PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	A study protocol for a national observational cohort investigating frailty, delirium and multimorbidity in older Surgical Patients: The Third Sprint National Anaesthesia Project (SNAP 3)
AUTHORS	Swarbrick, Claire; Poulton, Tom; Martin, Peter; Partridge, Judith; Moppett, Iain; SNAP 3 Project Team, SNAP 3 Project Team

VERSION 1 – REVIEW

REVIEWER	Beilby, Justin J.
REVIEW RETURNED	20-May-2022

GENERAL COMMENTS	This is a well crafted paper outlining an ambitious but important project. The protocol is well outlined in the article. I have a small number of specific suggestions particularly aimed at the international audience. I will emphasise the need for a detailed statistical review as part of the assessment of the article. My comments include: + Abstract I would suggest more detail be provided regarding the required target sample size. I am also interested that the authors consider adding a comment re the planned extension into Australia and NZ. This will position the study with the broader international focus. + Article Summary Useful key points. Really would like some more definitive and stronger comments regarding how the statistical methodology will adjust for the recruitment limitations + Introduction the summary is all quite logical but there is no mention of length of stay which is the primary outcome measure. Is this an oversight.? + Methods This is well documented but as I am unclear of how the National Institute of Academic Anaesthesia's Quality Audit and Research Coordinator (QuARC) and national Research and Innovation networks will work. Can another box/sentence be added for international reader. This will gauge how successful this hospital and researcher recruitment model will be. There is no overall discussion regarding ethics approval - is this hospital specific or to be completed nationally? + Conclusion I found the conclusion lack lustre. I would have preferred much more thought regarding next steps with the findings and how the findings could be quickly translated into clinical practice. Are there any other comparable studies in other disciplines where lessons could be learnt as the study is rolled out. ?

REVIEWER	McIsaac, Daniel The Ottawa Hospital, Anesthesiology and Pain Medicine
REVIEW RETURNED	16-Oct-2022

GENERAL COMMENTS	The Third Sprint National Anaesthesia Project (SNAP 3)- A Protocol for a National Observational Cohort Study of Frailty, Delirium and Multimorbidity in Older Surgical Patients Summary The authors present a protocol for a national audit project where data specific to older surgical patients will be collected over 5 weekdays with follow up to collect patient reported outcome data and linkage to routinely collected data. Included participants will be those 60 or older having a very general set of surgeries of various urgency classification. Exclusion largely centers on procedures of low invasiveness/complexity. Specific allowance is made to bridge issues of consent and capacity, which is a strength in this study and crucial for representativeness. Expectation is that the large majority (>90%) of non-private hospitals will be included. Specific variables under focus are the presence of frailty or multimorbidity before surgery, or the occurrence of delirium after surgery. The more distal 'primary' outcome is LoS, while PROMs, two version of days out of hospital and readmissions will also be collected. Details of in-hospital processes will be recorded, along with nested evaluation of workload and hospital structure/process features via survey. Overall, SNAP3 will be a foundational piece of science in having a broad and deep understanding of older surgical patients in the context of the Donebedian structure-process-outcome framework. Key aspects of evidence will likely be confirmed in a large and generalizable cohort, new knowledge gaps will be identified and novel insights into how older patients are cared for will be demonstrated. At the same time, there may be opportunities to build on these many strengths through consideration of some of the following points: 1-Person centred language: I am not sure the state of acceptance or requirement with person centered language in the UK related to older and vulnerable adults, but at least in North America (Canada and USA in my experience), leading societies have strongly encouraged and even required person centered language be used. In my experience this is also strongly preferred by older individuals. While I was glad to see the use of the term older (in place of elderly), I think that it may be preferable to refer to 'older adults with frailty', or 'older adults living with multimorbidity' instead of 'frail older', or 'multimorbid people' 2-Primary outcome-I suspect that it is too late to have much of an influence on this, but I was surprised to see LoS as the primary outcome. Clinically and data-wise, LoS tends to be high variance and is influenced by many heterogenous features beyond patient characteristics and what happens in hospital specifically for a given patient. There are also data directly from older patients that they do not prioritize LoS to the same extent that they prioritize many of the other outcomes that are planned for capture (especially those that reflect function and independence in day-to-day life; doi: 10.1007/s12630-022-02191-7). And in some core outcome sets for older people generally, LoS is at best a Tier 2 outcome that may be better reflected using DAH (10.1186/s12877-017-0701-3)

