PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Ability of postoperative delirium to predict intermediate-term postoperative cognitive function in patients undergoing elective surgery at an academic medical center: protocol for a prospective cohort study
AUTHORS	Aranake-Chrisinger, Amrita; Cheng, Jenny; Muench, Maxwell; Tang, Rose; Mickle, Angela; Maybrier, Hannah; Lin, Nan; Wildes, Troy; Lenze, Eric; Avidan, Michael

VERSION 1 – REVIEW

REVIEWER	Gregory L Bryson Department of Anesthesiology and Pain Medicine University of Ottawa Anada I reviewed the ENGAGES protocol as a member of the Canadian Perioperative Anesthesia Clinical Trials (PACT) Group. I provided feedback on the ENGAGES2 protocol (in development) as a PACT member.
REVIEW RETURNED	24-May-2017

GENERAL COMMENTS	The formatter including Table of Contents and Abbreviations does not appear to be consistent with BMJOpen's protocol format or reporting guidance from SPIRIT or STROBE. Suggest these be deleted. The ENGAGES and SATISFY-SOS trials form the basis of this protocol. It is surprising that they are first mentioned on page 7 with BMJOpen's publication of the ENGAGES protocol first cited on page 8. I would encourage you to introduce these trials in the Background. A citation for the SATISFY-SOS protocol would be welcome. The data from the systematic review described is not available to the reader. An appendix summarizing these studies would be welcome. As an aside, 21 of 28 studies is not 84% as indicated in the text. Please correct While the persistence of cognitive deficits a year or more following surgery is subject of some debate, these deficits are the primary outcome of this trial (P9L43) I encourage you to describe studies that document POCD 6 months or more postop and a plausible range of its frequency. The background describes the relationship between postoperative delirium and postoperative cognitive dysfunction (objective 1) and

	functional decline (objective 2). The association of postoperative delirium and dementia (objective 3) is not discussed. I encourage you to add a brief introduction to this research question in the Background. The objectives statement that conclude the introduction should indicate the measures that define the three outcomes in question. As delirium is a the exposure variable of interest, its measure should also be clearly defined. Methods Postoperative cognitive dysfunction receives fairly scant attention among the ENGAGES study outcomes in the trial's registration on ClinicalTrials.gov and in BMJOpen (doi: 10.1136/bmjopen-2016-011505). Indeed cognition is listed fourth among the "other measures" alongside several other clinically relevant outcomes. The association of delirium and dementia is not mentioned in either document. This would appear to be an exploratory analysis and should be described as such. Research on POCD has been made more complicated by variable tests, testing batteries, and analysis methods. The rationale for choosing only two elements of the ISPOCD testing battery should be explicit. I am sceptical of the benefit of multivariable regression with scores of cognitive and quality of life entered as continuous dependent variables. At a minimum, the protocol should state a minimal clinically important difference in these outcomes. Similarly, I am uncertain why duration of delirium has been chosen as an independent variable. A reference indicating that a continuous rather than binary measure of this fundamental element of the research question must be provided. If duration is the desired metric, then the research question must be reworded to reflect this. It is unclear if duration or incidence of delirium will be used in the analysis of Objectives 2 and 3. Regarding sample size. Peduzzi's "rule of ten" is not a sample size estimate per se. Furthermore, it refers to the number of events, not observations, per variable required (PMID 8970487). It is unclear how sample sizes for the continuous dependent variables will be addressed in the regression model.

REVIEWER	Claudia Spies Charité – Universitätsmedizin Berlin, Charité Centrum 7

	Campus Virchow-Klinikum und Campus Charité Mitte Klinik für Anästhesiologie mit Schwerpunkt operative Intensivmedizin Charitéplatz 1 10117 Berlin Germany
REVIEW RETURNED	30-May-2017

