PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Capturing intraoperative process deviations using a direct observational approach: The glitch method
AUTHORS	Morgan, Lauren; Robertson, Eleanor; Hadi, Mohammed; Catchpole, Ken; Pickering, Sharon; New, Steve; Collins, Gary; McCulloch, Peter

VERSION 1 - REVIEW

REVIEWER	Alexander F. Arriaga Brigham and Women's Hospital Ariadne Labs Harvard School of Public Health I have no conflicts of interest or disclosures relevant to this review.
REVIEW RETURNED	19-Aug-2013

GENERAL COMMENTS	This is a well-designed study and well-written manuscript presented by authors who are widely-published leaders in the field of surgical safety.[1-3] Their presentation of the "glitch method" helps us come to a better understanding of intraoperative process deviations based on data obtained through real-time measurement of operating room behavior. In a multi-center study of over 400 operations, the authors found that, on average, there are fourteen "glitches" per operation (a "glitch" is defined by the authors as a "deviation from the recognized process with the potential to reduce quality or speed, including interruptions, omissions and changes"). The majority of glitches occurred within the first half of the operation. Distractions were the most common glitches observed in all cases (Figure 2). These findings only emphasize the value of interventions such as safety pauses and checklists, which have been shown to reduce morbidity and mortality[4-7] and are increasingly being used in the operating room. I have a few questions and comments for the authors: 1. It is interesting that the range of glitches for a given case was 0 to 83 per operation. Some of this range of course represents the length of the operation, but there may be lessons that can be gathered from this finding. In manuscript page 10, lines 42-55, the reader can appreciate that this range was 1-63 for elective orthopedic surgery, 1-35 for trauma orthopedics, 2-49 in elective vascular surgery, and 1-83 in elective plastic surgery. Hence, across specialties, some cases had less than 3 glitches total. Do the authors have any speculation, other than operative duration or

	institution, on why some cases had so few glitches and others had a much greater quantity? The only comment I see relating to this is in manuscript page 12, lines 50-55: "The rate of distractions is the greatest in plastics which relates anecdotally to the discursive and fluid nature of the teams involved." If the authors/field observers have any further insight on the range of glitches, it would be an interesting addition to either the discussion section of this work or the topic of a future manuscript. a. Can the authors clarify the total range of glitches for a given case? In the abstract, the range is stated as 0-83. On page 10 of the manuscript, it states that "The average total glitch count per operation was 14, range 1-83." I wasn't sure if the range cited in page 10 was referring to a subset of the data. 2. The authors observed that glitches seemed to vary more by institution than by specialty. They appropriately point out that this observation was limited by the fact that there were large differences in the case distribution in each site (Table 3) with only 1 site performing trauma orthopedics, one site performing plastic surgery, and one (or more[?]) site(s) performing vascular surgery (manuscript page 13, lines 7-11: "Elective orthopaedic and vascular surgical procedures were observed in multiple sites, providing an opportunity for inter-site rate comparison amongst teams performing the same types of surgery"). a. Can the authors clarify how many sites had vascular surgical cases? (re: comparison of Table 3 and manuscript page 13, lines 7-11). b. On page 13 of the manuscript, the authors comment on the variation of glitch rate by institution. The passage (manuscript page 13, lines 5-29) reads as if they tested each pair of sites for differences in glitch rates but only reported the pairs that were statistically significant. They did provide a combined p-value of the relationship between glitch rate and institution ("There was a statistically significant differences [sic] in mean glitch rates per operation between the four sites [$p<0.001$]"), and they did point out that they did not adjust for multiple testing. However, it would be helpful to the reader for it to be stated that (if the authors tested it) that there were no significant differences between sites A and C, sites A and D, sites A and E, sites B and E, sites C and D, and sites C and E. If the authors did not test the pairs I listed above, then what made them test only the pairs that they reported? I do not think they need to remove the interesting and believable finding that there are differences in glitch rates by institution. A simple sentence would suffice clarifying that the other pairs were tested but not significant (if they were indeed tested). 3. Comment: I agree with the authors that the Hawthorne effect is not a large limitation to the study (manuscript page 16, lines 47-59: "…the observers quickly became well known to the theatre staff and as such became 'part of the furniture'…). Our research group has had similar anecdotal findings during

our operating room field observation work.[8-11]

In summary, this is a well-prepared manuscript describing a method to capture intraoperative process deviations though structured field observation. It was a multi-institutional study with a relatively large sample size for this type of work. It is worthy of publication.