	3-Causal pathways reflecting the 3 primary variables of frailty, multimorbidity and delirium: First, I fully agree that these variables are absolutely crucial (especially frailty and delirium; multimorbidity I find more challenging as it dichotomizes a heterogenous count where someone with well controlled HTN, DM and a mild NSTEMI with no residual cardiac derangements would be considered lower risk than someone with severe HF as a sole condition), so I fully support the focus on them. At the same time, there is great complexity in untangling the causal pathway between them, and this would necessarily impact the approaches to analysis. While the directionality of comorbidity<->frailty is not fully determined, at least under the accumulating deficits model (which informs all 3 of the frailty measures used in SNAP3), comorbidities generally precede the development of meaningful frailty. Next, delirium is a post op outcome that will be influenced by the presence of both baseline characteristics (like frailty and multimorbidity), and what happens in the OR and after surgery. The directed acyclic graph here gets quite extensive/complex! This leaves me concerned that several analyses may be at substantial risk of overadjustment bias if all of the relevant variables are just included as equivalent terms in the models without accounting for their order in time and over the causal pathway. I don't think there is a perfect approach here, but I do think that since this is the core focus of SNAP 3, a greater degree of causal pathway elucidation and matching this to appropriate approaches (effect modification, effect mediation, etc) is likely required and should be reasonably prespecified. 4-Delirium measurement: First, I recognize that measuring delirium is a massive undertaking as most events occur outside of daytime hours, and no highly accuate tools exist (ie, with high sens and spec) to identify delirum via chart review. At the same time, I was surprised by the number of different approaches described. Why are both the CAM-ICU and 4AT being used prospectively? Or are you using them differentially between critically ill vs non-critically ill pts? Is the chart review tool the same/similar to the one described by Inouye and colleagues (which is fairly specific but lower in terms of sensitivity-meaningful in terms or assessing misclassification bias? 10.1111/j.1532-5415.2005.53120.x). As one of the 3 core measures I think that greater description and justification could be warranted. 5-Frailty-Similarly, it isn't entirely clear why more than 1 frailty instrument is being used, and then why the 3 described (rEFS, CFS and eFI) are chosen (vs the many others that are out there like the Phenotype, RAI, etc). I completely agree that the CFS should be used. It really has risen to the top in terms of accuracy and feasibility in periop and other acute settings. The CFS is also CPOC guideline recommended. The rEFS has some limitations, especially possible loss of fidelity with reported vs measured components. Questions also emerge about training to administer the instruments. For example, online tutorials exist for the CFS (10.1093/ageing/afab258) -There also aren't, as I can see anyhow, any references for the section on P9 that described collection of frailty data. I know that one of the authors, Dr Partridge, has some great periop frailty reviews, obviously the UK's Dr Clegg has pioneered the eFI, and Dr. Rockwood's team has developed the original FI and CFS (and EFS). This is also a useful review (at least I hope it is) of clinical frailty tools applied in the perioperative setting (doi 10.1097/ALN.0000000000003257) that I'd think would back up the approaches proposed.