GENERAL COMMENTS	The manuscript 'Ability of postoperative delirium to predict intermediateterm postoperative cognitive function in patients undergoing elective surgery at an academic medical center: protocol for a prospective cohort study' describes the study protocol of a nested cohort study meant to investigate postoperative delirium (POD) as a predictor of cognitive decline and quality of life one year after surgery. As both end points might be influenced by preoperative frailty, co-morbidities and perioperative complications these potential confounders will be assessed as covariables at baseline. The authors intend to evaluate incidence of dementia as a further endpoint. The proposed study will gain important insight in factors predicting negative cognitive outcomes after surgery both in the clinical context as well as in the light of demographic changes in the ageing society. Based on a sound literature search identifying conflicting results, the authors succeed in pointing out the relevance of further studies to address the importance of their effort in performing the proposed study, but several issues should be addressed before considering this manuscript for publication: Major comments: 1. Study design: The study is planned as a nested prospective cohort study within the ENGAGES study. The authors intend to retrospectively enroll patients from the ENGAGES study that itself is a substudy of the SATISFY-SOS study. The reason to do so is availability of preoperative cognitive testing that has been performed within the parental studies (cognition, though is not listed as study end point in the clinical register on www.clinicaltrials.gov). Whereas ENGAGES recruits approximately 1200 patients, the study protocol proposed in this manuscript cuts this number down to 200 participants only with inclusion criteria being residence of the participants within 45 miles of the test center or planned for postsurgical clinical visits at 10-16 months after surgery. For the reader of this article who might not be familiar with the ENGAGES and SATISFY-SOS studies, a flow chart would be of great help to understand the design. We advise the authors to clarify, if they intend to perform an intention to treat or observed only analysis. The rate of loss to follow up has been reported as a study limitation and might be relevant to analysis. 2. Methods: A) The proposed nested cohort study combines data of two large studies with different focuses. Wheras ENGAGES is an interventional study (intraoperative EEG guided anaesthesia) with postoperative delirium as primary outcome, SATISFY-SOS is an observational study with a broad spectrum of long term outcome measures declared as primary end points. This approach of combining two different study types and include a nested, semi-prospective cohort ("the current study will retrospectively identify 200 participants…", page 8, line 14) seems promising at first sight, but might cause difficulties in statistical

	analyses. To give an example, the authors do not specify how they will consider the group randomization (EEG guided protocol vs. Standard anesthetic protocol) as potential confounder. B) The inclusion criteria „residence of the participants within 45 miles of the test center or planned for postsurgical clinical visits at 10-16 months after surgery" seem a rather pragmatic and feasibility choice bearing a potential confounding risk themselves. C) Choice of cognitive test battery: Cognitive change will be assessed by three cognitive tests, namely the TMT-A, TMT-B and stroop color word test. In 1995 Murkin et al published a consensus statement in that more cognitive domains than executive function and attention have been recommended for cognitive test batteries used in studies with the end point POCD (Murkin 1995). Lack of memory tests in the proposed study's cognitive test battery is our major concern, especially if incident dementia will be used as a further endpoint. In Alzheimer's disease research affection of memory domains (amnestic Mild cognitive impairment) as compared to affection of attention and executive function (non-amnestic Mild cognitive impairment) has been shown to more likely progress into dementia (Petersen 2005; Busse 2006; Jungwirth 2012; Vos 2013). Measurement of POCD as a clinical entity related to Mild cognitive deficit disorders in as well nomenclature as potentially pathophysiology should therefore not be based on a cognitive test battery consisting of non-amnestic tests only. D) Paperbased cognitive testing vs. computerized cognitive testing (page 9, line 57): The comments in C) seem to have been discussed within the group of authors of this manuscript. The implementation of the NIH Toolbox cognition battery seems a valid test battery coherent with the Murkin consensus statement. Why should the participants choose between a short cognitive test battery (10 minutes testing time) and a long cognitive test battery (25 minutes testing time). Testing time will impact concentration and test performance. Under-performers, potentially those at risk of fulfilling POCD criteria might choose for the short version and be falsely tested negative for the end point POCD. From the above comments Specific aim #1 as described on page 7, line 33 should be reviewed prior to publication, as it is not conform with current action taken to standardize cognitive testing in POCD research. E) Calculation methods: On page 6, line 50, the authors state "Instead of using an arbitrary threshold to dichotomize cognitive function as normal or impaired, it would be more informative to correlate these outcomes with cognitive function as a continuous variable or stratified into multiple groups". Despite their own statement the authors later choose a dichotomous approach to calcualte their POCD variable using Z-scores. The ISPOCD study's „Reliable change index", compare citation 4 in the proposed manuscript, introduces one Z-score based option to control for natural variability and learning effects and therefore prevent arbitrary cut-offs. Our advice to the authors is the implementation of a non-surgical control group if using composite Z-scores. F) Incident Dementia: The definition of this end point remains vague (Evaluation for dementia with the Short Blessed Test will be completed for all patients at baseline, and again between one to two years after surgery; page 10, line 11). Putting the diagnosis dementia based on a screening test and via telephone raises our

