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)	Making Shared Decision Making (SDM) a Reality: protocol of a large-scale long-term SDM implementation program at a Northern German University Hospital
AUTHORS	Danner, Marion; Geiger, Friedemann; Wehkamp, Kai; Rueffer, Jens Ulrich; Kuch, Christine; Sundmacher, L; Skjelbakken, T; Rummer, Anne; Novelli, Anna; Debrouwere, Marie; Scheibler, Fueloep

VERSION 1 – REVIEW

REVIEWER	Sophie Hill Centre for Health Communication and Participation, School of Psychology and Public Health, La Trobe University, Australia I am engaged in a funded project examining the implementation of shared decision making at 8 Victorian health services.
REVIEW RETURNED	24-Mar-2020

GENERAL COMMENTS	This is an outstanding protocol and the SDM community around the world will be eagerly awaiting results. I had one comment for consideration but it is up to the authors as to whether they make any changes. I am not requesting any. SDM and the development of decision aids will require clinicians to know more about the alternative options (to the main option they usually emphasise) than they did before and so they may need time and encouragement to understand and accept these. It cannot be assumed that there won't be some resistance. And if you encourage patients to ask questions, clinicians also need strong encouragement (and perhaps for some even communication skills training) to learn expanded skills in communicating with patients – rather than just answering questions. Thinking about this makes me suggest that decision aids might need to be improved/amended as you go along, as the pros and cons that patients identify become clearer or change. I apologise if you had already anticipated these issues – I really enjoyed reading the paper.

REVIEWER	Anne M Stiggelbout Leiden University Medical Center The Netherlands
REVIEW RETURNED	14-Apr-2020

GENERAL COMMENTS	This paper provides an extensive overview of a study that is more than halfway its execution. That makes my review mostly one aimed at correcting the text, since no changes can likely be made anymore. In several instances where I query e.g. the design, I therefore suggest that the authors argue why the chose/decided to do what

they did.
Further, regarding the text, I suggest having someone perform a strict edit for meandering text, as the manuscript would benefit from some streamlining, making it more succinct, and less informal to read. After that, an English text editor or native should read it carefully, since there are many instances where the text is awkard, or in my view as a non-native English speaker, erroneous.

The paper provides a nice showcase for a large implementation effort in a German hospital., of which I am envious. I like the Health economic impact analysis using propensity scoring.

As such the paper is not very innovative. It is not really implementation research, for it lacks, e.g., a good process evaluation, but also implementation strategies, but it is not an effectiveness study either, given the weak design.
Data analysis is missing.

Therefore I am not sure that this protocol would benefit the reader. Since the study is two thirds on its way by the time this would be published, why not await the results? . II am not sure that many others can follow this example, unfortunately, so there is no need to publish for that reason. Or else, rewrite it in a much less extensive version.

Abstract
from the abstract I immediately ask myself, what is the difference between 1 and 4 (clinicians in communication, HCPs in coaching)? [later I read it means “other HCPs”}

L. 14: is based on should read involves or entails,

Body of text
P7, L11 Similarly, it is an advantage to include clinicians, isn’t it?
P9, L 12-13 “In this study (…) implementation” reads as if the authors failed to implement a DA in their study, but they surveyed developers, please rephrase.

P9, L. 20 “No program”? They just referred to Northern Norway, ref 10. Further, Ammenthorpe and Dahl Steffenson (2019), in the Danish cancer hospital performed an endeavor quite similar to the one presented here.

The dates are October 2017 – October 2021, why publish this paper so late? In the meantime the above two studies were performed…

P10, L17: how do you know beforehand that exactly 83 ptDAs will be developed?
In this paragraph, it is confusing that clinicians seem to be physicians? Are the decision coaches not clinicians? In my understanding nurses, physiotherapists, etc are also termed clinicians.
It seems that integrative refers only to ptDA building, but it should also refer to development of the training and coaching. Was this not the case?

