate question as to what the sample size estimation would be if non-Bayesian analysis were to be used (see below my comments on the statistical analysis proposed in the protocol). The authors need to comment on this issue, and to make a much more extensive justification of their statistical methods. **Statistical analysis:** The authors propose to use Bayesian ordered logistic regression as the primary analytic test of the proposed primary outcome measure. If it is accepted that the 90-day hospital-free day count is the appropriate best outcome measure to determine the benefits of the SPRY treatments, then the numerical properties of that measure suggest that a rank-ordered multivariate regression analysis method is appropriate. Ordered logistic regression is probably the best method available if relatively simple analysis is to be undertaken. There are two problems that arise from this coice: 1. Firstly, the result that is produced by ordered logistic regression is an odds ratio. This odds ratio does not necessarily mean the same thing as an odds ratio arising from the binary logistic regression analysis with which most readers will be familiar, and it is also not the same as a relative risk ratio. It would be the responsibility of the authors to make this clear in any final report of the trial (that will eventually be produced). 2. The authors tacitly acknowledge this problem by their discussion of mean numbers of HFD in the sample size estimation

	description, even though they have elsewhere acknowledged that the usual multiple linear regression analyses that estimate mean group values are not appropriate in this instance. This is an irreducible problem that I also run into in this type of circumstance. The solutions include: a. Conducting and reporting both linear regression and ordered logistic regression analyses, and then attempting to reconcile any differences in interpretation of the results that may arise (this is transparent but ambiguous, and readers may be left uncertain about what the authors think is the result of the trial); b. Conducting and reporting the ordered logistic regression analysis, and then attempting to explain to the readers, particularly the clinicians who have to explain to patients what the trial has shown. The doctor-patient encounter during which fully-informed consent is given for use of the treatment is the only really important target for any randomised controlled trial, and if the result of the trial cannot be understood reliably by both parties, then it is questionable whether a useful result has been achieved. Secondary customers of the trial, such as regulatory authorities, are only secondarily important in these decisions, and they have their own responsibilities to provide an answer to this dilemma; c. Conducting and reporting the linear regression analysis, is easier to undertake, and if ordered logistic regression analysis is used as a verification process and that the results of the two analytic methods give compatible interpretations, then the mean difference plus uncertainty would be more useful for the users of the trial results. If the results are not compatible, then the problem reasserts itself. As I said, this may be an irreducible problem. 3. Secondly, the authors propose to use a Bayesian version of the ordered logistic regression analysis. “Probabilistic” statistics operate on the assumption that the study being analysed arises anew as though nothing is previously known about the subject of the study. The cost of this is that larger numbers of observations are likely to be required to achieve a given level of certainty about the answer to the question. Bayesian statistics attempt to include something of what is previously known about the answer. The cost of this is that what is known may be affected by publication bias and less formal biases to which humans are prone.

	4. If the authors have good reasons for use of a particular statistical method, this should be explained, including a discussion of the assumptions underlying each of the alternatives. These discussions ought to be included in any description of the results of the completed trial, so it might be useful for the authors to rehearse the arguments as part of their description of the protocol. 5. The authors should also address the question of how the results of Bayesian analysis as reported in their intended primary paper would be included in a meta-analysis of different trials of the proposed treatments. 6. The choice of statistical methodology is the prerogative of those conducting the trial. Their responsibility includes describing their methods in sufficiently clear language (i.e. plain health professional English) that will allow those who will implement the results of the final trial (i.e. clinicians) to make a reliable judgement about the veracity of those results. It is not the responsibility of clinicians to know and understand the details of arcane statistical methods. It is the responsibility of the statistician to help the clinicians to achieve the necessary level of understanding. In my judgement, the current text does not include the required level of clarity. The clarity is required in the primary results paper, but it might be useful if those describing the statistical methods and the results rehearse their explanations in this protocol paper. **Missing components of the analysis:** 1. The analysis plan may answer the questions posed by the trial, and the other SPRY domains will answer the related questions of effect of synergistic treatments. 2. However, what is missing is a refinement to the choice of patients who would most benefit from the proposed treatments. This could involve covariate analysis to identify a risk-score that may in the future help to select patients for the current proposed treatments, or further treatments that may be relevant for patient subgroups. 3. This would be an exploratory analysis, generating hypotheses for future studies. These studies may be conducted within the REMAP philosophy, or it may necessitate separate studies using more traditional methods. There may be information required for SPRY-type treatments for selected subgroups that are not currently or easily available from the electronic health record. The REMAP philosophy should not foreclose on future

	developments. 4. The authors may wish to consider including this type of analysis or describing current intentions for explicitly. **Meta-analysis compatibility:** 1. It is unusual for a single RCT to produce a result that a completely definitive result is produced that renders all other clinical trials of the same treatment unnecessary. The effect size would have to be so large as to be unarguable, and that the trial team were not aware of this during the trial such that the trial continued despite superiority having been demonstrated if an interim analysis had been performed. Thus, in most circumstances, any one RCT cannot be regarded as definitive by the wider health community. The results need to be combined, and this happens by the process of meta-analysis. 2. When meta-analysis is undertaken, it is likely that at least some of the other trials will use "probabilistic" analysis, and I do not know how compatible the results of the two statistical methods would be. Would the reporting of the results include information that would be compatible, or is the reporting of the uncertainty of the Bayesian analysis incompatible with the usual uncertainty reported in the "probabilistic" analyses when the meta-analysis is done using the reported summary results? 3. A solution to this potential incompatibility problem would be making available an anonymised selected-variable dataset for use by qualified professionals for meta-analyses conducted according to pre-specified PRISMA (or equivalent) protocols. This would allow re-analysis of all the trials using identical statistical methods, either Bayesian or "probabilistic" or both. It would be helpful if this question is addressed, even though the SPIRIT guidelines do not explicitly require this information. 4. It should be noted that the PRISMA guidelines do not include the requirement for a sample size estimation. This would involve an analysis at the end of the meta-analysis about how many more participants would need to be recruited in future trials before further recruitment becomes unnecessary. The appropriate end-point for such a sample size estimation might be the number of participants after which the judgements about whether to use the treatment at the lower and upper 95% confidence intervals are the same, when those judgements are made by the doctor-patient dyad, and also the community decision-making processes make those judgement when comparing any alternative use of the health service resources (based maybe on

	cost-utility or similar analyses along with suitable political processes. Only when such sample size estimations have been done can decisions about whether to conduct further RCTs be made independent of the convenience and benefits of the usual initiators of RCTs. 5. The SPRIT and CONSORT guidelines do not require explicit statements about cooperation with processes of meta-analyses. This is an oversight, possibly due to disagreements about how this should be done. Authors could and should remedy this oversight in the case of their own RCT by stating how they intend to undertake such a cooperation. **Minor comments:** 1. First sentence in introduction presumably should read 55 million (not thousand) population over 65. 2. Page 7, line 40 in **Trial Design:** should "recurrently" (or similar) be used for "perpetually"? Perpetually suggests going on forever. 3. SPIRIT Item 11c Interventions: Adherence: "Patients compliance is queried at follow up patient encounters (Table 3, Figure 2)" on page 13, line 17, whilst the table quoted states "…..both patient safety and study drug compliance is monitored via phone interview (contact point 2)" on page 30, line 52. It is not made clear how tablet counting can be verified via a phone interview, but this problem is not addressed in the main text.

VERSION 1 – AUTHOR RESPONSE

Overall Reviewer Comments:

This paper describes the protocol of a clinical trial currently being conducted (status: patients being recruited). The trial appears to be a substantial body of work being conducted by competent professionals in an appropriate setting. The trial protocol will have undergone extensive scrutiny and peer review. In this review, I am not intending to suggest any study redesign at this stage in its conduct: that is the prerogative of the trial team, and the institutional process that brought it into being.

The target condition, loss of resilience, is relevant to all who survive to old age. Thus, the authors are addressing a common problem. They are using surgery as a methodologically convenient stressor to test the effects of treatments for promotion of resilience. The trial does not identify at this point those people who are at highest or soonest risk of suffering from loss of resilience.

The trial uses two methodological processes that are relatively novel, which may have advantages over previous and current methodologies: 1) a computer algorithm process that guides and notifies the clinicians managing the patient within the routine electronic health record system that partially replaces a traditional RCT management centre; 2) statistical analysis using "Bayesian" principles and methods, as opposed to the more usual "probabilistic" principles and methods more familiar to practicing clinicians.

The amount and complexity of the information required to be useful in terms of reproducing the RCT in another hospital system seems to have overwhelmed the recommended word count for a publication (or at least, the usual understanding of the appropriate word count). Some required information is missing, and some of information required is rendered in language that is both imprecise and opaque. This may have been the result of trying to fit everything in, but without the necessary editorial effort.

I would suggest that the SPIRIT guidelines be used in a rigid manner (answer all the questions in the order in which they are numbered in the guidelines checklist), and place this in the separate additional materials document.

The text of the main paper would then be a description of the innovations, along with a discussion of each component of the protocol (e.g. advantages, disadvantages, whatever else seems to the authors to be important). "Our paper follows the standard SPIRIT guidelines (see additional materials), and these are the particular features of the trial that are of interest and why they are interesting." Of particular note is the use of multi-adaptive platform methodology (why it is used, advantages and disadvantages), how the trial methodology responds to the accumulated results following the interim analyses. Early in the paper, they might have an expanded discussion of resilience, how this trial is being used as an example of resilience in other contexts, how future selection of who should receive treatments for resilience promotion might be determined (for example). This paper would seem to me to be of much more interest than the paper in its current form.

The section of the paper currently in the additional materials document (i.e. the statistical methodology and sample size estimation) needs to be improved. A justification for use of Bayesian methodology needs to be made beyond the current apparent explanation of "that's the way we like to do this sort of analysis". This needs to be written in language that is understandable by the clinicians who may wish to establish trials using this methodology of the same or different treatments. It is the responsibility of the statisticians to write their descriptions in the appropriate language understandable by competent health practitioners, and that competence does not include knowledge of arcane details of statistical methods or language.

The SPIRIT guidelines protocol in the additional materials (as I am suggesting) should also include how the primary report of the results of the trial will be usable by those needing to perform metaanalysis that would include this trial. What are the authors' intentions about cooperation with metaanalyses.

If the authors are able to comply with these suggestions, then my opinion is that this paper would be suitable for publication.

