

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)	Change in staff perspectives on indwelling urinary catheter use after implementation of an intervention bundle in seven Swiss acute care hospitals: Results of a before/after survey study
AUTHORS	Niederhauser, Andrea; Züllig, Stephanie; Marschall, Jonas; Schweiger, Alexander; John, Gregor; Kuster, Stefan; Schwappach, David; Safe urinary catheterization collaboration group, progress!

VERSION 1 – REVIEW

REVIEWER	BJ Laan, SE Geerlings Amsterdam UMC, University of Amsterdam, Netherlands
REVIEW RETURNED	12-Feb-2019

GENERAL COMMENTS	<p>This is an important subject with an evaluation of staff perspectives towards urinary catheter use, and a good response rate with many staff members participating. However, we think the most important endpoint is not increasing knowledge or perception, but the use of this increased knowledge and perception during practice. Could it just be a learning effect without a change of practice? Does the authors have CA-UTI rates of both periods?</p> <p>Abstract – Conclusion: This is a hard statement. In my opinion the knowledge of catheter-associated risks and beliefs regarding restrictive catheter use increased slightly. Therefore, I suggest a milder statement.</p> <p>Methods – Setting: The authors stated that each hospital could decide which units participated in the project. This could introduced bias, because units who are willing to change would participate more often, this point about the external validation of the results must be mentioned into the Discussion part. Furthermore, could the author tell which units participated? Surgical or internal medicine?</p> <p>Methods – Questionnaire: Is this a validated questionnaire? In other words, has it also been tested in another group?</p> <p>Methods – Data analysis: I am not a statistician, but in the analysis of a Likert scale a mean and SD are showed. This makes some sense, but usually in this kind of data there is no normal distribution. Therefore, I would encourage the use of mode and/or seek advise of a statistician if the current approach is permitted in this case.</p> <p>Methods – Data analysis: Probably there is no different outcome when using paired analysis, such as McNemar test instead of Chi-square test, but the data is about staff perception of the same units.</p>
-------------------------	---

	<p>So, paired analysis makes sense.</p> <p>Results – Response rate: Although the response rate is high. I would like to know which staff did not participated. Was it mainly physicians?</p> <p>Results – Response rate and study sample: There is a difference in age, profession and work unit between the groups. Could this introduced bias, and did the authors made corrections for this difference?</p> <p>Discussion: Can the authors start the Discussion with their main findings?</p> <p>Discussion: It is interesting that the knowledge scores between participants with and without training did not change. If I am correct, there were no other interventions that increase knowledge for healthcare workers (HCWs) than the staff training, expect for the dissemination of an indication list. So, could you stated that the HCWs that did not participate in the staff training, are a kind of control group for the knowledge section? Or is a Hawthorne effect present? Are there other factors which could have increased the overall knowledge?</p> <p>Conclusion: What does an increased knowledge score of 0.6 mean? The effect size is 0.29, small to moderate, but I miss the clinical impact. Perception and behaviour seems to be increased mildly, and the self-reported responsibilities changed only slightly (not significant?). So, I miss a hard measurement of true change of behaviour. If the authors cannot provide these results, in my opinion they should state that this is only a first step in the evaluation of the intervention bundle. Hopefully we will see a decrease of infections rates in a future publication.</p>
--	--

REVIEWER	<p>Mary Jo Knobloch William S. Middleton Memorial Veterans Hospital, Madison, Wisconsin, USA University of Wisconsin-Madison, Madison, WI, USA</p>
REVIEW RETURNED	27-Feb-2019

GENERAL COMMENTS	<p>This is a well-written, easy to read paper. I like the use of the self-generated ID codes used, but thought the authors should further explain the technique they used for having participants generate codes.</p> <p>The authors should address what study design may have been beneficial instead of using the before/after cross-sectional, especially due to the fact that they had a large population. This would help others anticipating further research in this area.</p> <p>I wondered why a focus group or interviews could not be done, since this would provide a mixed-methods way to describe perceptions and perspectives.</p> <p>It appears that print versions of the survey were used. There was no explanation of how this worked, versus electronic. Again, further detail would be good for those seeking to replicate methods.</p>
-------------------------	--