P11, L. 6 “ before being used “ now in English refers to sample patients, not to the pdDAs, I think.
I suggest writing “four-year project” rather than ”4 year project”

	P11. L.20 these questions will be addressed: this could be more specific, how do the authors envisage to do this? Further, I find it confusing that for the implementation both NPT and CFIR are used. How are these related? Are they used for different aspects, or phases? If CFIR is already used, what does NPT add? Please clarify. Setting What does 200.000 cases mean? First outpatient visits? Hospitalizations? This is an unusual term, cases of what? Of illness? P12, L 13, each time new clinics are started? Sounds as if the hospital starts a new clinic, whereas I assume that a clinic enrolls in the program? From now on I will not mention things anymore that strike me in the English, but suggest that a native speaker read the text. P14,L.4 What is meant by sensitive here? Note that DA and EBpDA are used interchangeably. P15, L1: what is meant by this sentence: “As for the clinician training, etc.” Is this something else than the decision coaching? This entire section is long and informal. Some info is redundant, but then the content of the individual coaching sessions is missing. Or is that the clinician training in line 1? Of the PICS I would be interested in the number of items per scale, and the total scale scores/ranges. Further, for the PICS nothing is said about before and after the intervention, but I assume this holds, similarly to the Mappin? Ideally referring to the same consultation as that of the Mappin? (later I read more on this, but that is confusing, perhaps reordering of the text would help?) I wondered why the Collaborate is used and not the SDM-Q9, since the latter is the most widely used (providing lots of comparison data) and the Collaborate has extreme ceiling effects (necessitating so called Top scores), and reflects the effort of the clinician as perceived by the patient, and thereby is more a satisfaction measure (“the SDM was not very good, but at least the doc did her best….”). Top page 17: is this an informal way of presenting the power? I would like to have seen the sample size calculation here, to account for the N=16.00. (see further my comments below) As to the design, why a pre-post without a control group, why not an interrupted time series, or a stepped wedge design. Particularly looking at figure 2, that almost looks like a picture of a stepped wedge design! With such large numbers of clinicians to be trained, and multiple interventions, at least one of these alternative designs would have been possible and made a more compelling case. With the current attention to SDM, one can now not know whether the intervention worked or this was simply an effect of time? Please argue. P17, L. 12 what is SDM treatment? Power calculation

	L14: 9 items ranging from 0-4 and a satisfactory score is 1.5? Is that an error? Or are this one, as well as the PICS, divided by the number of items, leading to a total of 0-4 (and thus 1-4 for PICS)? This would explain the sample size calculation for the PICS as well. Please clarify. L19: “a sample size of about 40 for each group/clinic at each measurement (assuming a power of 80% and a level of significance of 5%, (one-sided, assuming positive effect).” How many clinics/ groups are there, for I assume you calculate a total sample, and divide by that to get to 40? Unless you expect an effect of nesting (intraclass correlation, with some clinics already doing better than others) and you need to correct for this design effect? Or is 40 needed in total, as it seems from the statement two lines below, that an effect can be seen at the individual clinic level? That seems highly unlikely? And then below it is said that “These numbers will allow to measure significant differences in the primary endpoint not only at the campus but also at the individual clinic level (at least in the larger clinics).” Are there clinics that are much larger than 40? How large would they need to be to assess an effect at individual clinic level? It would help me if this section on the power were rewritten to make it clear what the total N needs to be, and whether this can be simply the sum of all the clinics. And one sided testing is unusual, even with hypotheses (most research assumes one direction)! What are the numbers with two-sided? Why is the Mappin used if nothing is to be shown with it? I would expect it to be more sensitive to changes, so why not given an estimate of the power with the numbers you are expecting to obtain? Final page P19. I suggest deleting the formative assessment altogether, since as presented now, it does not provide sufficient information to the reader, and may not be too relevant in this respect? Or else, extend this section. I am not sure that I understand the role of the Implementation section. Is this meant as Process Evaluation, in addition to the other evaluations just described? Then this should be made clear, and the instruments described. If it is meant as part of the intervention, it should be moved up. I am not familiar with the CFIR so cannot judge Table 1 The figures are too small for the readers.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Reviewer Name: Sophie Hill
Institution and Country: Centre for Health Communication and Participation, School of Psychology and Public Health, La Trobe University, Australia

I am engaged in a funded project examining the implementation of shared decision making at 8 Victorian health services.