Overall Reviewer Comments Section - Response to Reviewer:

The reviewer has succinctly suggested a variety of adjustments which, we believe, have increased the readability and interest of our manuscript. We have generated a supplemental document (appendix 1) summarizing the trial information as it corresponds to the SPIRIT Guidelines. As the reviewer suggested, this has provided additional space within the manuscript for discussion surrounding resiliency as it pertains to surgical patients and justification as well as clarification of the primary Bayesian analysis methods. specific comments/questions in this section below.

Overall Comments Section - Specific Reviewer Comments and Associated Responses:

"I would suggest that the SPIRIT guidelines be used in a rigid manner (answer all the questions in the order in which they are numbered in the guidelines checklist), and place this in the separate additional materials document."

Response to Reviewer: Please see the SPIRIT Guidelines (Appendix 1) and associated changes to the manuscript, throughout the responses below.

Updated Manuscript:

Methods and Analysis

Our protocol follows the SPIRIT guidelines which are individually addressed in Appendix 1. The below content focuses on novel aspects of the SPRY Core protocol and associated SPRY-Metformin Domain-specific Appendix.

"The trial does not identify at this point those people who are at highest or soonest risk of suffering from loss of resilience."

"Early in the paper, they might have an expanded discussion of resilience, how this trial is being used as an example of resilience in other contexts, how future selection of who should receive treatments for resilience promotion might be determined (for example)."

Response to Reviewer:

We have expanded the discussion of loss of resiliency, or frailty, throughout the manuscript. Specifically, we have attempted to clarify for the reader why age is an important factor in the development of frailty, but the accumulation of deficits can be independent of patient age. Building off of the reviewer's comments, this allows the reader to better understand our inclusion and exclusion criteria as well as subgroup analysis.

Additionally, we have added a description of exploratory analyses to investigate the heterogeneity of treatment effects across key patient subgroups. See response to the reviewer comments under the **Trial Participants Section** below.

Updated Manuscript:

Background

By 2020, over 55 million Americans will be greater than 65 years of age [1]. The lifelong accumulation of stressors progressively leads to chronic disease and disability compromising homeostatic reserve. The complex interplay of cumulative medical, social, and functional generating these deficits, defined as frailty, are associated with but independent from age and leave individuals vulnerable to a physiologic insult further reducing resiliency [2,3].

Elderly patients, at risk of frailty, undergo over one third of all surgical interventions and have an increased rate of postoperative morbidity and mortality at all levels of surgical stress [3–7].

Updated Citations:

Joseph B, Pandit V, Zangbar B, Kulvatunyou N, Hashmi A, Green DJ, et al. Superiority of frailty over age in predicting outcomes among geriatric trauma patients: A prospective analysis. JAMA Surg. 2014;149(8):766–72.

Shinall MC, Arya S, Youk A, Varley P, Shah R, Massarweh NN, et al. Association of Preoperative Patient Frailty and Operative Stress with Postoperative Mortality. JAMA Surg. 2019;15213(1):1–9.

Shinall MC, Youk A, Massarweh NN, *et al.* Association of Preoperative Frailty and Operative Stress With Mortality After Elective vs Emergency Surgery. *JAMA Netw Open* 2020;**3**:10–3. doi:10.1001/jamanetworkopen.2020.10358

"A justification for use of Bayesian methodology needs to be made beyond the current apparent explanation of "that's the way we like to do this sort of analysis". This needs to be written in language that is understandable by the clinicians who may wish to establish trials using this methodology of the same or different treatments."

Response to Reviewer:

We hope to have clarified the language throughout the manuscript and associated appendixes to increase the readability of the manuscript, increase the transparency of our analysis plan, and to elaborate our justification for use of Bayesian methodology. For SPRY-Metformin, the Bayesian analysis was chosen over a frequentist approach to allow us to minimize the number of patients required to evaluate the effectiveness of multiple doses and durations of metformin. The Bayesian statistical analysis plan allows for borrowing of information on the treatment effect by placing a hierarchical prior distribution on this treatment across all doses, all durations, and the interaction between them. Further, to aid clinicians interested in expanding their own knowledge of the Bayesian methodology, we have also added the below references to the manuscript. We elaborated this discussion under the Statistical Analysis Section below.

Updated Manuscript:

Statistical Analysis

The primary analysis plan for SPRY-Metformin includes a Bayesian ordinal logistic regression analysis of 90-day HFD to allow for borrowing of information on the treatment effect across different doses and durations of Metformin to maximally inform the research questions while

minimizing the required patient sample size [8,9]. Complete documentation of the statistical analysis plan is including in the Statistical Analysis Appendix.

Statistical Analysis Appendix (Section 2.0)
In this setting, the Bayesian analysis makes use of non-informative prior distributions with regards to HFD distributions for each surgical strata and in this regard is very similar in nature to a frequentist ordinal logistic analysis. However, we chose to use a Bayesian analysis over a frequentist approach to allow for borrowing of information on the treatment effect across different doses and durations. This borrowing is done in the Bayesian setting by placing a hierarchical prior distribution on the treatment effects across all doses, all durations and the interactions between them.

Updated Citations:

Gelman A, Shalizi CR. Philosophy and the practice of Bayesian statistics. *Br J Math Stat Psychol* 2013;**66**:8–38. doi:10.1111/j.2044-8317.2011.02037.x
Agresti A. *Categorical Data Analysis*. 3rd ed. Wiley 2012.

Aims:

The immediate aim of the study is to determine the optimum duration and dose of post-op Metformin in older patients following surgery of different types; the trial is part of a multi-modality trial (patients randomised to multiple possibly synergistic therapies), and the protocol is reporting a single component of the protocol of that trial. It is unclear whether the effects of the treatments are to be examined in an RCT manner simultaneously or sequentially.

The broader aim is to evaluate and demonstrate a more elaborate, and possibly more effective, method of conduct of randomised controlled trials; using information technology to include in a more routine way the conduct of RCTs within normal hospital practice.

Comment on Aims:

1. **The medical condition being targeted by the therapies is age-related loss of reserve capacity within various systems of the body that would have previously allowed adequate response to a variety of stressors (e.g. illnesses or injuries).**
   a. **Surgery is being used in this trial as an example of such a stressor whose timing will be known precisely, and whose stressor strength can also be judged with reasonable precision.**
   b. **The target population implied by the inclusion criteria is older people, but it is likely that only a proportion of those initially eligible will have sufficient deficits in reserve capacity to justify the costs and potential adverse effects of the interventions. The trial has mechanisms for selecting sub-populations who have greater or lesser benefits.**
   c. **There may be interactions between the proposed multiple modalities being tested that go beyond simple additive or similar interactions.**
2. **The broader aim of demonstrating a possibly more effective way of conducting randomised controlled trials within highly developed health systems is laudable.**
   a. **This particular pathway of "embedding" the trial within routine health information systems within a hospital system requires a functioning electronic health record permitting machine reading of the health record of individual patients, with connectivity between the different components of the system. This would seem the goal of any electronic health record system, although it may not have been achieved everywhere to date.**
   b. **The aim of achieving "cultural embedding" is also appropriate, although no plans for auditing the degree of achievement of this are described, and this will be relevant to the interpretation of the trial results.**
   c. **If the researchers are able to achieve their proposed aims and can demonstrate to others how this methodology can be performed, a more intense identification of the treatments that work and do not work can be undertaken due to increased RCT volumes facilitated by the computer technologies. All health systems are recipients of the benefits of past RCT results, and it is arguable that all should contribute within**

their capacity to future RCT evaluations of treatments. There is an ethical obligation to offer to patients only treatments that have proven benefits that have been measured and not guessed. This development of methodology has the potential to contribute to this, as the authors suggest.

Aims Section - Response to Reviewer:

We appreciate the thoughtful comments and through understanding the reviewer has of both the manuscript and the trial. Please see our responses to specific comments/questions in this section below.

Aims Section – Specific Reviewer Comments and Associated Responses:

"It is unclear whether the effects of the treatments are to be examined in an RCT manner simultaneously or sequentially."

Response to Reviewer:

The flexibility of the REMAP platform is essential to our Core and Domain-Specific Protocol and therefore, we are very interested in clarifying this point and have made the following updates to the manuscript.

Updated Manuscript:

Background

Therefore, we report the first of many trial protocols evaluating perioperative therapies, both concurrently and sequentially, on this adaptive platform, SPRY-Metformin."

SPRY Core Protocol

In the REMAP design, patients can be randomized to one of many treatments within one of many simultaneously deployed domains resulting in multiple possible experimental treatment combinations. The Core Protocol allows for aggregation of the treatment response across different the simultaneously investigated domains and the multifactorial evaluation of synergistic or antagonistic combinations within each of the strata.

"The target population implied by the inclusion criteria is older people, but it is likely that only a proportion of those initially eligible will have sufficient deficits in reserve capacity to justify the costs and potential adverse effects of the interventions. The trial has mechanisms for selecting sub-populations who have greater or lesser benefits."

Response to Reviewer:

In response to other key comments from the reviewer, we have highlighted the importance of both age and loss of resiliency in the patients included within this study. We have extended our discussion with regards to exploring heterogeneity of treatment effects in key patient subgroups. This is discussed in full in the reviewers' comments within the **Trial Participants Section** below.

"There may be interactions between the proposed multiple modalities being tested that go beyond simple additive or similar interactions."

Response to Reviewer:

Please see response to the **Study Design Section** reviewer comments below with regards to the generalizability of the Bayesian analysis method and response adaptive randomization to multiple treatments/domains.

"The aim of achieving "cultural embedding" is also appropriate, although no plans for auditing the degree of achievement of this are described, and this will be relevant to the interpretation of the trial results."

Response to Reviewer:

Thank you for highlighting the importance of clinical embedding in our protocol. We also agree that the degree of cultural embedding may be important for result interpretation. In our final data analysis and presentation, we will include two informative data points regarding the enrolling clinics. First, in our final consort diagram, in our published results, we will include the number of patients enrolled per number screened for each clinic. Second, we will include descriptions the amount, if any, of clinical research staff required in each clinic for patient enrolment. For example,

some clinics have requested a member of the clinical research team be on site to help streamline patient enrolment while others function entirely autonomously.

Updated Manuscript:
SPIRIT Guidelines Appendix (Guideline, 9)
The study protocol is embedded within the workflow of both the electronic health record and the clinical care of patients. The final manuscript will include the list of enrolling clinics, the number of patients screened (both digitally and in-person) and enrolled per clinic, and the amount of clinical research staff support requested and required per clinic.