	<p>The authors mention sustainability - but offer no ideas on how to measure this over an extended period of time to determine sustained best practices. What about an interrupted time series approach?</p> <p>I felt this paper lacked the "so what" factor. If I'm going to implement a new evidence-based bundle, what does this paper tell me? The authors need to address why this is important to know, especially since sustainability is always an issue.</p> <p>I think this paper should be published, but only if authors can tell the reader what is new about their findings compared to what is known in the literature, and how this has implications for the reduction of infections.</p>
--	--

REVIEWER	<p>Dr Krista Wooller The Department of Medicine, The University of Ottawa, The Ottawa Hospital, Ottawa, Ontario, Canada</p>
REVIEW RETURNED	08-Mar-2019

GENERAL COMMENTS	<p>Thank you for the chance to review this interesting article which attempts to evaluate changes in staff knowledge and behavior related to indwelling urinary catheters after implementation of a IUC bundle as part of a quality improvement project.</p> <p>The authors should be commended for trying to measure changes in perception of culture and behavior via survey. The survey number and response rate is fairly reasonable. They write clearly and do demonstrate an increase in knowledge and improvement in the scales developed to measure the cultural aspects of the care environment. The discussion accurately points out the limitations of the study - there are very important limitation given we aren't sure who answered the survey and who didn't - however, as the authors point out they are unable to get that information so it is important they discuss it as they have. My suggestions for improvement for this paper are as follows:</p> <ul style="list-style-type: none"> - it would strengthen the ability to interpret the results if we knew more about the "training sessions" offered. Eg. how many hospitals actually implemented the practical training sessions. What did the theoretical training sessions look like? Can you tell us how many staff were eligible for training sessions vs how many actually completed them? - can you clarify exactly who were the eligible staff? The number is 3,245 as reported in the results. How many staff work at the pilot hospitals? - It would be useful to describe more about the validity of the survey tool (mentioned on line 35, page 6). How do you know this survey actually measures what you think it measures? How well are the "cultural factors" measured in a survey tool - do you have a good reference for people using surveys like this in the manner you have used it? - can you clarify how frequent vs infrequent users were defined for physicians? ie/ as a physician if you have patients with catheters but don't actually place them yourself are you a frequent or infrequent user? - unfortunately, reference 37 is another language and I can't interpret where the "importance of effect" classification came from.
-------------------------	--

	<p>I'm unfortunately not familiar with this. It seems to me just looking at the effect size by comparing pre to post that it is a rather small effect that we are seeing here. I agree it is statistically significant but I'm not sure it is really meaningful??</p> <p>- the discussion on line 26-27, page 13 states that "Our results show that resource-intensive education sessions need not be the only way to convey knowledge and awareness about safety issues and good practices". But I'm not sure you can really say this as I'm not convinced that you demonstrate what the intensity of the education sessions really was? If I am reading your methods correctly, it seems like some of the resource intensive training was optional and its not clear how many people it reached.</p> <p>Thank you for the opportunity to comment on your interesting paper.</p>
--	---

REVIEWER	Hsueh-Wen Chang National Sun Yat-Sen University Taiwan, Republic of China
REVIEW RETURNED	08-May-2019