	-In terms of agreement, while there is nothing wrong with doing these analyses, I don't think this is a huge gap. There are many papers that report data like these (and SRs of agreement in non periop settings) that generally show agreement is moderate at best, and often poor. The methods state that ICC will be used, but I am not clear on how this will be applied. Each of these tools is unique in terms of its scaling. Will they simply be dichotomized? If so, at what points. And if dichotomized why not use a Kappa statistic? Just questions that may be better asked by a biostatistician than a physician researcher, but one that informed readers may wonder about (I think). -Obj 5-What measures will be used to compare the differences in different frailty instruments as predictors? Are these just planned to compare effect sizes, or will you plan to use formal tests of predictive accuracy (eg, per Steyerberg PMID: 20010215 as recently operationalized for frailty instruments here 10.1016/j.bja.2022.07.019). Will this be a straight up frailty instrument vs frailty instrument, or frailty instrument added to some type of baseline model? 7-Delirium risk prediction model: Recent writings from the PROGRESS group (Riley, van Smeden, etc) have started to ask what seems like a very relevant question: Should we always be focused on deriving new risk prediction models when many (at best) internally validated models are already described? I don't mean to be difficult, but I do think this applies to this study. There are many systematic reviews of delirium risk prediction models published (eg, PMID 31413973 33354672 33354672 29705752 5516034 https://doi.org/10.1093/bja/aew476, etc). Might there be a role for external validation of some of the promising ones contained within? -Might there be enough data described across these many studies to at least clearly identify the clinically relevant variables that could be prespecified and avoid use of variable selection techniques that tend to lead to overfit and poor external performance? -Might there be substantive issues with your data that could be challenging for model derivation and subsequent validation given the point above about the multiple delirium measures being used. I am not trying to be difficult of obstructive, just wondering if there may be more optimal approaches? -Finally, while you describe the use of the c-stat as a measure of 'quality', I wonder if you may be interested in overall predictive accuracy, as measured by a suite of measures like those described by Steyerberg (PMID: 20010215), and which should likely include the most clinically relevant measure, calibration? 8-Objective 6-Can you clarify this a bit. Is the intention to simply describe the unadjusted difference in rates of interventions and outcomes based on frailty status? Are you concerned about substantial unmeasured confounding that will bias these estimates, or are purely descriptive values your aim? 9-Indication bias: Many of your objectives, and I know the team is likely aware of this, are going to suffer from substantive issues of confounding/indication bias, which will be unavoidable in an observational setting (eg, Obj 4, 9, 11). I wonder if a bit more pre-specification and acknowledgement of likely key sources of unmeasured confounding could be commented upon, along with the expected direction of these biases where such an estimate can made? This may help to better frame your results in advance and help to guide future reporting that will need to acknowledge these limitations.

	10-Weekend measures: Very sorry if I misunderstood this, but as I read this it appeared that all data will be collected M-F. There is then a comment about using IPW to forecast/project estimates to the weekend. However, if no weekend data are available, how would such a model even be constructed? Again, sorry if I missed something about weekend data collection that may make this more obvious 11-Minor comments to consider around reporting: - Objectives: This may be minutiae, and maybe I am wrong, but I believe in your primary objective frailty and multimorbidity will be measures of prevalence, whereas delirium will be an incidence -Might it be possible to add some specificity in the abstract's methods and analysis section to communicate a bit more clearly about your planned analyses for your lead objectives? E.g., We will estimate the prevalence of frailty and multimorbidity, and incidence of delirium, with 95%CIs. Unadjusted and multivariable adjusted regression models will be sued to estimate associations between primary exposures and outcomes… Something like that? -Strengths and Limitations: L15-I wonder if 'bias' might be more descriptive than 'skew' here? -Intro: Might you be able to provide some quantitative estimates about the effect sizes currently reported for your key measures to provide readers a bit more context (eg, P5L13-20)? -I know definitions and opinions vary, but at least as I've understood it based on what consensus is out there, frailty is more than a loss of physiologic reserve, but represents multidimensional loss of reserve (10.1093/gerona/gls119) -Might it be possible to share your key operationalized data points/case report forms as appendices with the protocol? -P8L10: Hoping you could clarify two points: 1) Why are you measuring both DAOH and DAH? They will be very, very similar, but add complexity to your study (I think). From a patient perspective DAH may be more meaningful as it does not focus only on hospital contributors to non-home days; 2) Sorry if this seems difficult, but I don't think DAH/DAOH are quality of life measures. I think you could lump all of these (DAH/DAOH/HRQoL as a 'patient centered outcomes' category, but not all under the heading of QoL -P10L51: Minor detail, but many of your 2nd outcomes, as I understand them, will not be dichotomous. So, I think you may be more accurate in saying that you will generate measures of association and 95%CI using regression models appropriate for each type of dependent outcome data (as opposed to saying all will be ORs). Obviously this is much more clearly explained in Obj 4, so perhaps just use the same description in both places Overall, I recognize that the comments in this review are quite extensive. These are simply meant to be constructive suggestions as I recognize the absolute importance of this project and the impact that it has, and therefore felt it worthwhile to provide in depth feedback based on my perspective as a future reader and user of the results of SNAP 3. I sincerely thank the team for the efforts applied to conduct of this study and reporting the protocol.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Justin J. Beilby
Comments to the Author:

This is a well crafted paper outlining an ambitious but important project. The protocol is well outlined in the article. I have a small number of specific suggestions particularly aimed at the international audience. I will emphasise the need for a detailed statistical review as part of the assessment of the article.
My comments include:
+ Abstract
I would suggest more detail be provided regarding the required target sample size.
This has been added.