Alexander F. Arriaga, MD, MPH, ScD

Brigham and Women's Hospital

Ariadne Labs

Harvard School of Public Health

References:

1. Catchpole K, Mishra A, Handa A, McCulloch P. Teamwork and error in the operating room: analysis of skills and roles. Ann Surg 2008;247:699-706.

2. Mishra A, Catchpole K, McCulloch P. The Oxford NOTECHS System: reliability and validity of a tool for measuring teamwork behaviour in the operating theatre. Qual Saf Health Care 2009;18:104-8.

3. McCulloch P, Catchpole K. A three-dimensional model of error and safety in surgical health care microsystems. Rationale, development and initial testing. BMC Surg;11:23.

4. Haynes AB, Weiser TG, Berry WR, et al. A surgical safety checklist to reduce morbidity and mortality in a global population. N Engl J Med 2009;360:491-9.

5. de Vries EN, Prins HA, Crolla RM, et al. Effect of a comprehensive surgical safety system on patient outcomes. N Engl J Med 2010;363:1928-37.

6. Neily J, Mills PD, Young-Xu Y, et al. Association between implementation of a medical team training program and surgical mortality. JAMA 2010;304:1693-700.

7. van Klei WA, Hoff RG, van Aarnhem EE, et al. Effects of the introduction of the WHO "Surgical Safety Checklist" on in-hospital mortality: a cohort study. Ann Surg 2012;255:44-9.

8. Hu YY, Arriaga AF, Roth EM, et al. Protecting patients from an unsafe system: the etiology and recovery of intraoperative deviations in care. Ann Surg 2012;256:203-10.

	9. Hu YY, Peyre SE, Arriaga AF, Roth EM, Corso KA, Greenberg CC. War stories: a qualitative analysis of narrative teaching strategies in the operating room. Am J Surg 2011;203:63-8. 10. Hu YY, Arriaga AF, Peyre SE, Corso KA, Roth EM, Greenberg CC. Deconstructing intraoperative communication failures. J Surg Res 2012;177:37-42. 11. Hu YY, Peyre SE, Arriaga AF, et al. Postgame analysis: using video-based coaching for continuous professional development. J Am Coll Surg 2011;214:115-24.

REVIEWER	Dr Tobias Gauss Consultant Surgical Critical Care Unit Department of Anaesthesia and Critical Care Hôpital Beaujon HUPNVS, APHP, Clichy, France No conflict of interest to declare
REVIEW RETURNED	23-Aug-2013

THE STUDY	1) It appears not clear which interventions have been included and which not, see comment 7) and 10). 2) see indicated in comment 2) and 3) and 5) 3) see comments 6) and 9) 4) with regard to the consort checklist only the items that have been mentioned before
GENERAL COMMENTS	The authors have attempted to develop a method to screen, quantify and categorize events, so called glitches that disturb the surgical workflow and could ultimately lead to adverse events and complications. This effort is pertinent and welcome, since to my knowledge, no such tool exists so far. Although there is an abundant literature on the subject, it appears necessary to homogenize the description of perioperative events. The question is pertinent and the methodology appears adapted. Questions and comment in details: 1) P6, L 14, I suggest to start with a description of the development of the method and validation. This appears easier to follow for the reader. 2) P6, L53, the description of the development and validation of the method appears quite short. Although the reader can of course refer to the description in Morgan et al 2011, a resume of the main elements appears necessary: how long was the validation period, how many interventions, which specialty, how were the categories chosen, etc. It would also be interesting what the baseline glitch rate, as well as the baseline complication rate was.