Reviewer: This is an outstanding protocol and the SDM community around the world will be eagerly awaiting results.
I had one comment for consideration but it is up to the authors as to whether they make any changes. I am not requesting any.
SDM and the development of decision aids will require clinicians to know more about the alternative options (to the main option they usually emphasise) than they did before and so they may need time and encouragement to understand and accept these. It cannot be assumed that there won't be some resistance. And if you encourage patients to ask questions, clinicians also need strong encouragement (and perhaps for some even communication skills training) to learn expanded skills in communicating with patients – rather than just answering questions.
Thinking about this makes me suggest that decision aids might need to be improved/amended as you go along, as the pros and cons that patients identify become clearer or change.

I apologise if you had already anticipated these issues – I really enjoyed reading the paper.

Answer: Thank you for your kind words. We actually take patient perspective on the pros and cons of treatment alternatives into account at several points throughout the development of decision aids (DA). First, we do so-called needs assessments with UKSH patients in a specific DA topic to structure the decision aid according to patient needs and to take patient needs into account when evidence is searched for. Second, eachecision aid is user tested after completion. In this respect, we absolutely agree with you: any additions or changes that patients might suggest in these steps or even after implementation of the DA will be checked and potentially included in one of our regular DA updates. DA updates are planned to be done on a yearly basis.

Reviewer: 2
Reviewer Name: Anne M Stiggelbout
Institution and Country: Leiden University Medical Center The Netherlands Please state any competing interests or state 'None declared': None declared

Reviewer comment: This paper provides an extensive overview of a study that is more than halfway its execution. That makes my review mostly one aimed at correcting the text, since no changes can likely be made anymore. In several instances where I query e.g. the design, I therefore suggest that the authors argue why the chose/decided to do what they did.
Further, regarding the text, I suggest having someone perform a strict edit for meandering text, as the manuscript would benefit from some streamlining, making it more succinct, and less informal to read. After that, an English text editor or native should read it carefully, since there are many instances where the text is awkard, or in my view as a non-native English speaker, erroneous.

The paper provides a nice showcase for a large implementation effort in a German hospital., of which I am envious. I like the Health economic impact analysis using propensity scoring.
Answer : Thanks for this. We sincerely hope to be able to share and discuss our results in some not too far away "post Corona" times with you and the scientific community. Please find below our comments individually for your suggestions and notes.

Reviewer: As such the paper is not very innovative. It is not really implementation research, for it lacks, e.g., a good process evaluation, but also implementation strategies, but it is not an effectiveness study either, given the weak design.
Data analysis is missing.

Answer: We actually do consider process evaluation a crucial part of the project. However, we might have underreported on this part since the remaining part is already quite lengthy and the process evaluation ought to be part of later publications. We agree with your second comment, but we had good reasons to go for a long-term implementation study instead of testing the combined effect of our complex intervention in an RCT. Since the evidence supporting the effectiveness of the individual S2C components (trainings, DAs in use, patient activation) is already strong (see respective references in text), we decided to go for a large-scale long-term implementation study as was suggested as a last step in intervention development by Campbell et al. in their "Framework for design and evaluation of complex interventions to improve health" (Campbell 2000). This is also in line with the requirements of the Innovation Fund, the sponsor and peer reviewer of our project. Our intent was to realize the intervention in a comprehensive real life hospital setting to learn more about differences and synergies between clinical departments, where and why SDM implementation on its different levels works well/not so well, etc. This is why process evaluation is so important and closely related to our implementation strategies. The CFIR Table 1 is used in its original meaning to create a knowledge base of development and implementation factors of our complex intervention (www.cfirguide.org). In the project, it is used on the one hand to document the baseline at UKSH for project development and implementation. On the other hand, we use the CFIR to follow-up on implementation and assess whether and to what degree the constructs will influence (positively or negatively) the success of implementation or where modification are needed. Since we agree with your critique here, we decided to…: 1. Change the title with respect to "large-scale long-term SDM implementation program", which more appropriately reflects what we do, and 2. Add more detail on what our study is in the introduction and 3. Provided detail on the process evaluation.