Study Design:

Once commended, the core SPRY trial will use the same protocol processes in what looks like a continuous sequence of recruitments measuring the effects of a number of different treatment modalities. The treatment(s) being offered to each patient will depend on randomisation to whichever treatments are being studied at the time of the recruitment. The trial design involved recurrent analysis of the results, with closure of the trials for certain patients and certain treatments as the success or futility of those treatments are demonstrated by the interim analyses. Metformin appears to be the first treatment being commenced.

Comments:

1. **The rules, if these exist, governing the commencement of the different treatments in the core trial are not described. The apparent (possibly default) assumption behind this decision might be that there are minimal and simple interactions between the different treatment modalities, and that the low-cost modalities (e.g. a metformin tablet) would be a reasonable place to start.**
2. **It would seem appropriate for the authors to describe how and why the different components (Domains) of the SPRY-Core trial will interact, since the treatments are likely to be used together for the same purpose once the Spry trials reach a successful conclusion.**

Study Design Section - Response to Reviewer:

Thank you for highlighting these important aspects of the REMAP study. Please see our responses to specific comments/questions in this section below.

Study Design Section - Specific Reviewer Comments and Associated Responses:

"The rules, if these exist, governing the commencement of the different treatments in the core trial are not described. The apparent (possibly default) assumption behind this decision might be that there are minimal and simple interactions between the different treatment modalities, and that the low-cost modalities (e.g. a metformin tablet) would be a reasonable place to start."

Response to Reviewer: As the reviewer suggests, metformin was chosen to be the first domain tested on the SPRY Core protocol for many reasons. First, the authors of this manuscript and our institutional collaborators have significant interest in understanding and harnessing the anti-inflammatory properties of metformin. These interests are both clinical, in relation to frailty and surgical outcomes, and mechanistic leading to the generation of the clinical trial protocol and associated biorepository. Second, metformin is safe, inexpensive, and has a well-established side effect profile and therefore minimizing in-trial, drug related safety uncertainties.

Potential new emerging domains or therapies to be tested on the SPRY core protocol are reviewed by the Trial Steering Committee and can be added to the SPRY core protocol as a Domain-specific Appendix similar to the example of the metformin domain. As the reviewer has highlighted, concurrent domains need to interact both safely and in a way that can be integrated into the statistical analysis plan.

Updated Manuscript:

Trial design
Specifically, SPRY will recurrently assess multiple, Trial Steering Committee (TSC) approved treatments in multiple surgical and disease subtypes using response adaptive randomization and a comprehensive statistical analysis plan to create a self-learning health system.

SPRY Core Protocol
We provided details herein on the first SPRY Core Protocol Domain (SPRY-Metformin). Other, New domains will be added to the Core Protocol as emerging therapies become available. The TSC will consider the scientific validity of each domain, safety of concurrent therapies, and current enrollment rates when deciding to introduce a new domain concurrently or following existing domains. A new domain is introduced as a Domain-specific Appendix to the SPRY Core Protocol. This Domain-specificAppendix will be generated outlining potential interactions between multiple domains within the primary statistical analysis of efficacy, if deemed clinically appropriate. If multiple domains have been introduced, response adaptive randomization will be based on best performing combinations of therapies within the multiple domains and incorporate potential interactions.

"It would seem appropriate for the authors to describe how and why the different components (Domains) of the SPRY-Core trial will interact, since the treatments are likely to be used together for the same purpose once the Spry trials reach a successful conclusion."

Response to Reviewer:
As new therapies emerge and are introduced into the SPRY Core Protocol, statistical interactions (multiplicative and additive) will be included within the primary Bayesian analysis if it is deemed to be clinically appropriate by the TSC. Additionally, response adaptive randomization will be used across all possible combinations including, treatment randomization and patient randomization to the most effective therapy or combination of therapies. A comment on the generalizability of the proposed Bayesian primary analysis to multiple treatments as well as a description of how response adaptive randomization will take place if there are multiple therapies / domains has been included in the manuscript, as seen below.

Updated Manuscript:
Trial Design
A new domain is introduced as a Domain-specific Appendix to the SPRY Core Protocol. This Domain-specific Appendix will be generated outlining potential interactions between multiple domains within the primary statistical analysis of efficacy, if deemed clinically appropriate. If multiple domains have been introduced, response adaptive randomization will be based on best performing combinations of therapies within the multiple domains and incorporate potential interactions.

Digital platform:

1. **The Multi-Adaptive Platform is a major innovation in this study. The laudable intention of this is to reduce the resource barriers to the conduct of randomised controlled trials.**
2. **However, one of the potential costs of this strategy might be to limit the degree of participant selection to only those participant characteristics that can be retrieved from the electronic health record (EHR). There may be participant characteristics that may require specific-measure screening that are not reliably available in the HER: if they are recorded in free-text data-fields, the absence of recording may be due to factor not present, or factor present by examination not done. It might be helpful if the extent of this limitation, is discussed in the paper.**
3. **The details of the digital platform used need to be described in more detail. The authors may wish to think about the nature of their platform, and how the generic specifications can be described as part of a protocol for this type of study. The SPIRIT guidelines do not specify what details should be described, presumably because few people knew of the possibility of such a trial innovation.**
4. **Also, to what extent does the digital platform represent proprietary software requiring purchase and/or building of the platform, and to what extent is the platform intended to be open source software. This question goes to the cost and difficulty of reproducing the results of the trial in different hospital systems, and different countries.**

Digital Platform Section - Response to Reviewer:

We appreciate the reviewer's interest in the digital embedding and the SPRY-Application. Please see our responses to specific comments/questions in this section below.

Digital Platform Section - Specific Reviewer Comments and Associated Responses:

"…to limit the degree of participant selection to only those participant characteristics that can be retrieved from the electronic health record (EHR). There may be participant characteristics that may require specific-measure screening that are not reliably available in the HER: if they are recorded in free-text data-fields, the absence of recording may be due to factor not present, or factor present by examination not done. It might be helpful if the extent of this limitation, is discussed in the paper."

Response to Reviewer: The reviewer adequately highlights, harnessing the potential increased efficiency of automated patient identification using EHR data has inherent limitation. Although many potential strategies exist for the screening of patients in an outpatient clinic, a commonly used protocol may include research staff reviewing the and/all pertinent health information prior to the clinical encounter. This technique, sometimes colloquially referred to as "chart scrubbing", allows systematic identification of patients who may meet inclusion and do not meet exclusion criteria for in person screening. As the reviewer has suggested, the SPRY-Application can only exclude patients from the list of potential SPRY-Metformin candidates if data are appropriately documented in automated fields. Therefore, SPRY-Application based patient identification has equal sensitivity but less specificity than the "chart scrubing." We have clarified this in the manuscript as described below.

Further, as the SPRY-Application guides the review of patients' inclusion and exclusion criteria any and all discrepancies identified through patient interview generate a prompt for the provider to update/correct the EHR.

Updated Manuscript:

Digital Embedding

Potential trial participant identification begins with the SPRY-Application screening. The SPRY-Application reviews SPRY specific, EHR data for each patient with a scheduled appointment at enrolling preoperative SPRY-Metformin clinics (**Figure 3**).

The SPRY-Application guides the clinician through the stepwise informed consent process. Then, pertinent clinical biorepository EHR data auto-populates screening information within the SPRY-Application for review and conformation with the patient (**Figure 4**).

"The details of the digital platform used need to be described in more detail."

"Also, to what extent does the digital platform represent proprietary software requiring purchase and/or building of the platform, and to what extent is the platform intended to be open source software."

Response to Reviewer: The SPRY-Application interacts with EHR data systematically abstracted from raw, structured data fields within our commercially available EHR. This system parallels data abstraction used for widely published retrospective clinical research studies both nationally and internationally. As previously published, and now cited in this manuscript, and available through the EHR support systems structured patient data are formatted into tables accessible by SQL Server. However, there is, of course, institutional specific variation in customizable EHR formatting and use. Therefore, the data abstracted into the data repository for review by the SPRY-Application is, at least to some extent, institutionally specific. At this time, we do not plan to provide open source code, but encourage collaboration from interested investigators who contact our group.

In this manuscript, we are hoping to document our protocol and highlight the novelty appreciated by the reviewer. We have added language to expand the technical aspects of the digital platform for transparency without overwhelming the general clinician reader.

Updated Manuscript:

Digital Embedding

At UPMC, a two-factor authentication system safeguards all private patient information accessed through a single Citrix Workspace (Fort Lauderdale, FL) in accordance with Health Insurance Portability and Accountability Act. Like all protected data and programs within the healthcare system, the SPRY-Application resides behind this institutional firewall. Here, the SPRY-Application is distinct from, but communicates with EHR data. The SPRY-Application accesses the clinical research data repository within UPMC Clinical Analytics and is managed by Biostatistical and Data Management Core in the Department of Critical Care Medicine at UPMC. The data repository abstracts structured, raw data from the inpatient (CERNER Co., Kansas City, MO) and outpatient (Epic Systems Co., Madison, WI) EHR in real time via SQL Server and generates accessible data tables. The data extraction process parallels the methodology used traditionally for retrospective EHR data collection and research [10–12]; however, these data are updated in real time.

Updated Citations:

Milinovich A, Kattan MW. Extracting and utilizing electronic health data from Epic for research. *Ann Transl Med* 2018;**6**:42–42. doi:10.21037/atm.2018.01.13

Reitz KM, Marroquin OC, Zenati MS, *et al.* Association between metformin exposure and postoperative outcomes in diabetic adults. *JAMA Surg* 2020.