	question if it fullfills sound neurologic criteria and may require more detailed work up. Furthermore, how will informed consent be guaranteed for those participants who are affected by cognitive disturbance due to dementia? As Eight-item Interview to Differentiate Aging and Dementia (AD-8) (page 9, line 24) is performed at baseline testing, will those participants with cognitive impairment at baseline be excluded from enrollment? What is the cut-off value? Minor Comments: 1. Page 6, line 11: “It [POD] is a neurological syndrome characterized by a combination of features, which can include an acute change, fluctuating course, disordered thinking, altered consciousness and inattention.” DSM-5 criteria are the gold standard of POD diagnosis. “Acute change and fluctuating course are criteria that have to be fulfilled to put the diagnosis POD. 2. We recommend the following reference to be included in this manuscript: Inouye SK, Marcantonio ER, Kosar CM, et al. The short-term and long-term relationship between delirium and cognitive trajectory in older surgical patients. Alzheimers Dement. 2016;12(7):766-775. doi:10.1016/j.jalz.2016.03.005. 3. Page 11, line 30: We suggest to include length of surgery as factor in the regression model as well. Although the proposed study protocol raises an important need of studies on long term perioperative cognitive trajectories and continues important considerations as published by one of the authors before (Nadelson 2014), to our understanding, some methodologic weaknesses prevent it from reaching some of its claimed specific aims (page 7), especially specific aim #3. We recommend the authors to clarify their methodology on the cognitive end points before the manuscript may be considered for publication or change title and focus back to an excellent study with a focus on quality of life measures. I hope that these comments help in your effort of the further handling of the manuscript. References Murkin JM, Newman SP, Stump DA, Blumenthal JA (1995) Statement of consensus on assessment of neurobehavioral outcomes after cardiac surgery. Ann Thorac Surg 59: 1289-1295. Petersen RC, Morris JC. Mild cognitive impairment as a clinical entity and treatment target. Arch Neurol. 2005;62(7):1160-1163; discussion 1167. doi:10.1001/archneur.62.7.1160. Busse A, Hensel A, Gühne U, Angermeyer MC, Riedel-Heller SG. Mild cognitive impairment: Long-term course of four clinical subtypes. Neurology. 2006;67(12):2176-2185. doi:10.1212/01.wnl.0000249117.23318.e1. Jungwirth S, Zehetmayer S, Hinterberger M, Tragl KH, Fischer P. The validity of amnestic mci and non-amnestic mci at age 75 in the prediction of alzheimer’s dementia and vascular dementia. Int Psychogeriatr. 2012;24(6):959-966. doi:10.1017/S1041610211002870. Vos SJB, Rossum IA van, Verhey F, et al. Prediction of alzheimer

	disease in subjects with amnestic and nonamnestic mci. Neurology. 2013;80(12):1124-1132. doi:10.1212/WNL.0b013e318288690c. Nadelson MR, Sanders RD, Avidan MS. Perioperative cognitive trajectory in adults. Br J Anaesth. 2014;112(3):440-451. doi:10.1093/bja/aet420.

VERSION 1 – AUTHOR RESPONSE

REVIEWER 1
Comments and Responses:

1. The formatter including Table of Contents and Abbreviations does not appear to be consistent with BMJ Open's protocol format or reporting guidance from SPIRIT or STROBE. Suggest these be deleted.

Response: These sections have been deleted.

2. The ENGAGES and SATISFY-SOS trials form the basis of this protocol. It is surprising that they are first mentioned on page 7 with BMJOpen's publication of the ENGAGES protocol first cited on page 8. I would encourage you to introduce these trials in the Background. A citation for the SATISFY-SOS protocol would be welcome.

Response: We have added a paragraph to the end of the Literature Review section. The two parent studies are further elaborated upon in the Study Design section and a flow chart has been added to further clarify the design. We have also included the NCT number for SATISFY-SOS.

3. The data from the systematic review described is not available to the reader. An appendix summarizing these studies would be welcome. As an aside, 21 of 28 studies is not 84% as indicated in the text. Please correct

Response: The systematic review data has been made available in appendix A. The text has been edited to "21 out of the 28 relevant articles (75%) ".

4. While the persistence of cognitive deficits a year or more following surgery is subject of some debate, these deficits are the primary outcome of this trial (P9L43) I encourage you to describe studies that document POCD 6 months or more postop and a plausible range of its frequency.

Response: We have added several sentences to the literature review section to further describe the time line of postoperative cognition.

5. The background describes the relationship between postoperative delirium and postoperative cognitive dysfunction (objective 1) and functional decline (objective 2). The association of postoperative delirium and dementia (objective 3) is not discussed. I encourage you to add a brief introduction to this research question in the Background.

Response: Thank you for your suggestion, we have added a paragraph with background regarding this question.

6. The objectives statement that conclude the introduction should indicate the measures that define the three outcomes in question. As delirium is a the exposure variable of interest, its measure should also be clearly defined.

Response: The measures that define the outcomes are clearly stated in the Methods and Analysis section of the abstract. We have also modified this section to include the use of the CAM to measure delirium.

7. Postoperative cognitive dysfunction receives fairly scant attention among the ENGAGES study outcomes in the trial's registration on ClinicalTrials.gov and in BMJOpen (doi: 10.1136/bmjopen-2016-011505). Indeed cognition is listed fourth among the "other measures" alongside several other clinically relevant outcomes. The association of delirium and dementia is not mentioned in either document. This would appear to be an exploratory analysis and should be described as such.

Response: The association between delirium and dementia will indeed be an exploratory analysis as we have stated in Specific Aim 3.