Regarding your last point: We added information in Table 2 on outcome measures and how data analysis is planned for each outcome measure and restructured text accordingly. Data analysis (T1) was planned to start in early 2020, but had to be postponed due to the unexpected Corona situation. This makes a protocol publication still or even more important to inform about the project design.

Reviewer: Therefore, I am not sure that this protocol would benefit the reader. Since the study is two thirds on its way by the time this would be published, why not await the results? I am not sure that many others can follow this example, unfortunately, so there is no need to publish for that reason. Or else, rewrite it in a much less extensive version.
Answer: Please see below answering your comment on timing of publication in detail.

Abstract
Reviewer: from the abstract I immediately ask myself, what is the difference between 1 and 4 (clinicians in communication, HCPs in coaching)? [later I read it means "other HCPs"}
Answer: Thank you. We changed this to "physicians" on the one hand and "other health care professionals" on the other. We used the term "clinicians" meaning "physicians working in a hospital setting" – sorry, I think this was a language misinterpretation on our side and we appreciate your advice. We also changed the word "clinics" to "clinical departments" (recommendation of our native speaker reviewing the paper now)

Reviewer: L. 14: is based on should read involves or entails,
Answer: Thanks, we changed this.

Reviewer: Body of text
P7, L11 Similarly, it is an advantage to include clinicians, isn't it?
Answer: added

Reviewer: P9, L 12-13 "In this study (…) implementation" reads as if the authors failed to implement a DA in their study, but they surveyed developers, please rephrase.
Answer: Thanks, we rephrased the sentence.

Reviewer: P9, L. 20 "No program"? They just referred to Northern Norway, ref 10.
Further, Ammenthorpe and Dahl Steffenson (2019), in the Danish cancer hospital performed an endeavor quite similar to the one presented here.
Answer: Thanks, we added references to a few similar projects that are
ongoing (Dahl Steffensen 2018; Sondergaard 2019) or planned (Scholl 2018). However, it seems that to our program is unique in that it addresses SDM interventions for several target groups at a time: physicians, other health care professionals / nurses, and patients. Also, more than 80 decision aids tailored to specific clinical departments' needs and their patients are developed at the same time.

Reviewer: The dates are October 2017 – October 2021, why publish this paper so late? In the meantime the above two studies were performed…
Answer: Yes, you are absolutely right. However, in this study, we were able to start the intervention and complete team recruiting only mid 2018 due to a delay in ethics approval at UKSH. Since this large-scale project intervention is a tremendous effort in each component (training physicians and other HCPs in SDM, develop decision aids, prepare and institute the ASK-3 program), which needed much more preparation/roll out time and man power than we had anticipated. We

did simply not find the time twrite and submit a study protocol. However, we still consider it of great value at this point in time to inform the scientific community about this project and indeed we do hope that this project will deliver insights for other researchers planning comparable projects and encourage scientific exchange with these. Analyses will equally be delayed now due the specific Corona situation and the shut-down that hit all clinical departments in the German health care setting.

Reviewer: P10, L17: how do you know beforehand that exactly 83 ptDAs will be developed?
Answer: Based on our budgetary constraints, we calculated the maximum possible number of decision aids to be developed per clinical department at UKSH (more in larger departments, less in smaller departments). So, this number (83) resulted from this calculation and is our goal to be met by the end of study duration. We added a rationale for this in the text. Thanks.

Reviewer: In this paragraph, it is confusing that clinicians seem to be physicians? Are the decision coaches not clinicians? In my understanding nurses, physiotherapists, etc are also termed clinicians.
Answer: see previous comments

Reviewer: It seems that integrative refers only to ptDA building, but it should also refer to development of the training and coaching. Was this not the case?
Answer: In DA development, it was very important to integrate patients and physicians to develop DAs along their needs and increase acceptance/use. However, the trainings had been previously developed and tested in studies (see references by Geiger, Kasper, Berger-Höger). At that stage (training development), feedback from physicians has been continuously integrated. Regarding the decision coach training, we did also include input from HCPs at initial stages of training development.

Reviewer: P11, L. 6 "before being used " now in English refers to sample patients, not to the pdDAs, I think. I suggest writing "four-year project" rather than "4 year project"
Answer: Thanks, wording changed.