GENERAL COMMENTS	<p>Major:</p> <ol style="list-style-type: none"> The authors use t test for paired samples to compare change between time points. It seems a repeated design has been employed. However, the sample of 1,579 subjects in the baseline survey is not the same as the 1,527 subjects in the follow up survey. Table 1 shows the difference of the characteristics of between the two sample periods. A paired or repeated design should be based on the survey results of the same subject over time. The results of an analysis on 420 subjects with matched ID would be served as a valid design. The data of Self reported responsibilities concerning IUC management are categorical. Was the comparison of these data also based on repeated samples? This needs to be clarified, and statistical methods stated. <p>Minor:</p> <ol style="list-style-type: none"> P values should be presented in Table 1. <p>Knowledge:</p> <ol style="list-style-type: none"> The mean number of correctly answered knowledge increase significantly from T0 to T1. Given a total 15 items, as the mean was 10.4 on T0 and 11.0 on T1, the difference is hardly 1 item. Furthermore, the effect sizes for all significant differences are under 0.5, only small to medium effect. But with such a large sample size, it shows $P < 0.001$. First paragraph, However, their knowledge scores at follow up still remained lower compared to the scores of physicians, staff members with managerial function and infrequent users. Knowledge scores increased more for participants with a matched ID compared to unmatched participants. P values
-------------------------	---

	<p>should be provided for these two statements comparing scores between groups</p> <p>4. Second paragraph, Among all matched respondents, 102 (25.3%) indicated having participated in both theoretical and practical training; 130 (32.2%) respondents had participated in either theoretical or practical training and 172 (42.6%) respondents had participated in no training. There add up to 404 subjects, not 420 shown in Table 1. Also, the results of the mixed model shows only scores of 402 subjects were used if subject was treated as a random factor and time was treated as a fixed factor This need an explanation. Perc</p> <p>option of practices and culture</p> <p>5. Similar to comment 2 of Knowledge section, the magnitude of increase of mean score is small For example, for over a ll sample , the mean score increased from 5.3 to 5.5., effect size 0.31, small to medium</p> <p>Responsibilities</p> <p>Responsibilities</p> <p>6. The he “****” in in FFig 1.ig 1. SShould should be be explainedexplained..</p> <p>7. It is good if the authors can provide P values for Alt is good if the authors can provide P values for Appeppendix 4ndix 4..</p> <p>8. “Positive changes in mean scores were observed for all three constructs Positive changes in mean scores were observed for all three constructs (attitudes, subjective norms and perceived behavioral control). They were (attitudes, subjective norms and perceived behavioral control). They were particularly strong for iparticularly strong for items assessing perceived social expectations to use tems assessing perceived social expectations to use catheters restrictively. The positive trends could be observed for professional catheters restrictively. The positive trends could be observed for professional group, managerial function and frequency of catheter placement.group, managerial function and frequency of catheter placement.” P values P values should be provide for these statements. should be provide for these statements.</p> <p>9. “There was no notable difference between the results for matcThere was no notable difference between the results for matched and hed and unmatched participants.unmatched participants.” Is Is ththis basedis based on difference of incren difference of increeas ease between two between two groups? A P values also should be provided.groups? A P values also should be provided.</p>
--	--