8. Research on POCD has been made more complicated by variable tests, testing batteries, and analysis methods. The rationale for choosing only two elements of the ISPOCD testing battery should be explicit.

Response: We are not trying to identify POCD, but evaluating specific cognitive domains that are primarily affected in delirium. To clarify, we have added "Given that delirium is predominantly a disorder of attention and executive function, we will focus our investigation on these cognitive domains" to the justification section.

9. I am sceptical of the benefit of multivariable regression with scores of cognitive and quality of life entered as continuous dependent variables. At a minimum, the protocol should state a minimal clinically important difference in these outcomes.

Response: With categorization of these variables, we would lose information and possibly skew results. For example, while some patients may have cognitive decline, the cognitive function in the cohort may be overall stable or improve. By categorizing cognition by an arbitrary threshold, we would emphasize decline while ignoring any potential improvement. We are interested in the ability of delirium to predict cognitive performance in the overall group, and do not think reporting incidences of decline defined by an arbitrary threshold would accurately represent the data. We have included minimal clinically important differences in the manuscript.

10. Similarly, I am uncertain why duration of delirium has been chosen as an independent variable. A reference indicating that a continuous rather than binary measure of this fundamental element of the research question must be provided. If duration is the desired metric, then the research question must be reworded to reflect this.

Response: Studies have suggested that the duration of delirium has important prognostic implications and postoperative cognition. We have included these references and adjusted the research question. We have also adjusted the analysis section to enter both POD incidence and duration into the model as categorical variables.

11. It is unclear if duration or incidence of delirium will be used in the analysis of Objectives 2 and 3.

Response: For aim 2, we will use both the incidence and duration of delirium. For aim 3, we will use the incidence of delirium. We have added this to the text.

12. Regarding sample size. Peduzzi's "rule of ten" is not a sample size estimate per se. Furthermore, it refers to the number of events, not observations, per variable required (PMID 8970487). It is unclear how sample sizes for the continuous dependent variables will be addressed in the regression model.

Response: Sample size calculations were updated for all three aims. The sample size for the multivariable regression was calculated using G*POWER, and has been updated in the manuscript. The sample size for the time to event analysis was calculated using PS, and has also been adjusted in the text.

REVIEWER 2

Major comments:

1. Study design: The study is planned as a nested prospective cohort study within the ENGAGES study. The authors intend to retrospectively enroll patients from the ENGAGES study that itself is a substudy of the SATISFY-SOS study. The reason to do so is availability of preoperative cognitive testing that has been performed within the parental studies (cognition, though is not listed as study end point in the clinical register on www.clinicaltrials.gov). Whereas ENGAGES recruits approximately 1200 patients, the study protocol proposed in this manuscript cuts this number down to 200 participants only with inclusion criteria being residence of the participants within 45 miles of the test center or planned for postsurgical clinical visits at 10-16 months after surgery. For the reader of this article who might not be familiar with the ENGAGES and SATISFY-SOS studies, a flow chart would be of great help to understand the design. We advise the authors to clarify, if they intend to perform an intention to treat or observed only analysis. The rate of loss to follow up has been reported as a study limitation and might be relevant to analysis.

Response: Flow chart has been added for clarity. Our study will include the randomization group in the statistical analysis as a potential confounder as suggested below. The randomization assignment, not the actual treatment, will be factored in the intention to treat analysis.

2. Methods:
A) The proposed nested cohort study combines data of two large studies with different focuses. Wheras ENGAGES is an interventional study (intraoperative EEG guided anaesthesia) with postoperative delirium as primary outcome, SATISFY-SOS is an observational study with a broad spectrum of long term outcome measures declared as primary end points. This approach of combining two different study types and include a nested, semi-prospective cohort ("the current study will retrospectively identify 200 participants…", page 8, line 14) seems promising at first sight, but might cause difficulties in statistical analyses. To give an example, the authors do not specify how they will consider the group randomization (EEG guided protocol vs. Standard anesthetic protocol) as potential confounder.

Response: We agree, and have included the group randomization as a factor in the multivariable regression.

B) The inclusion criteria "residence of the participants within 45 miles of the test center or planned for postsurgical clinical visits at 10-16 months after surgery" seem a rather pragmatic and feasibility choice bearing a potential confounding risk themselves.

Response: We agree that this was a feasibility issue - since the cognitive tests require a researcher to be physically present with the participant, phone interviews alone are not sufficient to determine cognitive ability. For many patients that live out of state, it was not feasible to ask participants to come

back to the facility or to have one of the research team travel long distances. It is possible that patients who live farther away are more likely to have greater or lesser decline in cognition/quality of life.

To address this limitation, we will compare the baseline characteristics of participants eligible for this study to those who were not eligible to determine if there are any potential confounding factors signifying a concern for sampling bias. We have adjusted the statistics section to include this comparison.