Reviewer: P11. L.20 these questions will be addressed: this could be more specific, how do the authors envisage to do this? Further, I find it confusing that for the implementation both NPT and CFIR are used. How are these related? Are they used for different aspects, or phases? If CFIR is already used, what does NPT add? Please clarify.
Answer: Thanks. We clarified in the text how these were used:

- NPT did drive our study implementation plans at the microlevel from the beginning. It focuses on what individuals involved do and what they think about what they do.
- CFIR is a broader construct that emphasizes different levels of factors important to our study planning and implementation that go beyond NPT and provide a rather "macro"-perspective on study implementation in this complex intervention effort: intervention characteristics, the setting (outer setting, inner setting), characteristics of individuals, and process.

CFIR was used in this publication to provide the reader a rather "comprehensive" picture of what plays a role for project project planning and realization. It entails the multilevel ecological factors at study initiation.

Reviewer: Setting
What does 200.000 cases mean? First outpatient visits? Hospitalizations? This is an unusual term, cases of what? Of illness?
Answer: We clarified this to "cases treated each year", meaning the number of cases treated each year, which can be one patient treated twice per year = 2 cases (but only one patient)

Reviewer: P12, L 13, each time new clinics are started? Sounds as if the hospital starts a new clinic, whereas I assume that a clinic enrolls in the program?
Answer: Thanks, wording was changed.

Reviewer: From now on I will not mention things anymore that strike me in the English, but suggest that a native speaker read the text.
Answer: Thanks, we had the paper reviewed by a native English speaker. Changes due to this review are not highlighted.

Reviewer: P14, L.4 What is meant by sensitive here?
Answer: Preference-sensitive

Reviewer: Note that DA and EBpDA are used interchangeably.
Answer: Thanks. Hopefully EbPDA used throughout now.

Reviewer: P15, L1: what is meant by this sentence: "As for the clinician training, etc." Is this something else than the decision coaching?
Answer: it is meant: "Comparable to how it works for the physician training". But we deleted this part of the sentence since it is not needed here. But indeed: the physician training is a bit different from the HCP training in decision coaching, which we try to specify in the text now.

Reviewer: This entire section is long and informal. Some info is redundant, but then the content of the individual coaching sessions is missing. Or is that the clinician training in line 1?
Answer: The training of "decision coaches" is similar to the physician training. However, it focuses more on the administration and discussion of a decision aid together with the patient. We tried to clarify this in the text. Overall we also tried hard to streamline the text where possible/necessary, taking your suggestions into account.

Reviewer: Of the PICS I would be interested in the number of items per scale, and the total scale scores/ranges.
Answer: Thank you, we added information on the number of items per scale and instrument characteristics and added Table 2 to provide more detail.

Reviewer: Further, for the PICS nothing is said about before and after the intervention, but I assume this holds, similarly to the Mappin? Ideally referring to the same consultation as that of the Mappin? (later I read more on this, but that is confusing, perhaps reordering of the text would help?)
Answer: The PICS is part of the **patient questionnaire**, which is measured at three points in time throughout the project (T0, T1, T2). Mappin'SDM is an observer-instrument and is measured only twice, at T0 and T1 (see figure 3 and new Table 2). We will clarify that PICS is part of the patient questionnaire and provide more detail on outcome measures in table 2. PICS and Mappin'SDM are not measured in the same sample.

Reviewer: I wondered why the Collaborate is used and not the SDM-Q9, since the latter is the most widely used (providing lots of comparison data) and the Collaborate has extreme ceiling effects (necessitating so called Top scores), and reflects the effort of the clinician as perceived by the patient, and thereby is more a satisfaction measure ("the SDM was not very good, but at least the doc did her best….").
Answer: Primary outcome measurement is provided by PICS and MAPPIN'SDM from different perspectives (PICS=patient perspective, Mappin'SDM=oberserver rating of specific patient-physician interaction). The German PICS questionnaire was used since our outcome measurement is not directly related to one specific decision interaction between a patient and a physician but rather to a bundle of interactions that the patient encountered throughout his/her last hospital stay. In addition, we preferred to use PICS since the patient more broadly is asked to reflect on his/her own perception of and participation in interactions. The ColloaboRATE was administered only as an additional (secondary) outcome measure to patients to be able to compare results to other international studies increasingly using it. But we totally agree with you regarding its potential limitations. The SDM-Q9 was not used since it directly refers to one specific decision interaction and it also is limited in our view given that it suffers from the same limitation you mention for CollaboRATE above: it reflects the effort of the clinician as perceived by the patient.