I am also interested that the authors consider adding a comment re the planned extension into Australia and NZ. This will position the study with the broader international focus.
Added.

+ Article Summary
Useful key points.
Really would like some more definitive and stronger comments regarding how the statistical methodology will adjust for the recruitment limitations
The editor has requested a reduction in the number and length of the key points. We are very happy to adjust the key points to the satisfaction of the editor and Dr Beilby. We had this statement in mind: We will conduct sensitivity analyses using estimates of the numbers of elective and emergency surgery carried out at weekends analyses to gauge the likely bias due to the expected overrepresentation of elective surgery in our sample.

This sensitivity analysis is also mentioned under Objective 1.

+ Introduction
the summary is all quite logical but there is no mention of length of stay which is the primary outcome measure. Is this an oversight.?
Thank you – yes it was. We have added a brief sentence to introduce this.

+ Methods
This is well documented but as I am unclear of how the National Institute of Academic Anaesthesia's Quality Audit and Research Coordinator (QuARC) and national Research and Innovation networks will work. Can another box/sentence be added for international reader. This will gauge how successful this hospital and researcher recruitment model will be.
Thank you. We have added some more detail:
The QuARC network consists of one or more research / audit interested anaesthetists in every NHS hospital who act as a contact, and in many cases also as the local lead investigator for Health Services Research Centre (HSRC) projects. There is also national network of research and innovation support in the UK NHS, which facilitates research support for eligible studies. As a consequence…

There is no overall discussion regarding ethics approval - is this hospital specific or to be completed nationally?
Details of the ethical approval have been added.

+ Conclusion

I found the conclusion lack lustre. I would have preferred much more thought regarding next steps with the findings and how the findings could be quickly translated into clinical practice. Are there any other comparable studies in other disciplines where lessons could be learnt as the study is rolled out. ?

The editor has requested for there not be a conclusion so we have deleted this section. We agree that the implementation of findings are important, but this is not a formal part of our protocol.

We are grateful to Professor Beilby for his comments and suggestions.

Reviewer: 2
Dr. Daniel McIsaac, The Ottawa Hospital
Comments to the Author:
The Third Sprint National Anaesthesia Project (SNAP 3)- A Protocol for a National Observational Cohort Study of Frailty, Delirium and Multimorbidity in Older Surgical Patients
Summary
The authors present a protocol for a national audit project where data specific to older surgical patients will be collected over 5 weekdays with follow up to collect patient reported outcome data and linkage to routinely collected data.
Included participants will be those 60 or older having a very general set of surgeries of various urgency classification. Exclusion largely centers on procedures of low invasiveness/complexity. Specific allowance is made to bridge issues of consent and capacity, which is a strength in this study and crucial for representativeness. Expectation is that the large majority (>90%) of non-private hospitals will be included.
Specific variables under focus are the presence of frailty or multimorbidity before surgery, or the occurrence of delirium after surgery. The more distal 'primary' outcome is LoS, while PROMs, two version of days out of hospital and readmissions will also be collected. Details of in-hospital processes will be recorded, along with nested evaluation of workload and hospital structure/process features via survey.
Overall, SNAP3 will be a foundational piece of science in having a broad and deep understanding of older surgical patients in the context of the Donebedian structure-process-outcome framework. Key aspects of evidence will likely be confirmed in a large and generalizable cohort, new knowledge gaps will be identified and novel insights into how older patients are cared for will be demonstrated.
At the same time, there may be opportunities to build on these many strengths through consideration of some of the following points:

1-Person centred language: I am not sure the state of acceptance or requirement with person centered language in the UK related to older and vulnerable adults, but at least in North America (Canada and USA in my experience), leading societies have strongly encouraged and even required person centered language be used. In my experience this is also strongly preferred by older individuals. While I was glad to see the use of the term older (in place of elderly), I think that it may be preferable to refer to 'older adults with frailty', or 'older adults living with multimorbidity' instead of 'frail older', or 'multimorbid people'
Thank you for this reflection which we agree with – the manuscript has been adjusted accordingly.