C) Choice of cognitive test battery: Cognitive change will be assessed by three cognitive tests, namely the TMT-A, TMT-B and stroop color word test. In 1995 Murkin et al published a consensus statement in that more cognitive domains than executive function and attention have been recommended for cognitive test batteries used in studies with the end point POCD (Murkin 1995). Lack of memory tests in the proposed study's cognitive test battery is our major concern, especially if incident dementia will be used as a further endpoint. In Alzheimer's disease research affection of memory domains (amnestic Mild cognitive impairment) as compared to affection of attention and executive function (non-amnestic Mild cognitive impairment) has been shown to more likely progress into dementia (Petersen 2005; Busse 2006; Jungwirth 2012; Vos 2013). Measurement of POCD as a clinical entity related to Mild cognitive deficit disorders in as well nomenclature as potentially pathophysiology should therefore not be based on a cognitive test battery consisting of non-amnestic tests only.

Response: Given that delirium is primarily a disorder of attention and executive function, we chose to focus on these cognitive domains. We will not be measuring or classifying patients as having postoperative cognitive decline, but simply comparing cognitive function in these 2 domains. While we agree that memory testing would be informative, we do not have baseline data available for this cognitive domain. We are however also collecting a battery of tests from the NIH toolbox at the postoperative assessment, which we can report as descriptive data. We will also add a secondary analysis comparing the results of the SBT, for which we have baseline data, and includes memory component. We have included this to the "pre-specified additional analyses" section.

D) Paperbased cognitive testing vs. computerized cognitive testing (page 9, line 57): The comments in C) seem to have been discussed within the group of authors of this manuscript. The implementation of the NIH Toolbox cognition battery seems a valid test battery coherent with the Murkin consensus statement. Why should the participants choose between a short cognitive test battery (10 minutes testing time) and a long cognitive test battery (25 minutes testing time). Testing time will impact concentration and test performance. Under-performers, potentially those at risk of fulfilling POCD criteria might choose for the short version and be falsely tested negative for the end point POCD. From the above comments Specific aim #1 as described on page 7, line 33 should be reviewed prior to publication, as it is not conform with current action taken to standardize cognitive testing in POCD research.

Response: We originally planned to complete all of the tests listed, however decided to use the abbreviated version for all participants to decrease assessment time and increase participation rates. The manuscript has been updated to reflect the shorter battery of tests.
E) Calculation methods: On page 6, line 50, the authors state "Instead of using an arbitrary threshold to dichotomize cognitive function as normal or impaired, it would be more informative to correlate these outcomes with cognitive function as a continuous variable or stratified into multiple groups". Despite their own statement the authors later choose a dichotomous approach to calcualte their POCD variable using Z-scores. The ISPOCD study's „Reliable change index", compare citation 4 in the proposed manuscript, introduces one Z-score based option to control for natural variability and

learning effects and therefore prevent arbitrary cut-offs. Our advice to the authors is the implementation of a non-surgical control group if using composite Z-scores.

Response: We will use Z-scores to combine the results of each cognitive test, but will not categorize the scores. We will leave the composite cognition score as a continuous variable. Since we will have baseline cognition scores and are comparing 2 surgical groups to each other, we would prefer not to use the RCI method for correction. We do not expect the tests we are using to have significant learning effects, given that they will only be administered twice and they will be administered approximately one year apart. We feel that using a non-surgical control group and the RCI may over-estimate decline since surgical patients will have more stress and anxiety in the preoperative period which likely affects performance on cognitive tests.

F) Incident Dementia: The definition of this endpoint remains vague (Evaluation for dementia with the Short Blessed Test will be completed for all patients at baseline, and again between one to two years after surgery; page 10, line 11). Putting the diagnosis dementia based on a screening test and via telephone raises our question if it fulfills sound neurologic criteria and may require more detailed work up. Furthermore, how will informed consent be guaranteed for those participants who are affected by cognitive disturbance due to dementia? As Eight-item Interview to Differentiate Aging and Dementia (AD-8) (page 9, line 24) is performed at baseline testing, will those participants with cognitive impairment at baseline be excluded from enrollment? What is the cut-off value?

Response: ENGAGES does not exclude patients based on the dementia screening tools, specifically because patients with a positive dementia screen represent a high risk population for postoperative delirium. After patients are consented, those who test positive for delirium with the CAM are excluded. For our study, we are using the SBT as a screening test for cognitive impairment. The SBT has 95% sensitivity and 65% specificity when compared with the MMSE, and is much faster to administer (this has been included in the text). While we cannot diagnose patients with dementia using this tool, it is nonetheless an effective screening method for this exploratory analysis. Patients will be consented over the phone for this aim, with approval from our IRB, as the test presents minimal risk and would not require informed consent in a clinical context. The patients have the right to refuse participation at any time.