VERSION 1 – AUTHOR RESPONSE

#	Reviewer comments	Authors response and changes to the manuscript
	Editorial Requests:	
1	Please include the study setting in the title.	Done
2	Abstract: please edit the second sentence of the participants section as it does not make sense currently.	Done, please see P2, L54-56
3	Please provide an English translation of the questionnaire as an additional file.	Done, we've added the questionnaire as appendix 2
	Reviewer 1	
4	<p>This is an important subject with an evaluation of staff perspectives towards urinary catheter use, and a good response rate with many staff members participating.</p> <p>However, we think the most important endpoint is not increasing knowledge or perception, but the use of this increased knowledge and perception during practice. Could it just be a learning effect without a change of practice? Does the authors have CAUTI rates of both periods?</p>	<p>Thank you for your positive comments. We agree that with our study design, we cannot assess if changes in knowledge or perception truly led to a change in practice. Direct observations of clinical practice or the combination of our survey data with health outcome measures (such as CAUTI) may have been beneficial.</p> <p>During our QI program, data on catheter utilization and catheter-associated complications was collected in all hospitals before and after the intervention. This surveillance data will be published elsewhere (paper submitted for publication) (P6, L169-170). We observed a reduction in catheter utilization rates after the intervention. We've added this information to the manuscript and discuss what this observation could mean in combination with the staff survey data (P16-17, L431-452). We will consider conducting further analyses to relate changes in specific clinical outcomes with specific changes in staff perceptions, for example subjective norms. (P18, L488-490).</p>
5	<p align="center">Abstract – Conclusion:</p> <p>This is a hard statement. In my opinion the knowledge of catheter-associated risks and beliefs regarding restrictive catheter use increased slightly. Therefore, I suggest a milder statement.</p>	<p>Several reviewers have stated that observed effects on staff perspectives are rather small. We agree that this needs to be discussed more extensively in the manuscript. We have made substantial revisions to the entire discussion section (P14-18).</p> <p>We also edited the statement in the abstract and conclusion section (P2, L69-72 and P19, L523-525).</p>
6	<p align="center">Methods – Setting:</p> <p>The authors stated that each hospital could</p>	<p>Yes, surgical and internal medicine wards participated. When a ward participated, all units</p>

decide which units participated in the project. This could introduced bias, because units who are willing to change would participate more	composing the ward participated in the project. We added this information on P6, L162-163. We also added the potential threat to external
--	---

P2/6

	often, this point about the external validation of the results must be mentioned into the Discussion part. Furthermore, could the author tell which units participated? Surgical or internal medicine?	validity in the limitations section P19, L 518-520.
7	Methods – Questionnaire: Is this a validated questionnaire? In other words, has it also been tested in another group?	No, the questionnaire has not been tested in another group; it was developed specifically for this study and pretested among 43 physicians and nurses. We added a precision on P6, L184.
8	Methods – Data analysis: I am not a statistician, but in the analysis of a Likert scale a mean and SD are showed. This makes some sense, but usually in this kind of data there is no normal distribution. Therefore, I would encourage the use of mode and/or seek advise of a statistician if the current approach is permitted in this case.	We used mean/SD because it offers more variance to measure change over time and our sample size was large enough. We chose a 7-point scale to be able to treat the items as continuous variables.
9	Methods – Data analysis: Probably there is no different outcome when using paired analysis, such as McNemar test instead of Chi-square test, but the data is about staff perception of the same units. So, paired analysis makes sense.	We assume that this comment refers to the analysis of differences in sample composition between the two survey periods. We cannot assume that the subjects in the baseline survey are the same as the subjects in the follow-up survey, even if the survey was administered on the same units. This is underlined by the fact that we were only able to match the self-generated ID for 420 respondents in the two surveys, which represents 27.5% of all possible matches. We therefore use unpaired analysis to determine differences between samples and changes between time points. We added a precision on P8, L232-235. However, we also performed paired analysis on the sub-sample of the 420 respondents that could be matched based on their self-generated ID (P8, L247-252).
10	Results – Response rate: Although the response rate is high. I would like to know which staff did not participated. Was it mainly physicians?	Unfortunately, we have no information on non-respondents.

11	Results – Response rate and study sample: There is a difference in age, profession and work unit between the groups. Could this introduced bias, and did the authors made corrections for this difference?	We did not make corrections for this difference. The advantage of our design is that we can show the change on the individual participant level (same individuals participating twice), and change in the overall sample (group level) (see additions in the manuscript on P14, L339-347). On the group level, we cannot differentiate if sample differences were due to non-responses or staff fluctuation (P19, L516).
12	Discussion: Can the authors start the Discussion with their main findings?	Okay, please see P14, L337-347.
13	Discussion: It is interesting that the knowledge scores between participants with and without training did not change. If I am correct, there were no other interventions that increase knowledge for healthcare workers (HCWs) than the staff training, expect for the dissemination	Yes, when only looking at participants that could be matched, we can state that those who did not participate in trainings act as a control group for the knowledge section. Other factors that could have increased knowledge were the presence of champions, the dissemination of