Singer M, Deutschman CS, Seymour CW, *et al.* The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA* 2016;**315**:801–10. doi:10.1001/jama.2016.0287

Trial participants:

Participants will be included in SPRY-Metformin who are over 60 years or have 3 or more Charlson Comorbidities (plus other detailed exclusions). Stratification will occur by surgical type (376 surgical strata), co-morbidities and age>75 (possibly others not stated).
Comments:

1. **The magnitude of the benefits of the SPRY interventions will accrue in absolute terms to those participants with the more severe levels of loss of "resilience" (possibly with loss of effect in those with lowest resilience).**
2. **The magnitude of the benefits of the SPRY interventions will accrue in absolute terms to those participants with the more severe levels of surgical "stress".**
3. **This implies a great deal of heterogeneity in the benefits of the intervention in patients with different characteristics.**
4. **There is no explanation of the choice of the age criteria, or how sensitivity analysis might identify an appropriate age threshold for the decision to advise metformin use. Such sensitivity analysis might be considered for addition to the protocol (assuming that analysis has not commenced in this trial).**
5. **The protocol may allow identification of surgical groups which may receive more or less benefit, although the sample size simulation has assumed equal benefit in all surgical groups (an assumption that may not be correct).**

Trial Participants Section - Response to Reviewer:

The reviewer astutely highlights the complex interplay between patient age, frailty, surgical stress, and the potential beneficial effects of perioperative optimization. Increasing frailty, or decrease physiologic reserve, confers a progressive increase in postoperative morbidity and mortality. Of course, surgical interventions generating high levels of operative stress have a higher postoperative complication and mortality rate. However, as demonstrated by Shinall et. al. (JAMA Surgery, 2019 and JAMA Network Open, 2020), 180-day postoperative morbidity are >10% higher for patients lacking physiologic reserve, across all levels operative stress, as measured by the Operative Stress Score, or surgical speciality.

As discussed below, the patient population to be enrolled in SPRY-Metformin are at risk of having decreased physiologic reserve and therefore are at an increased risk of postoperative morbidity and mortality. However, we do expect the degree of preoperative frailty to affect both expected HFD and the associated treatment effect of the study drug. As with all clinical trials, we must balance the expected heterogeneity of the treatment effect with the external validity or applicability of the trial results. For example, we could have chosen to narrow our enrolment

criteria to a specific and low level of physiologic reserve, for which the average treatment effect might be greater and more uniform.

As such, we do not assume a common treatment effect across the broadly enrolled patient population on absolute terms, there are expected differences in HFD. Instead, the Bayesian analysis model allows for heterogeneity in the overall distribution of the outcome measure as a function of surgical strata (i.e. speciality) and assumes a common odds ratio treatment effect across surgical strata (resulting in different absolute treatment differences depending on underlying risk). Exploratory analyses will be performed to assess the heterogeneity of treatment effects across key patient subgroups including, age, frailty, operative stress, and surgical strata.

Please see our responses to specific comments/questions in this section below.

Trial Participants Section - Specific Reviewer Comments and Associated Responses:

"There is no explanation of the choice of the age criteria, or how sensitivity analysis might identify an appropriate age threshold for the decision to advise metformin use. Such sensitivity analysis might be considered for addition to the protocol (assuming that analysis has not commenced in this trial)."

Response to Reviewer: As the reviewer has identified, the goal of SPRY is to examine the potential effectiveness of perioperative therapies on postoperative outcomes. SPRY-Metformin investigates the effect of metformin on aged, frail patients. The global syndrome of frailty is driven by the complex accumulation of medical, functional, and social deficits. The accumulation of these deficits is both associated with, but independent from the age of a patient.

The age of 60 years old was chosen as inclusion criteria for two reasons. First, our team of surgeons, intensivist, and geriatricians felt this to have good face validity as a risk factor a surgical intervention resulting in accelerated aging. Second, published (Joseph et al, JAMA Surgery, 2014) and unpublished data abstracted from the National Surgical Quality Improvement Program (NSQIP) demonstrated that the mean Risk Analysis Index (RAI), a validated and widely published measure of frailty in surgical patients, is >29 (therefore categorized as frail) for patients >60 years of age. For those patients less than 60 years of age, comorbid conditions, including those captured by the Charlson Co-morbidity Index allow for a systematic EHR based identification of risk factors predicting frailty.

Updated Manuscript:

Study Population

At the first level, patients exposed to the stress of an elective surgical intervention are identified. At the second level, participants are evaluated against the inclusion and exclusion criteria of the SPRY-Metformin domain identifying patients who, i) can be safely exposed to metformin and ii) are at risk of decreased physiologic reserve (i.e., older age and/or medical comorbidity) conferring postoperative morbidity and mortality at all levels of surgical stress (**Table 1**) [3,10]. Patients randomized in SPRY-Metformin can also participate in either or both substudies (microbiome or motor) as well as additional future domains on the SPRY Core Protocol.

Updated Citations:

Joseph B, Pandit V, Zangbar B, Kulvatunyou N, Hashmi A, Green DJ, et al. Superiority of frailty over age in predicting outcomes among geriatric trauma patients: A prospective analysis. JAMA Surg. 2014;149(8):766–72.

Shinall MC, Arya S, Youk A, Varley P, Shah R, Massarweh NN, et al. Association of Preoperative Patient Frailty and Operative Stress with Postoperative Mortality. JAMA Surg. 2019;15213(1):1–9.

"The protocol may allow identification of surgical groups which may receive more or less benefit, although the sample size simulation has assumed equal benefit in all surgical groups (an assumption that may not be correct)."

Response to Reviewer: We have included a broad range of surgical interventions in order to maximize the external validity of effectiveness of metformin pharmacologic optimization in the perioperative period. As discussed above, for all major surgical interventions, the stress of an operation confers an increased risk of postoperative mortality and morbidity in patients with reduced physiologic reserve. Our current primary analysis allows for heterogeneity in the

underlying distribution of the outcome as a function of surgical subtypes and assumes a common odds ratio treatment effect. This common odds ratio treatment effect allows for different absolute treatment benefits based on the underlying risk. Additionally, we have included a description of exploratory analyses that will performed to assess the heterogeneity of treatment effects across key patient subgroups.

Within the Statistical Analysis Appendix, we have included a clearer description of assumption of the treatment effect is provided in the description of the statistical modelling and the heterogeneity of treatment effects across key patient subgroups.

Updated Manuscript:
Background
Elderly patients, at risk of frailty, undergo over one third of all surgical interventions and have an increased rate of postoperative morbidity and mortality independent of the magnitude of the surgical stressor [3–7].

Analysis plan
Exploratory analyses will investigate the heterogeneity of treatment effects across key patient subgroups including, patient age and frailty as well as operative stress and surgical strata.

Statistical Analysis Appendix (Section 2.0)
The ordinal logistic regression model accounts for underlying differences in the expected 90-day HFD distribution depending on surgical procedure or strata of the patient but assumes a common odds ratio treatment effect across the surgical strata. The odds ratio shift within an ordinal logistic regression model can be thought of similarly to an odds ratio within a logistic regression analysis of a dichotomous endpoint. Within ordinal logistic regression, we are simply performing multiple logistic regression analyses (one for each possible dichotomization of the data) and providing a weighted average of the odds ratios across these different dichotomizations. The assumption of a common odds ratio treatment effect across the different surgical subtypes translates into different absolute differences in the mean hospital free days within each surgical strata. For a common odds ratio across surgical strata, the larger the expected HFD within the strata the smaller the absolute mean difference in HFD between treatment and control.

Statistical Analysis Appendix (Section 2.5)
The heterogeneity of treatment effects across different key patient subgroups will be explored by allowing the common odds ratio per dose and duration,), to be subgroup dependent. Subgroups of interest include: Surgical specialty (see Table 4.1), operative stress level as defined by the Operative Stress Score,[7]surgical subtype, age (19-44, 45-64, 65-84, >84 years of age), and frailty based upon the prospectively calculated Revised Analysis Index.

Outcome measures:
Primary outcome: 90-day hospital free days (HFD)
Secondary outcomes: 1) Index hospital course: Incidence and duration of post-op ICU admission; Length of stay; Discharge location; in-hospital mortality; 2) 30-days post-op: Surgical site infection; Surgical Site events; Organ failure free days; 3) 12-month post-op: Re-admission, Re-operation, DVT, PE, Mortality.

Comments:
1. The primary outcome measure is being used as a proxy measure for resilience, which presumably (resilience) cannot be measured directly.
   a. **The question is whether the measure of hospital free days is an appropriate model for testing the different treatments that may be used for enhancing resilience. [I will also elaborate this concern later in the comments concerning the sample size simulations.]**
   b. **The authors do not describe extensively the medical model of "resilience" and the effects of the treatments, although they do suggest the following may be occurring. It**

may be argued that many diseases involve diminution of "resilience" in a focal component of bodily function, and that the Charlson Co-morbidity Index is counting the numbers of such components that have crossed some threshold of loss of "resilience" relevant to that component. For many, the process of life approaching death will involve all the body systems progressing at different rates towards "resilience" being overwhelmed by "stressors", and death occurring when an acute or subacute event occurs where the reserve capacity to repair the damage is no longer present.

c. **The authors claim that a long follow-up period is required to measure these effects, but that rests on the assumption that it is not possible to identify those at highest risk of early death, or delayed or inadequate recovery. However, it may be perfectly possible to select such a high-risk group, but that may require more elaborate screening of pre-operative surgical patients than is envisaged by the REMAP procedure chosen to address this particular clinical problem. The result of the trial design decision is that relatively large numbers of surgical patients will be recruited as participants of the trial who are at very-low risk of being affected by the lack of "resilience" being targeted by the SPRY treatments.**

d. **The 90-day HFD measure will be affected by a number of effects, not all of which are affected in the same way by loss of "resilience":**

   i. **The time from surgery to readiness for discharge;**
   ii. **Whether any post-operative rehabilitation services are provided, and whether those services are provided within the hospital environment or at some location outside the definition of in-hospital care;**
   iii. **Death within 90 days;**
   iv. **Whether readmission to hospital occurs as a result of failure to repair the surgical injury;**
   v. **Whether readmission to hospital occurs as a result of a new disease process made more likely by the lack of sufficient "resilience";**
   vi. **Whether readmission to hospital occurs as a result of an entirely new disease process unrelated to the surgical "stressors" or lack of "resilience".**

The effects of any treatments to reduce the effects of lack of "resilience" may affect those different processes to a different degree. With the large numbers of different surgeries in a broad range of patients, the heterogeneity in the components of the 90-day HFD measure have the possibility of creating significant confounding effects. The randomisation processes of randomised controlled trial are an attempt to create evenly matched populations in the different treatment groups. However, a heterogenous population that is intended to occur in this trial could result in treatment groups mismatched for confounders, even if there has been no defect in the randomisation process, simply due to the large number of subgroups that will be created. It might have been possible to quantify this effect by comparing the degree of confounder mismatching in each of the many simulated participant sample generated: however, this has not been reported (see later comments of sample size estimation).