Minor Comments:
1. Page 6, line 11: “It [POD] is a neurological syndrome characterized by a combination of features, which can include an acute change, fluctuating course, disordered thinking, altered consciousness and inattention.” DSM-5 criteria are the gold standard of POD diagnosis. “Acute change and fluctuating course are criteria that have to be fulfilled to put the diagnosis POD.

Response: Thank you for catching our mistake. The sentence has been edited to: “It is a neurological syndrome characterized by a combination of features, which must include an acute change, fluctuating course, inattention, and may include either disordered thinking or altered consciousness.”

2. We recommend the following reference to be included in this manuscript:
Inouye SK, Marcantonio ER, Kosar CM, et al. The short-term and long-term relationship between delirium and cognitive trajectory in older surgical patients. Alzheimers Dement. 2016;12(7):766-775. doi:10.1016/j.jalz.2016.03.005.

Response: This reference has been added in the Literature review section.

3. Page 11, line 30: We suggest to include length of surgery as factor in the regression model as well.

Response: Due to concerns for overfitting in the model, we will not include this factor.

Although the proposed study protocol raises an important need of studies on long term perioperative cognitive trajectories and continues important considerations as published by one of the authors before (Nadelson 2014), to our understanding, some methodologic weaknesses prevent it from reaching some of its claimed specific aims (page 7), especially specific aim #3.
We recommend the authors to clarify their methodology on the cognitive end points before the manuscript may be considered for publication or change title and focus back to an excellent study with a focus on quality of life measures.

I hope that these comments help in your effort of the further handling of the manuscript.

VERSION 2 – REVIEW

REVIEWER	Gregory L Bryson Department of Anesthesiology and Pain Medicine University of Ottawa Canada Have worked with senior author, M Avidan, through the Canadian Perioperative Anesthesia Clinical Trials (PACT) group.
REVIEW RETURNED	25-Aug-2017

GENERAL COMMENTS	I believe the authors have addressed all concerns raised in my initial review. I thank them for their efforts on this revised manuscript. One small addition is suggested ... In light of the recent ICMJE statement on Data Sharing (http://annals.org/aim/article/2630766/data-sharing-statements-clinical-trials-requirement-international-committee-medical-journal) I encourage the authors to update their "Reporting and Dissemination" section to address the reporting requirement of SPIRIT 31c "Plans, if any, for granting public access to the full protocol, participant-level dataset, and statistical code" I believe in open peer review, Gregory L Bryson, MD, FRCPC, MSc University of Ottawa

REVIEWER	Claudia Spies Charité – Universitätsmedizin Berlin, Charité Centrum 7 Campus Virchow-Klinikum und Campus Charité Mitte Klinik für Anästhesiologie mit Schwerpunkt operative Intensivmedizin Charitéplatz 1 10117 Berlin Germany

REVIEW RETURNED	29-Aug-2017

GENERAL COMMENTS	This is the revised manuscript 'Ability of postoperative delirium to predict intermediate-term postoperative cognitive function in patients undergoing elective surgery at an academic medical center: protocol for a prospective cohort study' by Aranake-Chrisinger and colleagues. While there are a number of changes in details, the main changes encompass the total number of patients to be recruited from the ENGAGES-study (130 instead of 200), notice of the Ethical committee approval of the Washington University in St. Louis, and additional adjustment for randomization status in the multivariable analyses. The authors did not change the method of cognitive assessment at follow-up, stating that their main interest was Trails A and B, and the Stroop Color and Word Test, which they evaluated with the ENGAGES team at study baseline. These tests will be accompanied by the Cognition Battery of the National Institutes of Health (NIH) Toolbox Assessment of Neurological and Behavioral Function as well as by tests evaluating attention, episodic memory, and executive function. Cognitive decline will be defined as a change in score of one standard deviation, and the sample size calculation is based on this difference. One may question the authors' assumption to define cognitive decline as a change in one standard deviation – some authors use 2 standard deviations [1,2]. Additionally and although the sample size calculation seems sound, we strongly recommend that this calculation is reviewed by a biostatistician. Frailty will be evaluated using the grip strength and the TUG. Here, the question still remains if this suffices to meet the requirements for a valid diagnosis of frailty as e.g. outlines by Fried and colleagues [3]. In the new manuscript the authors included a flow chart that guides the reader through recruitment procedures. In the part „statistical considerations", estimated loss to follow up rates are provided for each specific aim. Additionally, authors included „randomization group" as additional confounder in their statistical model (page 9). In their revised version the authors give additional information on assessment of memory function and dementia screening. The use of the NIH Toolbox cognition tests is a sound supplementation of the three cognitive tests (TMT and Stroop), that have been used for baseline assessment. The problem remains the retrospective design with no NIH toolbox assessment available at baseline. Nevertheless, the predictor is delirium and the outcome cognitive dysfunction at revisit. For the chosen setting of this study, the authors present a valid test battery to describe cognitive function in their cohort at the prospective time point with unavoidable limitations in baseline testing due to recruitment from a retrospective cohort. With regard to our previous concern to let the participant choose from a long-version and short-version of cognitive testing, the authors now select a fixed set of cognitive tests from NIH toolbox to all patients. We agree that specific aim 1 and title do not need adjustment prior to publication, as the term „POCD" is not used and „intermediate-term postoperative cognitive function" sufficiently explained in the methods. We also respect the author's decision to not implement a non-surgical control group, as this is a feasibility burden to their effort. Yet we strongly recommend discussing more into detail how the authors intend to control for natural variability and learning effects in repetitive testing.