P3/6

	of an indication list. So, could you stated that the HCWs that did not participate in the staff training, are a kind of control group for the knowledge section? Or is a Hawthorne effect present? Are there other factors which could have increased the overall knowledge?	the indication list or other efforts to raise awareness. In the revised discussion section, we elaborate this point on P14, L350-356. In response to reviewer comment #23, we also added some more details on the content of the trainings (P5-6, L145-158).
14	Conclusion: What does an increased knowledge score of 0.6 mean? The effect size is 0.29, small to moderate, but I miss the clinical impact. Perception and behaviour seems to be increased mildly, and the self-reported responsibilities changed only slightly (not significant?). So, I miss a hard measurement of true change of behaviour. If the authors cannot provide these results, in my opinion they should state that this is only a first step in the evaluation of the intervention bundle. Hopefully we will see a decrease of infections rates in a future publication.	Please see our response to comment #4. In the limitations section we now state that survey data and surveillance data should be linked in secondary data analysis, and we mention direct observations as an alternative method for measuring change of behavior (P18, L488-495).
	Reviewer 2	
15	This is a well-written, easy to read paper. I like the use of the self-generated ID codes used, but thought the authors should further explain the technique they used for having participants generate codes.	Thank you for your positive feedback. We added a more detailed description of the technique for generating self-generated IDs on P7, L211-217. In addition, we included the entire questionnaire as appendix 2. It includes an example of how to generate the ID.

16	The authors should address what study design may have been beneficial instead of using the before/after cross-sectional, especially due to the fact that they had a large population. This would help others anticipating further research in this area.	We edited the limitations sections and have included several suggestions for alternative study designs, notably conducting direct observation to assess compliance with catheter insertion protocols, and using a stepped-wedge design as an alternative to the single-group design to better assess the contribution of the intervention bundle to observed effects P18, L490-501.
17	I wondered why a focus group or interviews could not be done, since this would provide a mixed-methods way to describe perceptions and perspectives.	We agree that using a mixed-methods approach could have added more in-depth insights into staff perceptions. Unfortunately, our study design did not include an additional qualitative method. We added a section in the limitations sections P18, L501-504.
18	It appears that print versions of the survey were used. There was no explanation of how this worked, versus electronic. Again, further detail would be good for those seeking to replicate methods.	We added more details on the reasons why we used paper surveys in the limitations section, P19, L507-512.
19	The authors mention sustainability - but offer no ideas on how to measure this over an extended period of time to determine sustained best practices. What about an interrupted time series approach?	We mention the importance to sustain effects over time in the discussion section (P16, L425-430) and have added as per your suggestion more ideas on how to measure this in the limitations section (P19, L506).
20	I felt this paper lacked the "so what" factor. If I'm going to implement a new evidence-based bundle, what does this paper tell me? The authors need to address why this is important	We have substantially revised the entire discussion section and hope that we can demonstrate the additions that our paper brings to the existing literature (P14-18).
	to know, especially since sustainability is always an issue. I think this paper should be published, but only if authors can tell the reader what is new about their findings compared to what is known in the literature, and how this has implications for the reduction of infections.	
	Reviewer 3	
22	Thank you for the chance to review this interesting article which attempts to evaluate changes in staff knowledge and behavior related to indwelling urinary catheters after implementation of a IUC bundle as part of a quality improvement project. The authors should be commended for trying to measure changes in perception of culture and	Thank you for your positive feedback.