2. **The assignment of a value of -1 HFD for any death is intended as a way of getting around the problem of scoring of mortality, it assumes that any death occurred as a result of failure to recover from surgery (i.e. would include road trauma death). This arbitrary work-about will not allow different gradations of failure of resilience between different patients who die (e.g. after 30 days compared to after 75 days). A time-to-event analysis might get around this problem, but this would add further complexity (in terms of understanding what is going on).**

3. **It should be noted that the paper does not describe the method of determination of death (with date and cause of death), which would be expected in the protocol. In Tasmania, the State government regularly searches the Australian National Death Register and transfers this information to the local public hospital patient medical records (in order to avoid sending clinic appointments to people who have died): this may not include the causes of death in the hospital records. A similar system may operate in their local hospital system, but the authors need to describe how their local system works. It is a standard procedure in medical research that the full list of participants is sent to the national death register to determine whether they have died, and if so, the causes. Is this what is intended to happen?**

4. **Monitoring of participant status during their time in the clinical trial is to be conducted by examination of the electronic health record for encounters within the hospital system, and by phone interviews at 30- and 90-days following the date of surgery for encounters outside the trial hospital system. The paper does not describe how loss to follow-up will be handled (unless it is intended that the missing data will be filled by statistical multiple imputation), or how large differences in actual follow-up times occur compared with intended times will be handled. This lack of detail in the measurement of the primary outcome needs to be corrected, since presumably this information has already been specified in trial documentation. It should be noted that multiple imputation is usually conducted under the assumption that missing data has occurred at random, whereas missing follow-up data may be missing preferentially amongst those showing the greatest effects of lack of resilience. Is there pilot data available to demonstrate the size of this problem?**

Outcome Measure Section - Response to Reviewer:

In response to the throughfall reviewer feedback above, we have better described the proposed medical model for patients at risk of loss of resiliency and frailty throughout the above responses and manuscript edits. Here, we will describe how the 90-day hospital free day outcome is pertinent and ideally flexible to capture the effectiveness of treatments in SPRY and SPRY-Metformin. Please see our responses to specific comments/questions in this section below.

Outcome Measure Section - Specific Reviewer Comments and Associated Responses:

"The authors claim that a long follow-up period is required to measure these effects, but that rests on the assumption that it is not possible to identify those at highest risk of early death, or delayed or inadequate recovery. However, it may be perfectly possible to select such a high-risk group, but that may require more elaborate screening of pre-operative surgical patients than is envisaged by the REMAP procedure chosen to address this particular clinical problem. The result of the trial design decision is that relatively large numbers of surgical patients will be recruited as participants of the trial who are at very-low risk of being affected by the lack of "resilience" being targeted by the SPRY treatments."

Response to Reviewer: We agree with the reviewer, the long term follow up required for traditional aging studies (i.e., the TAME study) may not be required and is not an efficient experimental model. In aged patients with comorbidity, the physiologic stress of a surgical intervention is considered by the National Institute for Aging to be an age-accelerating cause of frailty. Therefore, the inclusion and exclusion criteria for SPRY targets patients at risk of loss of resiliency and therefore allowing us to efficiently assess therapies on in this REMAP trial. Throughout the manuscript, we have elaborated how our screening and enrolment patient population are selected as a high-risk group.

Updated Manuscript:

Patient Selection

At the first level, patients exposed to the stress of an elective surgical intervention are identified. At the second level, participants are evaluated against the inclusion and exclusion criteria of the SPRY-Metformin domain identifying patients who, i) can be safely exposed to metformin and ii) are at risk of decreased physiologic reserve (i.e., older age and/or medical comorbidity) conferring postoperative morbidity and mortality at all levels of surgical stress (**Table 1**) [3,10]. Patients randomized in SPRY-Metformin can also participate in either or both substudies (microbiome or motor) as well as additional future domains on the SPRY Core Protocol.

"The question is whether the measure of hospital free days is an appropriate model for testing the different treatments that may be used for enhancing resilience. [I will also elaborate this concern later in the comments concerning the sample size simulations.]"

"The 90-day HFD measure will be affected by a number of effects, not all of which are affected in the same way by loss of "resilience"."

Response to Reviewer: Patients with decreased physiologic reserve have an over 10-fold increased risk of postoperative complications, increased rates of return to the operating room, longer index hospitalizations, and a 50% increased rate of unplanned hospital readmissions.[7,13–17] These outcomes directly and objectively translate into

a reduction in hospital free days for aged, medically complex patients undergoing both low and high operative stress procedures.[7,16] Further, this outcome will capture significant postoperative complications (i.e., acute care hospital readmissions for surgical site infection) as well as additional postoperative progression of frailty (i.e., acute care hospital readmission for sarcopenia and osteopenia related falls and fractures). We chose this outcome, as opposed to a cumulation of postoperative complications related to the surgical event, to capture many potential events and therefore the overall loss of resiliency or increase in frailty.

The reviewer also pointed out the importance of precision of language in regards to the readmission calculation. HFD are affected by acute care, not rehabilitation, hospital readmissions. This language was added throughout the manuscript for clarity.

Updated Manuscript:

Endpoints

The primary endpoint of SPRY Core Protocol hospital free days (HFD) up to 90 days [33–36]. This composite endpoint is an ordered categorical variable defined as the number of days from the day of surgery to the 90 thereafter, during which the patient is alive and free of hospitalization and was chosen for three reasons. First, this composite variable quantifies the care required for patients with reduced physiologic reserve with an increased risk of both specific postoperative complications (i.e., wound infections) and overall progression of frailty <="" span="" style="font-family: Arial;">[10,37–41]. Second, HFD is weighted (i.e. -1) to address potential effects on mortality, independent of the cause, throughout the 90-day postoperative period [33]. Third, time out of the hospital quantifies clinical outcomes and the cost of resource utilization, but reflects postoperative events important to patients and their families [42]. Therefore, HFD captures any treatment associated enhancements in resiliency across surgical strata and is applicable to SPRY-Metformin and any domain on the SPRY Core Protocol. The predefined and validated secondary clinical endpoints are listed in **Table 2** and **Table 3** [42–45].

Updated Citations:

Wahl TS, Graham LA, Hawn MT, Richman J, Hollis RH, Jones CE, et al. Association of the modified frailty index with 30-day surgical readmission. JAMA Surg. 2017;152(8):749–57.

Rothenberg KA, Stern JR, George EL, Trickey AW, Morris AM, Hall DE, et al. Association of Frailty and Postoperative Complications With Unplanned Readmissions After Elective Outpatient Surgery. JAMA Netw open. 2019;2(5):e194330.

Davenport DL, Henderson WG, Khuri SF, Mentzer RM, Richardson JD, Shemin RJ, et al. Preoperative risk factors and surgical complexity are more predictive of costs than postoperative complications: A case study using the National Surgical Quality Improvement Program (NSQIP) database. Ann Surg. 2005;242(4):463–71.

Shah R, Attwood K, Arya S, Hall DE, Johanning JM, Gabriel E, et al. Association of frailty with failure to rescue after low-risk and high-risk inpatient surgery. JAMA Surg. 2018;153(5).

Fagenson AM, Powers BD, Zorbas KA, Karhadkar S, Karachristos A, Di Carlo A, et al. Frailty Predicts Morbidity and Mortality After Laparoscopic Cholecystectomy for Acute Cholecystitis: An ACS-NSQIP Cohort Analysis. J Gastrointest Surg. 2020.

"The randomisation processes of randomised controlled trial are an attempt to create evenly matched populations in the different treatment groups. However, a heterogenous population that is intended to occur in this trial could result in treatment groups mismatched for confounders, even if there has been no defect in the randomisation process, simply due to the large number of subgroups that will be created. It might have been possible to quantify this effect by comparing the degree of confounder mismatching in each of the many simulated participant sample generated: however, this has not been reported (see later comments of sample size estimation)."

Response to Reviewer:

Additionally, we believe that the majority of heterogeneity within the outcome of interest will be based on preoperative drug duration, age, and surgical strata. Therefore, randomization is performed based on pre-specified randomization tables that use block randomization within each strata. Further, randomization is stratified by enrolment site, patient age, and preoperative study drug duration. Additionally, the primary analysis method adjusts for the surgical strata as a covariate. Additional baseline covariates will be reported in the trial publication within each

treatment group to determine if the groups are mismatched on any important measured confounding variables.

Updated Manuscript:
Digital Embedding
Patients meeting all inclusion and no exclusion criteria are allocated to a treatment regimen based upon the established randomization tables uploaded to the SPRY-Application.

SPIRIT Guidelines (Guideline, 16a)
Randomization is performed based on pre-specified randomization tables that utilize block randomization within each strata. Randomization is stratified by enrolment site, patient age, and the preoperative duration of study drug exposure.

"The assignment of a value of -1 HFD for any death is intended as a way of getting around the problem of scoring of mortality, it assumes that any death occurred as a result of failure to recover from surgery (i.e. would include road trauma death). This arbitrary work-about will not allow different gradations of failure of resilience between different patients who die (e.g. after 30 days compared to after 75 days). A time-to-event analysis might get around this problem, but this would add further complexity (in terms of understanding what is going on)."

Response to Reviewer: We agree with the reviewer, both the cause and the exact time of death in the postoperative period is not addressed in the primary outcome. As surgeons, we certainly feel the burden of a postoperative mortality that occurs in the operating room, during the hospital admission, or even within the 28-day window (captured by some governing surgical bodies as directly related to the operative event). However, the true clinically meaningful difference between a postoperative day 30 or day 75 death, is likely devastating equivalent to patients and their families and therefore, represented equally in our outcome.

As discussed above, as our goal is to quantify resiliency an intraoperative event (a myocardial infarction or significant blood loss) and postoperative events (an incidental fall or motor vehicle accident) are likely to confer poorer outcomes and increased risk of mortality in patients without physiologic reserve. Therefore, the specific cause of death is less pertinent to our interpretation of the results.

We have also been more transparent in the manuscript with this assumption.