2-Primary outcome-I suspect that it is too late to have much of an influence on this, but I was surprised to see LoS as the primary outcome. Clinically and data-wise, LoS tends to be high variance and is influenced by many heterogenous features beyond patient characteristics and what happens in hospital specifically for a given patient. There are also data directly from older patients that they do not prioritize LoS to the same extent that they prioritize many of the other outcomes that are planned for capture (especially those that reflect function and independence in day-to-day life; doi: 10.1007/s12630-022-02191-7). And in some core outcome sets for older people generally, LoS is at best a Tier 2 outcome that may be better reflected using DAH (10.1186/s12877-017-0701-3)

As a project team we discussed which outcomes to measure at length. We agree with Dr McIsaac that LoS is influenced by a multitude of other factors. However, we justify its use as a primary outcome for the following reasons:
1) PPI data from previous studies demonstrates that LoS is important to patients. This was supported by the PPI input in our project team.
2) LoS is relevant for healthcare costs. We cannot ignore the financial and organisational impact of an ageing surgical population, particularly in the context of resource-constrained public (and likely private) healthcare.
3) By reporting a range of secondary measures (including DAH) patients, public, and researchers will be able to assess the findings across domains.
As Dr McIsaac points out, we are unable to change the outcome at this stage.

3-Causal pathways reflecting the 3 primary variables of frailty, multimorbidity and delirium: First, I fully agree that these variables are absolutely crucial (especially frailty and delirium; multimorbidity I find more challenging as it dichotomizes a heterogenous count where someone with well controlled HTN, DM and a mild NSTEMI with no residual cardiac derangements would be considered lower risk than someone with severe HF as a sole condition), so I fully support the focus on them. At the same time, there is great complexity in untangling the causal pathway between them, and this would necessarily impact the approaches to analysis. While the directionality of comorbidity<->frailty is not fully determined, at least under the accumulating deficits model (which informs all 3 of the frailty measures used in SNAP3), comorbidities generally precede the development of meaningful frailty. Next, delirium is a post op outcome that will be influenced by the presence of both baseline characteristics (like frailty and multimorbidity), and what happens in the OR and after surgery. The directed acyclic graph here gets quite extensive/complex! This leaves me concerned that several analyses may be at substantial risk of overadjustment bias if all of the relevant variables are just included as equivalent terms in the models without accounting for their order in time and over the causal pathway. I don't think there is a perfect approach here, but I do think that since this is the core focus of SNAP 3, a greater degree of causal pathway elucidation and matching this to appropriate approaches (effect modification, effect mediation, etc) is likely required and should be reasonably prespecified.
We agree entirely, and plan to elaborate the details in a statistical analysis plan prior to undertaking the modelling. We have revised to explain that we will use DAGs to clarify hypothesized causal relationships and to aid in selection of covariates.

4-Delirium measurement: First, I recognize that measuring delirium is a massive undertaking as most events occur outside of daytime hours, and no highly accuate tools exist (ie, with high sens and spec) to identify delirum via chart review. At the same time, I was surprised by the number of different approaches described. Why are both the CAM-ICU and 4AT being used prospectively? Or are you using them differentially between critically ill vs non-critically ill pts? Is the chart review tool the same/similar to the one described by Inouye and colleagues (which is fairly specific but lower in terms of sensitivity-meaningful in terms or assessing misclassification bias? 10.1111/j.1532-5415.2005.53120.x). As one of the 3 core measures I think that greater description and justification could be warranted.

We agree that the measurement of delirium is inherently problematic and for this reasons several approaches were taken. We have clarified this in greater detail in the manuscript. To clarify some of these points:
Prospective assessment for delirium was carried out either with 4AT or CAM-ICU. 4AT was completed in person prospectively as much as possible. CAM-ICU is only for those patients who were critically ill at the time of assessment, which is likely to be a small number of patients.