	<p>behavior via survey. The survey number and response rate is fairly reasonable. They write clearly and do demonstrate an increase in knowledge and improvement in the scales developed to measure the cultural aspects of the care environment. The discussion accurately points out the limitations of the study - there are very important limitation given we aren't sure who answered the survey and who didn't - however, as the authors point out they are unable to get that information so it is important they discuss it as they have. My suggestions for improvement for this paper are as follows:</p>	
23	<p>- it would strengthen the ability to interpret the results if we knew more about the "training sessions" offered. Eg. how many hospitals actually implemented the practical training sessions. What did the theoretical training sessions look like? Can you tell us how many staff were eligible for training sessions vs how many actually completed them?</p>	<p>We added more details on the training sessions on P5-6, L145-158. Unfortunately, it proved too difficult to elicit the exact number of participants in theoretical and practical trainings from each hospital.</p>
24	<p>- can you clarify exactly who were the eligible staff? The number is 3,245 as reported in the results. How many staff work at the pilot hospitals?</p>	<p>All staff members invited to participate in the survey worked at one of the seven pilot hospitals. All members of nursing and medical staff, except staff members not involved in direct patient care, healthcare workers in education and affiliated physicians were eligible to participate (P6, L180). Local project teams were responsible to identify and distribute the surveys to all eligible staff members (P8, L221-223).</p>
25	<p>- It would be useful to describe more about the validity of the survey tool (mentioned on line 35, page 6). How do you know this survey actually measures what you think it measures? How well are the "cultural factors" measured in a survey tool - do you have a good reference for people using surveys like this in the manner you have used it?</p>	<p>We ensured validity of the questionnaire by using psychological theory (theory of planned behavior) to model intention for behavior change and by pretesting the instrument among 43 physicians and nurses. Nevertheless, the moderate effects can raise the question if the items in our questionnaire adequately measure aspects of knowledge and perceptions required to reduce IUC use (content validity) (P17, L435-441).</p>

P5/6

26	<p>- can you clarify how frequent vs infrequent users were defined for physicians? ie/ as a physician if you have patients with catheters</p>	<p>Frequent user refers to the self-reported frequency of placing a catheter. We clarified this aspect on P8, L244-245.</p>
----	---	---

	but don't actually place them yourself are you a frequent or infrequent user?	
27	- unfortunately, reference 37 is another language and I can't interpret where the "importance of effect" classification came from. I'm unfortunately not familiar with this. It seems to me just looking at the effect size by comparing pre to post that it is a rather small effect that we are seeing here. I agree it is statistically significant but I'm not sure it is really meaningful??	The reference (now #38) has been replaced by a reference in English. Several reviewers have stated that observed effects on staff perspectives are rather small. We agree that this needs to be discussed more extensively in the manuscript. We have therefore made some substantial revisions to the entire discussion section (P14-18).
28	- the discussion on line 26-27, page 13 states that "Our results show that resource-intensive education sessions need not be the only way to convey knowledge and awareness about safety issues and good practices". But I'm not sure you can really say this as I'm not convinced that you demonstrate what the intensity of the education sessions really was? If I am reading your methods correctly, it seems like some of the resource intensive training was optional and its not clear how many people it reached.	We found it interesting to see that knowledge scores did not differ between participants with and without training. This indicates that other factors, such as the dissemination of the indication list or the presence of champions may have contributed to the observed effects. We have clarified our statement on P14, L350-356. We also added more details on the training sessions on P5-6, L145-158.
	Reviewer 4 (statistical comments)	
29	Major: 1. The authors use t-test for paired samples to compare change between time points. It seems a repeated design has been employed. However, the sample of 1,579 subjects in the baseline survey is not the same as the 1,527 subjects in the follow-up survey. Table 1 shows the difference of the characteristics of between the two sample periods. A paired or repeated design should be based on the survey results of the same subject over time. The results of analyses on 420 Subjects with matched ID would be served as a valid design. 2. The data of Self-reported responsibilities concerning IUC management are categorical. Was the comparison of these data also based on repeated samples? This needs to be clarified, and statistical methods stated.	We only used t-test for paired samples to compare change between time points for participants that could be matched based on the 8-digit ID (n=420). To determine changes between time points for the overall sample, we used unpaired analysis. So for example we used chi2 test to test for differences in sample characteristics (table 1) and self-reported responsibilities (figure 1) at T0 and T1 and we used t-test for independent samples to test for differences in mean scale scores between T0 and T1 (tables 2-4). We edited the methods section to clarify the different types of analyses used for the overall sample and the sub-sample of matched participants. P8, L231-235 and L247-252.
30	Minor: 1. P values should be presented in Table 1.	We have added p-values in table 1.
31	Knowledge: 2. The mean number of correctly answered knowledge increase significantly from T0 to T1. Given a total 15 items, as the mean was 10.4 on T0 and 11.0 on T1, the difference is hardly 1 item. Furthermore, the effect sizes for all	Several reviewers have stated that observed effects on staff perspectives are rather small. We agree that this needs to be discussed more extensively in the manuscript. We have made some substantial revisions to the entire discussion section (P14-18).