Updated Manuscript:
Endpoints
This composite endpoint is an ordered categorical variable defined as the number of days from the day of surgery to the 90 thereafter, during which the patient is alive and free of hospitalization and was chosen for three reasons. First, this composite variable quantifies the care required for patients with reduced physiologic reserve with an increased risk of both specific postoperative complications (i.e., wound infections) and overall progression of frailty (i.e., progressive sarcopenia resulting in a fall and hip fracture) resulting in fewer HFD [10,37–41]. Second, HFD is weighted (i.e. -1) to address potential effects on mortality, independent of the cause, throughout the 90-day postoperative period [33]. Third, time out of the hospital quantifies clinical outcomes and the cost of resource utilization, but reflects postoperative events important to patients and their families [42]. Therefore, HFD captures any treatment associated enhancements in resiliency across surgical strata and is applicable to SPRY-Metformin and any domain on the SPRY Core Protocol. The predefined and validated secondary clinical endpoints are listed in **Table 2** and **Table 3** [42–45].

"It should be noted that the paper does not describe the method of determination of death (with date and cause of death), which would be expected in the protocol."

"It is a standard procedure in medical research that the full list of participants is sent to the national death register to determine whether they have died, and if so, the causes. Is this what is intended to happen?"

Response to Reviewer: This pertinent information has been added to the SPIRIT Guidelines (Appendix 1) and included below.

Of note, as patients enrolled in this trial are receiving care at UPMC, a hospital system and insurance provider, any and all regulations will be followed for clinical outcomes. This includes updating our EHR with death dates and cause of death, as mandated by local and national requirements.

Updated Manuscript:
SPIRIT Guidelines Appendix (Guideline, 18a)
Patient vitality, date and cause of death, is monitored in three ways in the clinical research data repository: 1) Prospective patient interaction at established contact points, 2) updates of electronic health record documentation of die within a UPMC healthcare system-based facility (e.g. nursing or rehabilitation facilities, emergency departments, and/or acute care hospitals), 3) monthly updates of the Social Security Administrative Death data files. Notably, when compared to a prospective patient registry, our combine (2) EHR vitality status and (3) Social Security Administrative data file is 94% sensitivity and 92% specificity. Therefore, in combination with prospective patient monitoring the internal validity of postoperative mortality is accurate.

"The paper does not describe how loss to follow-up will be handled (unless it is intended that the missing data will be filled by statistical multiple imputation), or how large differences in actual follow-up times occur compared with intended times will be handled. … It should be noted that multiple imputation is usually conducted under the assumption that missing data has occurred at random, whereas missing follow-up data may be missing preferentially amongst those showing the greatest effects of lack of resilience. Is there pilot data available to demonstrate the size of this problem?"

Response to Reviewer:
With have added a missing data section to the Statistical Analysis Appendix that better describes our primary plan for missing data as well as our missing not at random imputation strategies as sensitivity analyses.

Updated Manuscript:
Analysis Plan
Sensitivity analysis will explore a per protocol analysis and alterative imputation strategies that do not make missing at random assumptions.

Statistical Analysis Appendix (Section 2.4)
All missing data will be imputed based upon the median observed 90-day HFD value for each treatment arm and preoperative duration. Sensitivity analyses will utilize the following different imputation strategies that do not assume missing at random (MAR):

- Impute all missing values as the median observed 90-day HFD value under the placebo group.
- Impute all missing values as the worse observed 90-day HFD value within that surgical strata.

Sample size estimation:

To avoid the problem of false negative measurements in RCTs of clinical treatments, adequate numbers of participants must be observed to completion. This process involves estimation of the minimum numbers of participants require to detect a minimum relevant effect size (based on judgments that can be understood when a doctor and patient discuss consent for treatment) with predetermined levels of certainty (rate of Type 1 and 2 errors that are acceptable). There is no place for wishful thinking ("what benefits might be hoped for or expected") in therapeutic agent RCTs that may be acceptable in other scientific contexts (such as testing the concepts of scientific theories). Prospective power estimation is only appropriate where there is no control over the numbers of cases that can be observed. There are at least two ways of performing sample size estimation for the detection of treatment effects:

1. **Calculations using the statistical tests (mathematical model) that best match the clinical problem (medical model). There are limited numbers of sample size calculation equations for different analysis methods, and there all assume that the population effects of treatment are uniform and known. At least, if the calculations are possible, they can be done quickly;**
2. **Simulation of large numbers of datasets, based on the best understanding of what the proposed participant population might look like, and how they are affected by treatment. This is a much more laborious process, but it does allow a much wider range of populations, effects and associations to be examined. There are two types of simulation:**
   a. **One using a hypothesised distribution of a study population, effects and associations, with the population distribution being derived from a known population (such as an audit of recent patients in the hospital who might be enrolled as participants in the proposed trial).**
   This would be based primarily on the numerical model of the trial;
   b. **One modelling individual patients, including what is known about the clinical problem and the hypothesised effects of treatment, again with the simulated datasets derived from audit results. This would be a combination of the numerical and medical models of the trial;**
   c. **The numerical model would be simpler to perform but would not necessarily give much information about tailoring treatment to specific patients, whilst the combined numerical/medical model would require much more data and knowledge to reliably make additional predictions for the benefit of specific patients.**

Comments:

1. **The use of simulations in this trial is an appropriate response to the uncertainties that exist about the distribution of the participant population. The question arises as to what was done, and whether it has been adequately described.**
2. **Samples were taken from the audit sample of observed patients.**
3. **The numbers of participants required to achieve a pre-specified degree of certainty for a prespecified effect size benefit were not calculated.**
4. **Instead, it appears that a pre-specified sample size was determined by some process, and then an analysis was undertaken to determine whether the power was adequate. Apparently, the power of the trial using the pre-specified numbers was adequate, but this result seems to have occurred by chance. If some other process occurred, the authors need to describe what has happened more precisely.**
5. **No clear justification was given about the choice of effect size being sought in the trial. It is likely to be difficult to give an appropriate clear justification, since the primary outcome is a proxy measure made up of component effects, some of which would be responsive to the treatment and some would not. Given that larger benefits will be confined to the biologically older participants, and that the stressors from different surgeries will vary in this respect, there is a level of uncertainty around the measurement of effect size that does not appear to have been acknowledged in the sample size determination process.**
6. **What I would expect to see from the description of the sample size simulation would be that a series of simulations would be conducted in the manner described that involve a series of different inputted effect sizes, and a series of different sample sizes. Each of the effect size and sample size condition combinations would then be repeated multiple times (possibly 100 to 10,000 times depending on the judgment of those conducting the analysis) with the planned statistical analysis applied to each of the generated samples. The proportion (the power) of simulations in each of the conditions that produced a positive result (the effect size measurement was greater than the pre-determined threshold, usually defined by the alpha value in "probabilistic" statistical analysis) would then be counted for each of the samples sizes. Those counts would then be analysed by non-linear regression (power as outcome and sample size as predictor variable) to determine the exact number of participants that would be required that correspond to the pre-determined level of power being sought for their RCT. The result would be an exact number (say 2,347 total participants). It would be very unlikely that the answer would be a significant number (2,000 or 2,500). The authors may have decided that they would round their numbers up to a significant number for reasons of tidiness, but if this is the case, they should have said so.**

a. **My point is that it is impossible to be certain from the description whether the uncertainties that had been pre-defined determined the numbers of participants (which is the appropriate way of doing this sample size estimation), or that a choice of the numbers of participants drove the determination of the levels of uncertainties that were to be tolerated (which is not the appropriate way of going about this exercise).**
b. **If it was the former that happened, a simple improvement in the description in the protocol paper is all that would be required.**
c. **If it was the latter, either a re-analysis of the simulation data would be required if this could be done without repeating the simulations, or the whole sample size estimation may need to be repeated if the final trial result is to be taken as a definitive answer to the question of the utility of the trial treatment as applied to the trial population.**
d. **Please note that I think that this is less of a problem than my comments above might appear to imply. RCTs are only going to be definitive as a single study when the effect size is much larger than the minimum effect size that is judged before the trial to sufficient to justify applying the treatment to patients. In most circumstances, multiple RCTs are provide such a definitive answer, and those multiple trials need to be combined by metaanalysis (see comments below concerning how this RCT is compatible with meta-analysis). So long as the authors are committed to being cautious in their interpretation of their final results, I do not regard this issue as being the major problem that is usually thought.**

7. **A further issue with the sample size estimation process is the impact of the use of Bayesian analysis versus the more usual "probabilistic" statistical analysis methods. No justification is made of the use of the Bayesian methods, although they may be justifiable. One of the possible justification might be a reduced sample size requirement. It is a legitimate question as to what the sample size estimation would be if non-Bayesian analysis were to be used (see below my comments on the statistical analysis proposed in the protocol). The authors need to comment on this issue, and to make a much more extensive justification of their statistical methods.**

Sample Size Estimation Section - Response to Reviewer:

We, of course, agree with the reviewer in his thoughtful description of the importance of an established analysis plan and associated sample size calculation. Our results can only be interpreted in the context of the expected type I and type II error set a priori, driving our calculated sample size.

We in part chose a Bayesian analysis plan and adaptative randomization to minimize the sample size required. In this superiority trial, our type I error is fixed throughout the interim and final analysis to minimize the risk of false positive results. Similarly, the type II error is set for the trial overall. At interim analysis and adaptive randomization, patient allocation is adjusted: fewer patients to poorly performing or "losing" dose(s) and more patients to the best performing or "winning" dose(s). Inherent to this strategy, we can uncover a positive result with a minimal overall sample size at the cost of potentially making a false negative conclusion about "losing" dose(s). In other words, the trial is purposefully allocating statistical power to winning doses and away from losing.

Please see our responses to specific comments/questions in this section below.

Sample Size Estimation Section - Specific Reviewer Comments and Associated Responses:

"The numbers of participants required to achieve a pre-specified degree of certainty for a prespecified effect size benefit were not calculated. Instead, it appears that a pre-specified sample size was determined by some process, and then an analysis was undertaken to determine whether the power was adequate. Apparently, the power of the trial using the pre-specified numbers was adequate, but this result seems to have occurred by chance. If some other process occurred, the authors need to describe what has happened more precisely.

“What I would expect to see from the description of the sample size simulation would be that a series of simulations would be conducted in the manner described that involve a series of different inputted effect sizes, and a series of different sample sizes.”

“My point is that it is impossible to be certain from the description whether the uncertainties that had been pre-defined determined the numbers of participants (which is the appropriate way of doing this sample size estimation), or that a choice of the numbers of participants drove the determination of the levels of uncertainties that were to be tolerated (which is not the appropriate way of going about this exercise).”