Reviewer: Top page 17: is this an informal way of presenting the power? I would like to have seen the sample size calculation here, to account for the N=16.000. (see further my comments below)
Answer: Since the information is confusing/superfluent here, we took it out and present it together with other information in the section "Sample Size Calculation" now.

Reviewer comment: As to the design, why a pre-post without a control group, why not an interrupted time series, or a stepped wedge design. Particularly looking at figure 2, that almost looks like a picture of a stepped wedge design! With such large numbers of clinicians to be trained, and multiple interventions, at least one of these alternative designs would have been possible and made a more

compelling case. With the current attention to SDM, one can now not know whether the intervention worked or this was simply an effect of time? Please argue.
Answer:
Thank you for sharing these thoughts with us. We had similar discussions before designing the project. In addition to what we previously stated:

1. Due to the size of our target intervention unit – an entire University hospital - a comparative study randomizing comparable hospitals was neither feasible nor affordable in this project / given the budgetary constraints.

2. We are not convinced that interrupted time series or a stepped wedge design within the hospital would have been adequate in this study. The main problem is that we assumed to have rather strong variations in outcome measures between the clinical departments at baseline (which was confirmed in our baseline assessment). This might be due to the nature of clinical questions, clinician's (typical) attitudes in the different disciplines, the different patient populations, etc. For example, the department of general surgery had the best PICS levels at baseline. All our observational MAPPIN' results, field notes, document analyses (e.g. number of participants in trainings) indicate strongly that these results, however, might be biased for several reasons. A stepped wedge-design requires several assumptions to be made, e.g. normal distribution of outcome variables and independence of effect measurement from timing of measurement, which did not appear to be given with our program intervention and in our setting for the reasons described above. Comparing individual departments simply before and after the intervention, however, might still give a fairly realistic idea of the program's success.

   An interrupted time series needs the intervention to be done in a clearly defined and rather short period. As the 4 different components start at different points in time (e.g. decision coaching starts after completion of EbPDA development in a department) it would have been difficult to define these points in time. As we measure three times, at T0, at T1 – directly after finishing all interventions in a department, and at T2 shortly before the project ends, we will probably have a very simple form of interrupted time series in this study.

In this study, we apply a very simple before-after test, knowing about all its limitations. We have discussed the study design prospectively with methodology-experts in the German Institute of Quality and Efficiency in Health Care (IQWiG; Dr. Sauerland and Dr. Lange). They both agreed to the chosen design. We are fully aware, that we will not measure real effects with this design. However, it will give us some information if we e.g. observe changes from baseline in some departments and none in others. The whole design should probably be regarded as a kind of benchmarking-effort comparing intra-department measurements/changes across departments (combined with insights from process evaluation).

Reviewer: P17, L. 12 what is SDM treatment?
Answer: Thanks. changed into "SDM-based treatment"

Reviewer: Power calculation
L14: 9 items ranging from 0-4 and a satisfactory score is 1.5? Is that an error? Or are this one, as well as the PICS, divided by the number of items, leading to a total of 0-4 (and thus 1-4 for PICS)? This would explain the sample size calculation for the PICS as well. Please clarify.
Answer: This section is not about the PICS but about the MAPPIN'SDM assessment. While PICS is a survey instrument answered by patients, the MAPPIN'SDM assessment is rated by independent raters. We clarified this in the new table 2 now. From our previous studies using the MAPPIN'SDM observer scale, we can tell that this score or higher indicates a clinically relevant level of SDM. In addition, 1.5 marks the transition from "1=minimal attempt of particular SDM skill" to "2=basic competence of particular SDM skill" according to the rater manual.