When it was not possible to record a prospective assessment using the 4AT or CAM-ICU, we asked researchers to record if the results of these assessments were recorded via a retrospective notes

review. The notes review also looked for common phrases/words mapped to the DSM 5 diagnostic criteria for delirium. The notes review has limitations of both sensitivity and specificity. We will have the opportunity to estimate this prospectively to an extent given that a large number of participants will have a prospective 4AT and notes review completed.

5-Frailty-Similarly, it isn't entirely clear why more than 1 frailty instrument is being used, and then why the 3 described (rEFS, CFS and eFI) are chosen (vs the many others that are out there like the Phenotype, RAI, etc). I completely agree that the CFS should be used. It really has risen to the top in terms of accuracy and feasibility in periop and other acute settings. The CFS is also CPOC guideline recommended. The rEFS has some limitations, especially possible loss of fidelity with reported vs measured components. Questions also emerge about training to administer the instruments. For example, online tutorials exist for the CFS (10.1093/ageing/afab258)
-There also aren't, as I can see anyhow, any references for the section on P9 that described collection of frailty data. I know that one of the authors, Dr Partridge, has some great periop frailty reviews, obviously the UK's Dr Clegg has pioneered the eFI, and Dr. Rockwood's team has developed the original FI and CFS (and EFS). This is also a useful review (at least I hope it is) of clinical frailty tools applied in the perioperative setting (doi 10.1097/ALN.0000000000003257) that I'd think would back up the approaches proposed.
Thank you for raising this and for flagging the systematic review which informed this study, and which has now been appropriately referenced. As with the choice of delirium tools the decision to use different frailty tools was in part pragmatic – different services use different tools and this study aimed to record current practice in addition to actually measuring frailty. Although literature has become clearer about use of CFS, when this study was referred to the ethics committee the CPOC-BGS guidelines had not yet been published.
At the time of protocol finalisation we did not know how widely eFI was used, either at geographical level (which primary care practices record it) and perhaps more usefully whether it was reaching into the surgical pathways. If eFI turns out to be a good automated tool in this population then this may have implications for how pathways are designed.
We have edited the manuscript accordingly.

-In terms of agreement, while there is nothing wrong with doing these analyses, I don't think this is a huge gap. There are many papers that report data like these (and SRs of agreement in non periop settings) that generally show agreement is moderate at best, and often poor. The methods state that ICC will be used, but I am not clear on how this will be applied. Each of these tools is unique in terms of its scaling. Will they simply be dichotomized? If so, at what points. And if dichotomized why not use a Kappa statistic? Just questions that may be better asked by a biostatistician than a physician researcher, but one that informed readers may wonder about (I think).
We agree that these agreements (or lack thereof) have been reported by others. There are a couple of reasons for us keeping these analyses (some good, some just being honest). The 'good' reasons are that:

- we believe this is the largest perioperative cohort measuring both the frailty scales and a suite of outcomes.
- good science supports reproduction of previous findings and it is efficient for us to do this within the wider study

The 'keeping us honest' reason is that if we didn't report these associations someone is highly likely to either ask us to do so, or to request the data to do it themselves. We would rather be transparent about this as a pre-planned analysis rather than post hoc.
Regarding the actual analysis,
It is possible to measure consistency using ICC after rescaling each measure of frailty to have a common mean and variance. On reflection, however, we think that pairwise Spearman's correlations better fit the aims of our study, and have changed the manuscript accordingly. We have also added a

specification and justification for measuring agreement between dichotomized versions of the frailty measures.

-Obj 5-What measures will be used to compare the differences in different frailty instruments as predictors? Are these just planned to compare effect sizes, or will you plan to use formal tests of predictive accuracy (eg, per Steyerberg PMID: 20010215 as recently operationalized for frailty instruments here 10.1016/j.bja.2022.07.019). Will this be a straight up frailty instrument vs frailty instrument, or frailty instrument added to some type of baseline model?

Thank you for suggesting this reference, which wasn't published at the time we submitted this protocol for review. We have added this. The analysis section for Obj 5 states that we will compare the three frailty measures with respect to the analyses specified for Objectives 1-3, which involve univariate and bivariate analyses only. We have added a sentence to clarify that we won't extend the three-way comparison to the multivariable models under Obj 4.