Response to Reviewer: As the reviewer suggests, clinical trial simulations were used to optimize clinical trial design and to determine the sample size needed. For SPRY-Metformin, this includes the best thresholds for early success, dose dropping, and futility stopping and to obtain at least 80% power for a clinically meaningful treatment effect and a one-sided 2.5% type I error under the null distributions. Assumptions within these simulations were data-driven and based. We retrospectively abstracted data from patients identified within the UPMC health system undergoing similar interventions as those proposed in SPRY; therefore, these data were as similar to actual trial data as possible. Simulations were provided under a wide range of these clinical trial parameters to optimize the design with the design team. As it was not the goal of the manuscript or Statistical Analysis Appendix to describe the design process in its entirety, operating characteristics within the Statistical Analysis Appendix are provided only for the final proposed design. We have included and expanded a description of this process within the Statistical Analysis Appendix and moved specific content to the SPIRIT Guidelines Appendix.

Updated Manuscript:

Simulations and Sample Size Generation

Clinical trial simulations are used to optimize clinical trial design (best thresholds for early success, dose dropping, and futility stopping), to determine the sample size needed within this trial to obtain at least 80% power for a clinically meaningful treatment effect and a one-sided 2.5% type I error under the null distributions, and to quantify additional operating characteristics of the SPRY-Metformin trial. Utilizing pertinent retrospective UPMC EHR data, virtual patient datasets were created based on the observed distributions of the primary endpoint, 90-day HFD, within each stratum. Clinical trial simulations randomized patients to study drug and numerous trials were virtually executed, including all interim analysis and randomization adaptations. For simulated patients randomized to placebo, we assumed the primary outcome to be distributed similar to the observed 90-day HFD distribution per surgical strata within the UPMC EHR data. For simulated patients randomized to metformin, the distributions of 90-day HFD within UPMC HER data per surgical strata were shifted towards higher values of 90-day HFD being more likely based on a common percent reduction in 90-day hosptial days (90- [90-day HFD]). For examples of how the distributions were shifted see Appendix 2, Figure 4.1 and Table 4.2. The minimum clinically meaningful effect size was assumed to be a common percent decrease of 15% in 90-day mean hospital days for the highest treatment dose. This was chosen because it is sensitive to absolute differences in hospital days and treatments may have a larger absolute benefit for those procedures that are expected to result in more hospital days (Appendix 2, Table 4.2) [34].

Statistical Analysis Appendix (Section 4.0)

Clinical trial simulations are used to provide example trial results, to optimize clinical trial design (best thresholds for early success, dose dropping, and futility stopping) and to determine the sample size needed within this trial to obtain at least 80% power for a clinically meaningful treatment effect and a one-sided 2.5% type I error under the null distributions. Simulations were provided under a wide range of these clinical trial parameters to optimize the design with the design team. Operating characteristics are provided for the final design herein.

“No clear justification was given about the choice of effect size being sought in the trial. It is likely to be difficult to give an appropriate clear justification, since the primary outcome is a proxy measure made up of component effects, some of which would be responsive to the treatment and some would not. Given that larger benefits will be confined to the biologically older participants, and that the stressors from different surgeries will vary in this respect, there is a level of uncertainty around the measurement of effect size that does not appear to have been acknowledged in the sample size determination process.”

Response to Reviewer: We have provided a clearer description of the treatment effect and justification of the effect size. The trial was powered to detect a clinically meaningful reduction in the primary outcome. We chose to use a 15% reduction in mean hospital days as the common treatment effect across all subgroups. Importantly, the percent reduction results in different absolute effects in mean hospital days depending on the surgical subtype (see Appendix 2 Table 4.2). For example, among many patients undergoing total knee arthroplasty, a 15% reduction in mean hospital days would result in a half of a day reduction in hospital days (3.4 days in hospital vs. 2.9 days). Yet, for comparison patients undergoing an endovascular aortic repair, the expected reduction in hospital days is 1.6 (10.8 days in hospital vs. 9.2 days). Therefore, although this assumed clinically meaningful treatment effect is stable across all subgroups and patients, the percent reduction flexibly reflects important changes independent of the surgical intervention.

Updated Manuscript:

Simulations and Sample Size Generation

The minimum clinically meaningful effect size was assumed to be a common percent decrease of 15% in 90-day mean hospital days for the highest treatment dose. This was chosen because it is sensitive to absolute differences in hospital days and treatments may have a larger absolute benefit for those procedures that are expected to result in more hospital days (Appendix 2, Table 4.2) [34].

Statistical Analysis Appendix (Section 4.1)

The trial was powered assuming a common treatment effect of 15% reduction in mean hospital days across all surgical subtypes. The common treatment effect is specified as a percent reduction in mean hospital days to help elicit the minimal clinically meaningful treatment effect from the trial design team. The reduction of 15% in hospital days is thought to be the minimal clinically meaningful treatment effect within this patient population. This common percent reduction across surgical subtypes, results in different absolute effects in mean hospital days depending on the surgical subtype (see Table 4.2). In particular, for one of the most common surgical subtypes of Total Knee Arthroplasty, this would result in a half of a day reduction in hospital days (3.4 days in hospital vs. 2.9 days). In comparison, under Endovascular aortic replacement, the expected reduction in hospital days is 1.6 (10.8 days in hospital vs. 9.2 days). A percent reduction that that is at least 15% would result in a savings ranging from 0.2 – 3.9 hospital days for the 10 most common surgical types.

"A further issue with the sample size estimation process is the impact of the use of Bayesian analysis versus the more usual "probabilistic" statistical analysis methods. No justification is made of the use of the Bayesian methods, although they may be justifiable. One of the possible justification might be a reduced sample size requirement. It is a legitimate question as to what the sample size estimation would be if non-Bayesian analysis were to be used (see below my comments on the statistical analysis proposed in the protocol). The authors need to comment on this issue, and to make a much more extensive justification of their statistical methods."

Response to Reviewer:

We have included a more robust discussion of our reasoning for choosing a Bayesian approach both in the manuscript and the Statistical Analysis Appendix.

Updated Manuscript:

Statistical Analysis

The primary analysis plan for SPRY-Metformin includes a Bayesian ordinal logistic regression analysis of 90-day HFD to allow for borrowing of information on the treatment effect across different doses and durations of Metformin to maximally inform the research questions while minimizing the required patient sample size [8,9]. Complete documentation of the statistical analysis plan is including in the Statistical Analysis Appendix. (**Appendix 2**).

See responses to the **Statistical Analysis** comments from the review, below.

Statistical Analysis:

The authors propose to use Bayesian ordered logistic regression as the primary analytic test of the proposed primary outcome measure. If it is accepted that the 90-day hospital-free day count is the appropriate best outcome measure to determine the benefits of the SPRY treatments, then the numerical properties of that measure suggest that a rank-ordered multivariate regression analysis method is appropriate. Ordered logistic regression is probably the best method available if relatively simple analysis is to be undertaken. There are two problems that arise from this choice:

1. **Firstly, the result that is produced by ordered logistic regression is an odds ratio. This odds ratio does not necessarily mean the same thing as an odds ratio arising from the binary logistic regression analysis with which most readers will be familiar, and it is also not the same as a relative risk ratio. It would be the responsibility of the authors to make this clear in any final report of the trial (that will eventually be produced).**
2. **The authors tacitly acknowledge this problem by their discussion of mean numbers of HFD in the sample size estimation description, even though they have elsewhere acknowledged that the usual multiple linear regression analyses that estimate mean group values are not appropriate in this instance. This is an irreducible problem that I also run into in this type of circumstance. The solutions include:**
   a. **Conducting and reporting both linear regression and ordered logistic regression analyses, and then attempting to reconcile any differences in interpretation of the results that may arise (this is transparent but ambiguous, and readers may be left uncertain about what the authors think is the result of the trial);**
   b. **Conducting and reporting the ordered logistic regression analysis, and then attempting to explain to the readers, particularly the clinicians who have to explain to patients what the trial has shown. The doctor-patient encounter during which fully-informed consent is given for use of the treatment is the only really important target for any randomised controlled trial, and if the result of the trial cannot be understood reliably by both parties, then it is questionable whether a useful result has been achieved. Secondary customers of the trial, such as regulatory authorities, are only secondarily important in these decisions, and they have their own responsibilities to provide an answer to this dilemma;**
   c. **Conducting and reporting the linear regression analysis, is easier to undertake, and if ordered logistic regression analysis is used as a verification process and that the results of the two analytic methods give compatible interpretations, then the mean difference plus uncertainty would be more useful for the users of the trial results. If the results are not compatible, then the problem reasserts itself. As I said, this may be an irreducible problem.**
3. **Secondly, the authors propose to use a Bayesian version of the ordered logistic regression analysis. "Probabilistic" statistics operate on the assumption that the study being analysed arises anew as though nothing is previously known about the subject of the study. The cost of this is that larger numbers of observations are likely to be required to achieve a given level of certainty about the answer to the question. Bayesian statistics attempt to include something of what is previously known about the answer. The cost of this is that what is known may be affected by publication bias and less formal biases to which humans are prone.**
4. **If the authors have good reasons for use of a particular statistical method, this should be explained, including a discussion of the assumptions underlying each of the alternatives. These discussions ought to be included in any description of the results of the completed trial, so it might be useful for the authors to rehearse the arguments as part of their description of the protocol.**
5. **The authors should also address the question of how the results of Bayesian analysis as reported in their intended primary paper would be included in a meta-analysis of different trials of the proposed treatments.**
6. **The choice of statistical methodology is the prerogative of those conducting the trial. Their responsibility includes describing their methods in sufficiently clear language (i.e. plain health professional English) that will allow those who will implement the results of the final trial (i.e. clinicians) to make a reliable judgement about the veracity of those results. It is not the responsibility of clinicians to know and understand the details of arcane statistical methods. It is the responsibility of the statistician to help the clinicians to achieve the necessary level of understanding. In my judgement, the current text does not include the required level of clarity. The clarity is required in the primary results paper,**

but it might be useful if those describing the statistical methods and the results rehearse their explanations in this protocol paper.

Statistical Analysis Section - Response to Reviewer:

Response to Reviewer:

We have updated the Statistical Analysis Appendix to further clarify the description and justification of the Bayesian ordinal logistic regression primary analysis method. This includes a detailed description of the treatment effect that is assumed within the analysis and will be estimated in the trial.