	The authors did not raise further discussion on the ethical considerations concerning dementia/cognitive impairment at baseline. Minor Comments: 1. On page 3, lines 9- 11, the authors state: “It [POD] is a neurological syndrome characterized by a combination of features, which can include an acute change, fluctuating course, disordered thinking, altered consciousness and inattention.” This definition is maybe not incorrect but slightly imprecise. If you look at the correct wording of the DSM-5 definition, one might prefer to use a phrase such as (the following sentence is a suggestion of a non-native speaker and concerns the use of certain wordings, not the grammar): “It [POD] is a neurological syndrome characterized by a combination of features, which requires a disturbance in attention and awareness, representing a change from a baseline status, a fluctuating course, at least one additional cognitive symptom and no preexisting medical condition that might explain its presence. In the context of coma, delirium should not be diagnosed.” 1. Goldman JG, Holden S, Bernard B, Ouyang B, Goetz CG, Stebbins GT. Defining optimal cutoff scores for cognitive impairment using MDS Task Force PD-MCI criteria. Movement disorders : official journal of the Movement Disorder Society. 2013;28(14):1972-1979. doi:10.1002/mds.25655.2. 2. Goldman JG, Holden S, Ouyang B, Bernard B, Goetz CG, Stebbins GT. Diagnosing PD-MCI by MDS Task Force criteria: how many and which neuropsychological tests? Movement disorders : official journal of the Movement Disorder Society. 2015;30(3):402-406. doi:10.1002/mds.26084. 3. Fried LP et al. (2001) Frailty in older adults: evidence for a phenotype. J Gerontol A Biol Sci Med Sci 56:M146-156.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 1

Reviewer Name: Gregory L Bryson

Institution and Country: Department of Anesthesiology and Pain Medicine, University of Ottawa, Canada

Competing Interests: Have worked with senior author, M Avidan, through the Canadian Perioperative Anesthesia Clinical Trials (PACT) group.

I believe the authors have addressed all concerns raised in my initial review. I thank them for their efforts on this revised manuscript. One small addition is suggested ...

Comment: In light of the recent ICMJE statement on Data Sharing (http://annals.org/aim/article/2630766/data-sharing-statements-clinical-trials-requirement-international-committee-medical-journal) I encourage the authors to update their "Reporting and Dissemination" section to address the reporting requirement of SPIRIT 31c "Plans, if any, for granting public access to the full protocol, participant-level dataset, and statistical code"

Response: Thank you, we agree. We have included a data sharing statement in the protocol.

I believe in open peer review,
Gregory L Bryson, MD, FRCPC, MSc
University of Ottawa

Reviewer: 2
Reviewer Name: Claudia Spies
Institution and Country: Charité – Universitätsmedizin Berlin, Charité Centrum 7; Campus Virchow-Klinikum und Campus Charité Mitte Klinik für Anästhesiologie mit Schwerpunkt operative Intensivmedizin Charitéplatz 1
10117 Berlin, Germany
Competing Interests: None declared

Comment: This is the revised manuscript 'Ability of postoperative delirium to predict intermediate-term postoperative cognitive function in patients undergoing elective surgery at an academic medical center: protocol for a prospective cohort study' by Aranake-Chrisinger and colleagues. While there are a number of changes in details, the main changes encompass the total number of patients to be recruited from the ENGAGES-study (130 instead of 200), notice of the Ethical committee approval of the Washington University in St. Louis, and additional adjustment for randomization status in the multivariable analyses. The authors did not change the method of cognitive assessment at follow-up, stating that their main interest was Trails A and B, and the Stroop Color and Word Test, which they evaluated with the ENGAGES team at study baseline. These tests will be accompanied by the Cognition Battery of the National Institutes of Health (NIH) Toolbox Assessment of Neurological and Behavioral Function as well as by tests evaluating attention, episodic memory, and executive function. Cognitive decline will be defined as a change in score of one standard deviation, and the sample size calculation is based on this difference. One may question the authors' assumption to define cognitive decline as a change in one standard deviation – some authors use 2 standard deviations [1,2].
A: Many different thresholds have been used to define cognitive decline. While some authors use two standard deviations, others have used one standard deviation [1]. Since a minimal clinically important difference in cognition scores has yet to be defined, we chose to use the more conservative change of one standard deviation for this study.
Additionally and although the sample size calculation seems sound, we strongly recommend that this calculation is reviewed by a biostatistician.