7-Delirium risk prediction model: Recent writings from the PROGRESS group (Riley, van Smeden, etc) have started to ask what seems like a very relevant question: Should we always be focused on deriving new risk prediction models when many (at best) internally validated models are already described? I don't mean to be difficult, but I do think this applies to this study. There are many systematic reviews of delirium risk prediction models published (eg, PMID 31413973 33354672 33354672 29705752 5516034 https://doi.org/10.1093/bja/aew476, etc). Might there be a role for external validation of some of the promising ones contained within?

We agree that simply adding another model into the delirium ecosystem is unlikely to be helpful on its own. We disagree (slightly) that there sufficient good models already exist. Interestingly a recent review of general inpatient delirium models specifically called for development of robust models.

Lindroth H, Bratzke L, Purvis S, et al Systematic review of prediction models for delirium in the older adult inpatient BMJ Open 2018;8:e019223. doi: 10.1136/bmjopen-2017-019223

Furthermore, most models are based on relatively small cohorts, or single specialty studies or administrative dataset, distinct from the data available to SNAP3.

We are open to external validation of existing models (limited by concordance of measured predictors). We haven't prespecified this in the protocol as we felt it is far enough down the secondary / tertiary analyses to be formally specified at a later date.

-Might there be enough data described across these many studies to at least clearly identify the clinically relevant variables that could be prespecified and avoid use of variable selection techniques that tend to lead to overfit and poor external performance?
We agree. This was intended to be included in our step 1 but perhaps 'candidate' was doing too much heavy lifting here. We have revised:
… candidate predictors, identified from previous studies and clinical insight, and the probability of delirium…

-Might there be substantive issues with your data that could be challenging for model derivation and subsequent validation given the point above about the multiple delirium measures being used. I am not trying to be difficult of obstructive, just wondering if there may be more optimal approaches?

The comment is not perceived as being obstructive or difficult at all. If we have understood correctly, Dr McIsaac may have concerns that if our delirium outcome is not the same as another study (more or less specific / sensitive) then it becomes difficult to validate externally. We agree – again to an

extent. We can (and will) report any derived tool on the basis of a sub or superset of delirium outcomes (e.g. in person 4AT positive only or 'any' delirium for instance). Given that there is currently no gold standard or core outcome set for delirium detection we hope this will be helpful.

-Finally, while you describe the use of the c-stat as a measure of 'quality', I wonder if you may be interested in overall predictive accuracy, as measured by a suite of measures like those described by Steyerberg (PMID: 20010215), and which should likely include the most clinically relevant measure, calibration?
We have clarified that we will also investigate model calibration and calculate the Brier score.

8-Objective 6-Can you clarify this a bit. Is the intention to simply describe the unadjusted difference in rates of interventions and outcomes based on frailty status? Are you concerned about substantial unmeasured confounding that will bias these estimates, or are purely descriptive values your aim?

For this objective we simply wish to describe the variation in models of care across the participating hospitals.

9-Indication bias: Many of your objectives, and I know the team is likely aware of this, are going to suffer from substantive issues of confounding/indication bias, which will be unavoidable in an observational setting (eg, Obj 4, 9, 11). I wonder if a bit more pre-specification and acknowledgement of likely key sources of unmeasured confounding could be commented upon, along with the expected direction of these biases where such an estimate can made? This may help to better frame your results in advance and help to guide future reporting that will need to acknowledge these limitations.
Again, we agree, and we fully acknowledge all the limitations of an observational study.
Examination of associations between baseline variables and outcomes will potentially be confounded by
[a] other clinical/patient characteristics – age (will be included in models as standard), socioeconomic status (co-confounding with frailty/multimorbidity and very difficult to measure)
[b] processes of care i.e. POPS models, medical input, anaesthetic preparation, Level 2/3 usage
We are capturing all these data. However, by definition it is somewhat difficult to estimate the size and magnitude of unmeasured confounders.