Additionally, we have added a description of which clinically relevant and easily interpretable, for both patients and clinicians, summaries will be provided of the treatment effect from the trial. In particular, treatment effects will be summarized from the model as a common odds ratio across surgical subtypes, as well as translated into expected mean differences in HFD for each surgical subtype enrolled in the trial. Finally, we will report raw mean (and SD) differences in HFD for each surgical subtype, under each dose and duration. These raw estimates will not take into account the borrowing of information across doses and durations.

Please see **Overall Comments Section** and **Trial Participants Section** above for further elaboration and additional manuscript updates.

Updated Manuscript:

Statistical Analysis Appendix (Section 2.3.1)

The effect of each dose, *t*, and duration, *d*, will be summarized by reporting the posterior mean and 95% CI of the odds ratio, (common across all surgical strata). Additionally, we will translate the posterior mean odds ratio into expected mean differences in HFD for each surgical subtype enrolled in the trial. Finally, we will report raw mean (and SD) differences in HFD for each surgical subtype, under each dose and duration. These raw estimates will not take into account the borrowing of information across doses and durations.

Missing components of the analysis:

1. **The analysis plan may answer the questions posed by the trial, and the other SPRY domains will answer the related questions of effect of synergistic treatments.**
2. **However, what is missing is a refinement to the choice of patients who would most benefit from the proposed treatments. This could involve covariate analysis to identify a risk-score that may in the future help to select patients for the current proposed treatments, or further treatments that may be relevant for patient subgroups.**
3. **This would be an exploratory analysis, generating hypotheses for future studies. These studies may be conducted within the REMAP philosophy, or it may necessitate separate studies using more traditional methods. There may be information required for SPRY-type treatments for selected subgroups that are not currently or easily available from the electronic health record. The REMAP philosophy should not foreclose on future developments.**
4. **The authors may wish to consider including this type of analysis or describing current intentions for explicitly.**

Missing Components of the Analysis Section – Response to Reviewer:

Response to Reviewer:

In response to the above comments and questions from the reviewer, we have elaborated on the frailty mechanism and heterogeneity of treatment effect throughout the manuscript and associated appendixes. As we have discussed elsewhere in the responses to the reviewer, data abstracted for analysis is both automated, through the digital embedding of the trial, and confirmed through face-to-face patient interactions, as prompted by the SPRY-Application. Additionally, through both our simulations and our associated retrospective analysis (Reitz et. al., *JAMA Surgery* 2020), we have validated pertinent data.

Meta-analysis compatibility:

1. **It is unusual for a single RCT to produce a result that a completely definitive result is produced that renders all other clinical trials of the same treatment unnecessary. The effect size would have to be so large as to be unarguable, and that the trial team were not aware of this during the trial such that the trial continued despite superiority having been demonstrated if an interim analysis had been performed. Thus, in most circumstances, any one RCT cannot be regarded as definitive by the wider health community. The results need to be combined, and this happens by the process of meta-analysis.**
2. **When meta-analysis is undertaken, it is likely that at least some of the other trials will use "probabilistic" analysis, and I do not know how compatible the results of the two statistical methods would be. Would the reporting of the results include information that would be compatible, or is the reporting of the uncertainty of the Bayesian analysis incompatible with the usual uncertainty reported in the "probabilistic" analyses when the meta-analysis is done using the reported summary results?**
3. **A solution to this potential incompatibility problem would be making available an anonymised selected-variable dataset for use by qualified professionals for meta-analyses conducted according to pre-specified PRISMA (or equivalent) protocols. This would allow re-analysis of all the trials using identical statistical methods, either Bayesian or "probabilistic" or both. It would be helpful if this question is addressed, even though the SPIRIT guidelines do not explicitly require this information.**
4. **It should be noted that the PRISMA guidelines do not include the requirement for a sample size estimation. This would involve an analysis at the end of the meta-analysis about how many more participants would need to be recruited in future trials before further recruitment becomes unnecessary. The appropriate end-point for such a sample size estimation might be the number of participants after which the judgements about whether to use the treatment at the lower and upper 95% confidence intervals are the same, when those judgements are made by the doctor-patient dyad, and also the community decision-making processes make those judgement when comparing any alternative use of the health service resources (based maybe on cost-utility or similar analyses along with suitable political processes. Only when such sample size estimations have been done can decisions about whether to conduct further RCTs be made independent of the convenience and benefits of the usual initiators of RCTs.**
5. **The SPRIT and CONSORT guidelines do not require explicit statements about cooperation with processes of meta-analyses. This is an oversight, possibly due to disagreements about how this should be done. Authors could and should remedy this oversight in the case of their own RCT by stating how they intend to undertake such a cooperation.**

Meta-Analysis Capability Section - Response to Reviewer:

Response to Reviewer:

We have added a description of which summary statistics will be provided of the treatment effect from the proposed trial. Notably, as the reviewer suggests these data will allow for meta-analysis, as reported information will include raw estimates without the Bayesian probabilistic borrowing to be considered for meta-analysis with a trial using a frequentist approach.

For pertinent updates to the Statistical Analysis Appendix (Section 2.3.1), please see the above **Statistical Analysis Section**.

Updated Manuscript:

SPIRIT Guideline (Guideline 31a)

In particular, treatment effects will be summarized from the model as a common odds ratio across surgical subtypes, as well as translated into expected mean differences in HFD for each surgical subtype enrolled in the trial. These treatment effect estimates will be from the Bayesian primary analysis model that allows for borrowing of information across doses and durations of the

treatment. We will report raw mean (and SD) differences in HFD for each surgical subtype, under each dose and duration. These raw estimates will not take into account the borrowing of information across doses and durations and should be compatible with other trial publications for use in future meta-analyses.

Minor comments:

First sentence in introduction presumably should read 55 million (not thousand) population over 65.

Response to Reviewer: Thank you for identifying this error.

Updated Manuscript:
Background
"By 2020, over 55 million Americans will be greater than 65 years of age [26]."

"Page 7, line 40 in Trial Design: should "recurrently" (or similar) be used for "perpetually"? Perpetually suggests going on forever."

Response to Reviewer: This language was adjusted for accuracy, as suggested by the reviewer.

Updated Manuscript:
Trial design
Specifically, SPRY will recurrently assess multiple, Trial Steering Committee (TSC) approved, domains in multiple surgical strata and disease subtypes using response adaptive randomization and a comprehensive statistical analysis plan to create a self-learning health system.

"SPIRIT Item 11c Interventions: Adherence: "Patients compliance is queried at follow up patient encounters (Table 3, Figure 2)" on page 13, line 17, whilst the table quoted states "…..both patient safety and study drug compliance is monitored via phone interview (contact point 2)" on page 30, line 52. It is not made clear how tablet counting can be verified via a phone interview, but this problem is not addressed in the main text."

Response to Reviewer: As the reviewer suggested above, additional details of our trial protocol have been compiled into an appendix sequentially addressing the questions within the SPIRIT guidelines supplemental information (appendix 1). Providing this additional clarity

Updated Manuscript:
The above text has been omitted from the manuscript for brevity and the following information can be found in the SPIRIT Guidelines (Appendix 1):
Following confirming all inclusion and no exclusion criteria is met enrollment, and randomization study drug is provided to patients from established stock at each enrolling sight. Each study drug kit comes with the dosage specific number of 500mg of metformin ER or 500mg metformin ER matched placebo pills (i.e., two tablets per day for 1000mg metformin daily randomization). Patients allocated to the 1500mg arm are prescribed two 500mg tablets for seven days before ramping up to the full three tablet dose [6]. In the placebo arm, the same ramp up procedure and multiple dosages are used maintaining the blinded nature of this study.

Study drug is maintained throughout the duration of the preoperative period into the postoperative period and for 90 days thereafter. Notably, the medication is not discontinued or held, unless deemed medically necessary by the research or clinical team, in the perioperative period.

VERSION 2 – REVIEW

REVIEWER	Iain K Robertson College of Health and Medicine, University of Tasmania, Tasmania,

	Australia
REVIEW RETURNED	06-Aug-2020

GENERAL COMMENTS	1. The authors have made a significant attempt to address my concerns about comprehensibility and readability to the ordinary clinician. Although not perfect, the law of diminishing returns suggests to me that I should accept what has been done as satisfactory. 2. The separation of the details defining compliance with SPIRIT guidelines appears successful, and has made explicit two relevant details: a) that a highly detailed protocol document exists (as I expected) governing the conduct of the trial, which would be made available only after presumably satisfactory negotiations over such access to specific groups/organizations; b) that it is not intended for participant-level, anonymised data to be made available for the purposes of conduct of meta-analysis by appropriately skilled public-good organizations. 3. The authors have gone some way to providing an explanation of their statistical methodologies, and I judge this to be sufficient in the context of this paper. 4. The authors have not provided a fully satisfactory explanation for the use of Bayesian statistical techniques, but what is written is an improvement. Further, it suggests that there are efficiency gains available by making such a choice: for a given level of precision in the estimates of the effect sizes obtained, fewer patients would be required, or that greater precision could be obtained from the same numbers of patients (and trial resources). No reference is given for this implied assumption (Statistical Analysis Appendix, page 2, second paragraph). It may be that no appropriate reference is available. I am not concerned by this deficit, but the authors may wish to include a methodological analysis to be performed on the final data of this study comparing the precision of the results obtained from the Bayesian methods with the results obtained from non-Bayesian methods: this might be performed using the sort of simulation analysis used for the sample size justification. Please note that this is not a request to redesign the trial, but merely an elaboration to be conducted at some point in the future to provide retrospective justification for the choices made to the wider medical community. 5. The inclusion of a requirement for clinical trial registration in the Helsinki Declaration was made on the basis of ensuring that all available clinical trial data relevant to clinical decision should be made publically available to avoid biases in decision-making. There is also the need to arrive at reliable decision-making in the most efficient manner (soonest and with fewest trial resources expended and participants inconvenienced). The use of summary data in meta-analyses is inherently less efficient than meta-analyses based on participant-level data. 6. The whole point of the complex information management system described in this protocol, along with the use of Bayesian methods, is to improved the efficiency of the trial process. The decision about not making participant-level data would seem to reduce the efficiency of the trial process. At present, making participant-level data is not a Helsinki Declaration requirement, but it is possible that it may become one in the future. I am not asking for an answer to such a hypothetical question in this publication, but the authors might wish to consider how they would deal with this at some point in the future. In summary, I consider the manuscript a satisfactory presentation of

	the protocol of the clinical trial, assuming the authors are of the same opinion. It is suitable for publication in this form.