Response: We have reviewed the sample size calculation with a biostatistician. While the power calculation is appropriate, he has advised us to increase the sample size given the number of factors we plan to include in the regression model to avoid overfitting. We will plan to recruit 200 patients; this will also give us greater power for the primary outcome of the study.

Comment: Frailty will be evaluated using the grip strength and the TUG. Here, the question still remains if this suffices to meet the requirements for a valid diagnosis of frailty as e.g. outlines by Fried and colleagues [3].

Response: As part of the preoperative evaluation and SATISFY-SOS, information regarding weight loss, endurance, and physical activity level are routinely collected. We have included this statement in the protocol. These variables have also been included in table 1.
In the new manuscript the authors included a flow chart that guides the reader through recruitment procedures. In the part „statistical considerations“, estimated loss to follow up rates are provided for each specific aim. Additionally, authors included „randomization group“ as additional confounder in their statistical model (page 9). In their revised version the authors give additional information on assessment of memory function and dementia screening. The use of the NIH Toolbox cognition tests is a sound supplementation of the three cognitive tests (TMT and Stroop), that have been used for baseline assessment. The problem remains the retrospective design with no NIH toolbox assessment available at baseline. Nevertheless, the predictor is delirium and the outcome cognitive dysfunction at revisit. For the chosen setting of this study, the authors present a valid test battery to describe cognitive function in their cohort at the prospective time point with unavoidable limitations in baseline testing due to recruitment from a retrospective cohort.

Comment: With regard to our previous concern to let the participant choose from a long-version and short-version of cognitive testing, the authors now select a fixed set of cognitive tests from NIH toolbox to all patients. We agree that specific aim 1 and title do not need adjustment prior to publication, as the term „POCD“ is not used and „intermediate-term postoperative cognitive function“ sufficiently explained in the methods. We also respect the author’s decision to not implement a non-surgical control group, as this is a feasibility burden to their effort. Yet we strongly recommend discussing more into detail how the authors intend to control for natural variability and learning effects in repetitive testing.

Response: Both the Stroop and Trails making tests have high test-retest reliability, and will have less variability due to error than less reliable tests [2]. The TMT does not demonstrate practice effects across larger time intervals, such as one year [3]. While the Stroop test has shown practice effects with repetitive testing, most studies administered the test many times and with shorter time intervals. If there is a learning effect between the baseline and one-year Stroop testing, it should be present in both the delirium cohort and the control cohort; thus any difference between groups is unlikely to be due to learning effects.
The authors did not raise further discussion on the ethical considerations concerning dementia/cognitive impairment at baseline.

Minor Comments:

1. On page 3, lines 9- 11, the authors state: “It [POD] is a neurological syndrome characterized by a combination of features, which can include an acute change, fluctuating course, disordered thinking, altered consciousness and inattention.”
This definition is maybe not incorrect but slightly imprecise. If you look at the correct wording of the DSM-5 definition, one might prefer to use a phrase such as (the following sentence is a suggestion of a non-native speaker and concerns the use of certain wordings, not the grammar):

“It [POD] is a neurological syndrome characterized by a combination of features, which requires a disturbance in attention and awareness, representing a change from a baseline status, a fluctuating course, at least one additional cognitive symptom and no preexisting medical condition that might explain its presence. In the context of coma, delirium should not be diagnosed.”

Response: The sentence has been rephrased for clarity.

1. Goldman JG, Holden S, Bernard B, Ouyang B, Goetz CG, Stebbins GT. Defining optimal cutoff scores for cognitive impairment using MDS Task Force PD-MCI criteria. Movement disorders : official journal of the Movement Disorder Society. 2013;28(14):1972-1979. doi:10.1002/mds.25655.2.
2. Goldman JG, Holden S, Ouyang B, Bernard B, Goetz CG, Stebbins GT. Diagnosing PD-MCI by MDS Task Force criteria: how many and which neuropsychological tests? Movement disorders : official journal of the Movement Disorder Society. 2015;30(3):402-406. doi:10.1002/mds.26084.
3. Fried LP et al. (2001) Frailty in older adults: evidence for a phenotype. J Gerontol A Biol Sci Med Sci 56:M146-156.

1. Rasmussen LS, Larsen K, Houx P, Skovgaard LT, Hanning CD, Moller JT. The assessment of postoperative cognitive function. Acta anaesthesiologica Scandinavica 2001;45:275-89.
2. Dikmen SS, Heaton RK, Grant I, Temkin NR. Test-retest reliability and practice effects of expanded Halstead-Reitan Neuropsychological Test Battery. J Int Neuropsychol Soc 1999;5:346-56.
3. Spreen O, Strauss E. A compendium of neuropsychological tests : administration, norms, and commentary. 2nd ed. New York: Oxford University Press; 1998.