Reviewer: L19: "a sample size of about 40 for each group/department at each measurement (assuming a power of 80% and a level of significance of 5%, (one-sided, assuming positive effect)." How many clinics/ groups are there, for I assume you calculate a total sample, and divide by that to

get to 40? Unless you expect an effect of nesting (intraclass correlation, with some clinics already doing better than others) and you need to correct for this design effect? Or is 40 needed in total, as it seems from the statement two lines below, that an effect can be seen at the individual clinic level? That seems highly unlikely?
Answer: Thanks for this. There are 27 clinical departments at UKSH and involved in the study and in the Patient questionnaire/PICS evaluation. We calculated the sample size per unit per measurement time assuming a before/after comparative study design. One unit equals one UKSH department (=40 patients analyzed, about 60 sampled). To have samples bigger than 40 from the majority of departments (some are too small and it would take more than a year to get a sample size >40) it would need us to ideally have returned questionnaires from 1080 patients overall per measurement time, which would correspond to an overall sample size of 1600 patients and assuming a response rate between of about 60-70 % of the target population at each point in time (T1, T2, T3).
The question about nesting effects is a very good one. Although we assume that there might be some nesting/intraclass correlation at the campus evaluation level, these effects might not be relevant for our analyses since we do not compare departments against each other or intervention departments against control departments but limit ourselves to compare SDM levels at one department or the entire campus after the intervention to SDM levels before the intervention.

Reviewer: And then below it is said that “These numbers will allow to measure significant differences in the primary endpoint not only at the campus but also at the individual department level (at least in the larger clinics).” Are there clinics that are much larger than 40? How large would they need to be to assess an effect at individual clinic level?
Answer: This means that we will have returned questionnaires from about 1000-1100 patients at each time point (T1, T2 and T3) and on average 40 returned patient questionnaires per clinical department. For larger clinical departments we might have larger numbers than 40 and for smaller ones smaller. 40 is just the minimum number of returned patient questionnaires per clinical department to be able to run this subgroup-analysis.

Reviewer: It would help me if this section on the power were rewritten to make it clear what the total N needs to be, and whether this can be simply the sum of all the clinics.
And one sided testing is unusual, even with hypotheses (most research assumes one direction)! What are the numbers with two-sided?
Answer: Numbers with a two-sided test are higher, 50 per group. We are not convinced that a two-sided test is necessary. If power is calculated to be 80% in a one-sided test and if it is quite clear, that an intervention will nor reduce SDM, power should be sufficient to find a difference, even if calculates one-sided. We revised the sections on sample size calculation in the manuscript.

Reviewer: Why is the Mappin used if nothing is to be shown with it? I would expect it to be more sensitive to changes, so why not given an estimate of the power with the numbers you are expecting to obtain?
Answer: Only 7 of these clinical departments will be part of the Mappin analysis. Actually we do want to show that after the SDM program is completed, more than 80% will have a score higher than 1.5. So, we defined a response criterion here. We tried to clarify in text and added new Table 2 for clarification. In general, we considered the observed SDM-level of physicians/patients in their encounters (MAPPIN’SDM O’dyad) as a surrogate and perceived involvement in care (PICS) by patients as patient-important outcome. Therefore, we decided to choose PICS as the primary outcome for power calculation. For Mappin’SDM, sample size is given by the number of physicians in the included clinical departments. This is also further specified in the text now.

Reviewer: Final page
P19. I suggest deleting the formative assessment altogether, since as presented now, it does not provide sufficient information to the reader, and may not be too relevant in this respect? Or else, extend this section.
Answer: We extended this section since it is indeed an important part of our project.

Reviewer: I am not sure that I understand the role of the Implementation section. Is this meant as Process Evaluation, in addition to the other evaluations just described? Then this should be made clear, and the instruments described. If it is meant as part of the intervention, it should be moved up.
Answer: This section is taken out, the section on process evaluation was extended. Thanks, we very much appreciate making this point.

Reviewer: I am not familiar with the CFIR so cannot judge Table 1
Answer: see explanation above. And we extended on this in the main manuscript as well. We hope that the use and usefulness of the CFIR table is clearer now. We consider CFIR a very important instrument in our project planning and implementation.

Reviewer: The figures are too small for the readers. **Answer:** will be extended. Thanks.