	significant differences are under 0.5, only small to medium effect. But with such a large sample size, it shows $P < 0.001$ nevertheless.	
32	3. First paragraph, "However, their knowledge scores at follow-up still remained lower compared to the scores of physicians, staff members with managerial function and	You are right; we actually did not compare scores between groups at follow-up but only changes in scores within groups over time. This is why we deleted this sentence on P11
	infrequent users. Knowledge scores increased more for participants with a matched ID compared to unmatched participants." P values should be provided for these two statements comparing scores between groups.	
33	4. Second paragraph, "Among all matched respondents, 102 (25.3%) indicated having participated in both theoretical and practical training; 130 (32.2%) respondents had participated in either theoretical or practical training and 172 (42.6%) respondents had participated in no training." There add up to 404 subjects, not 420 shown in Table 1. Also, the results of the mixed model shows only scores of 402 subjects were used if "subject" was treated as a random factor and "time" was treated as a fixed factor. This need an explanation.	n=16 subjects had missing values on the items assessing participation in training. We added this to the text (P11, L 296).
34	Perception of practices and culture 5. Similar to comment 2 of Knowledge section, the magnetite of increase of mean score is small. For example, for overall sample, the mean score increased from 5.3 to 5.5., effect size 0.31, small to medium	Several reviewers have stated that observed effects on staff perspectives are rather small. We agree that this needs to be discussed more extensively in the manuscript. We have made some substantial revisions to the entire discussion section (P14-18).
35	Responsibilities 6. The "*" in Fig 1. Should be explained.	The * indicates significant changes between time periods at $p < 0.05$. This is explained in the figure legends, please see P2, L542
36	7. It is good if the authors can provide P values for Appendix 4.	As requested by the reviewer, we performed an independent samples t-test for each item and added the p-values to appendix 5. Please note, for completeness we also performed t-tests for each item in appendix 4.
37	8. "Positive changes in mean scores were observed for all three constructs (attitudes, subjective norms and perceived behavioral control). They were particularly strong for items assessing perceived social expectations to use catheters restrictively. The positive trends could be observed for professional group, managerial function and frequency of catheter	We added p-values for appendix 5

	placement." P values should be provide for these statements.	
38	9. "There was no notable difference between the results for matched and unmatched participants." Is this based on difference of increase between two groups? A P values also should be provided.	Yes, the sentence is based on the difference in increase between the two groups. We performed difference-in-difference analyses to examine if the increase between matched and unmatched participants over time. This was done in order to examine if it is adequate to analyze data on a group-level data (unmatched participants) to evaluate changes between two time points. We clarify this in several places in the manuscript: - methods section (P9, L254-257) - new chapter in the results section (P13-14, L330-335) - discussion of findings (P14, L339-347)

VERSION 2 – REVIEW

REVIEWER	Mary Jo Knobloch University of Wisconsin Madison
REVIEW RETURNED	27-Aug-2019
GENERAL COMMENTS	It appears that authors have done a much better job of explaining details, addressing sustainability and the "so what" factor that was missing in the first iteration.