10-Weekend measures: Very sorry if I misunderstood this, but as I read this it appeared that all data will be collected M-F. There is then a comment about using IPW to forecast/project estimates to the weekend. However, if no weekend data are available, how would such a model even be constructed? Again, sorry if I missed something about weekend data collection that may make this more obvious
This was indeed not clear. We have obtained numbers of elective and emergency surgeries at representative hospitals, which we used in the process of planning our study, and we plan to re-weight data based on those estimates. We have changed the text to clarify this.

11-Minor comments to consider around reporting:
- Objectives: This may be minutiae, and maybe I am wrong, but I believe in your primary objective frailty and multimorbidity will be measures of prevalence, whereas delirium will be an incidence
This is correct and thank you for spotting that.

-Might it be possible to add some specificity in the abstract's methods and analysis section to communicate a bit more clearly about your planned analyses for your lead objectives? E.g., We will estimate the prevalence of frailty and multimorbidity, and incidence of delirium, with 95%CIs. Unadjusted and multivariable adjusted regression models will be sued to estimate associations between primary exposures and outcomes… Something like that?
Thank you. We have followed these suggestions.

-Strengths and Limitations: L15-I wonder if ‘bias’ might be more descriptive than ‘skew’ here?
We agree and we would have changed the manuscript accordingly were it not for reducing the quantity and length of these points.

-Intro: Might you be able to provide some quantitative estimates about the effect sizes currently reported for your key measures to provide readers a bit more context (eg, P5L13-20)?
We have added detail from previous studies including findings of the impact of frailty and multimorbidityon outcomes.

-I know definitions and opinions vary, but at least as I’ve understood it based on what consensus is out there, frailty is more than a loss of physiologic reserve, but represents multidimensional loss of reserve (10.1093/gerona/gls119)
Manuscript amended accordingly.

-Might it be possible to share your key operationalized data points/case report forms as appendices with the protocol?

We have included an appendix with the data points.

-P8L10: Hoping you could clarify two points: 1) Why are you measuring both DAOH and DAH? They will be very, very similar, but add complexity to your study (I think). From a patient perspective DAH may be more meaningful as it does not focus only on hospital contributors to non-home days; 2) Sorry if this seems difficult, but I don’t think DAH/DAOH are quality of life measures. I think you could lump all of these (DAH/DAOH/HRQoL as a ‘patient centered outcomes’ category, but not all under the heading of QoL

DAH and DAOH will of course be closely correlated, we have two reasons for measuring both. First, in the older patient failure to return home is a feared outcome (particularly if permanent) which is completely missed by DAOH. For more general studies DAOH is far easier to collect and likely represents the bulk of the non-hospital time, but we felt that it was important to (try) to collect DAH. Second, DAH is an emerging measure which is favoured by some (for the reasons above) but as yet has relatively limited validation. It is harder to collect, so we hope we will be able to demonstrate either that it adds value that is worth the effort (as above) or perhaps it is so closely linked to DAOH that it is not worth the bother.

We have changed the text:

We will also determine the ‘days at home’ (DAH) and ‘days alive and out of hospital’ (DAOH) at 120 days as a measure of the process of recovery that has been shown to be of importance to patients,45. Days alive and out of hospital is available from central records, and hence easier to collect at scale, but excludes time in residential or nursing home care, outcomes which are feared by older patients. Days at home, is more difficult to capture, but more closely aligns with what patients want from a good recovery. A possible by product of the study is a demonstration of whether the collection of DAH is worth the additional research burden.

-P10L51: Minor detail, but many of your 2nd outcomes, as I understand them, will not be dichotomous. So, I think you may be more accurate in saying that you will generate measures of association and 95%CI using regression models appropriate for each type of dependent outcome data (as opposed to saying all will be ORs). Obviously this is much more clearly explained in Obj 4, so perhaps just use the same description in both places

Thank you. We have made the statement for Obj 2 more generic to match obj 4.

Overall, I recognize that the comments in this review are quite extensive. These are simply meant to be constructive suggestions as I recognize the absolute importance of this project and the impact that it has, and therefore felt it worthwhile to provide in depth feedback based on my perspective as a future reader and user of the results of SNAP 3. I sincerely thank the team for the efforts applied to conduct of this study and reporting the protocol.

We are very grateful to Dr McIsaac. His comments are universally constructive, and we hope we have addressed them satisfactorily.