	3) P6, L20, How many observers were there in total? Were the observers identical to the researchers that conceived the categories? If so they might be biased towards their own methodology. They same logic would apply if the researches also conducted the initial validation of the methodology described in the paper Morgan et al 2011. This seems to be the case P7, L8, as the observers also conducted the final review of the categorization of all glitches. A bias towards the conceived categories cannot be excluded. Was any independent person involved in the observation as well in the final review? This appears of importance to me, as this may question the external validity of the method. 4) P6, L 38: The authors state that technical skills and WHO checklist adherence were collected, but these data are not clearly provided later in the result section. Is non-compliance with the WHO-Checklist also a glitch? 5) P6, L58 and table 1: It appears to me that the glitch categories do not include glitches such as deviation from planned surgical strategy or any unexpected clinical events (allergy, unexpected bleeding, etc) or unexpected surgical difficulty, but maybe do I not understand the meaning of all categories. Was rescheduling considered as a glitch? Do the glitches only apply to intraoperative glitches or was the whole perioperative period included? In my opinion perioperative (pre-, intra- and postoperative) organisational difficulties and communication breakdowns can be very disturbing impacts on the workflow and patient safety. Were glitches with regard to the anaesthetic management taken into account? Personally I would suggest that not all suggested glitch-categories carry the same amount of risk for potential harm. Did the authors attempt or think about rating the categories and glitches? 6) P7, L3 Does the indicated kappa take into account that there were several raters involved? A kappa of 0,7 appears low under the assumption that the researchers themselves have used it and perform a joint review in the case of disagreement. It would be interesting to dispose of the kappa of the prospective data presented in this paper. 7) P7, L37 Why has visceral surgery not been included? Visceral surgery is one of the most common types of surgery. Including visceral surgery would probably increase the external validity. In this context, with regard to table 3, orthopaedics were observed in all five centres, but trauma, plastic and vascular, only at one site each. Why could not more observations be made at the other sites? 40% of interventions have been observed at one site with a disproportionate amount of observations made in orthopaedics. This could be a source of bias. (Minor remark, the numbers on duration is a bit cumbersome to read.

	8) P7, L39 It would be interesting to know more about the kind of surgery, the surgical difficulty, the experience of the operators, the percentage of emergency surgery and to dispose of some clinical data, such as median ASA. 9) P9, L8, if I understand it correctly, no power analysis to determine a required number of observations for an estimated precision is provided, which is all the more striking, since the authors dispose of preliminary data. 10) P 9, L 19, 429 operations over 16 months does not appear a lot to me. Why not more over a shorter period of time? Who made choice over which 11) P 10, L 45 and further on, The glitch rate per hour and specialty is an interesting concept and appears intuitive and useful as indicator. Nevertheless it would be interesting to see a link to the level of competence of the operators, skill of team, level of teamwork, surgical and clinical difficulty, etc. 12) It seems that the authors do not present any data that could link the glitches to measurable and pertinent clinical events that may have an impact on quality of care, patient safety or even complications. This was of course not the scope of the paper, but if this were the case this would be a pity given the important effort they have made. The work would be all the more interesting. Overall the questions is pertinent, the work and contribution to the field potentially useful, but a test of the external validity of the tool in other institutions appears appropriate before putting it into use.

REVIEWER	Douglas A. Wiegmann, PhD Department of Industrial and Systems Engineering University of Wisconsin-Madison USA
REVIEW RETURNED	26-Aug-2013

THE STUDY	What is the outcome of interest? Update the references. Over extend findings/implications
RESULTS & CONCLUSIONS	Authors fail to integrate findings to other studies Authors downplay previous research to over sell their methods/finding.
GENERAL COMMENTS	The goal of this research project was to develop a methodology for identifying “technical process imperfections during surgical procedures” in order to assess their impact on patient outcomes and evaluate the effectiveness of interventions. There are several positive aspects of this study. The first being the authors’ attempt to assess inter-rater reliability of the observational approach utilized. The authors were able to demonstrate acceptable levels of reliability across different types of raters (human factors vs. medical) and raters standing in different locations. This is somewhat conflicting with the results of previous research that indicates that an observer’s location in the operating room impacts the types of

	events that they observe. Research also suggests that observers with medical backgrounds tend to observe more clinically related events than human factors observers (who tend to observe more work system factors). Unfortunately, the authors do not address their findings in light of the pervious research. Another positive aspect of this research project is that the authors were able to create a generic system for documenting “glitches” across procedures. Previous methodologies have been procedure specific and researchers have had to significantly modify existing data collection tools to accommodate the nuances associated with individual surgical procedures. The ability for both researchers and practitioners to have access to a generic “off the shelf” data collection tool could be of great benefit for those who do not have the resources to develop their own data collection tool. However, the authors fail to discuss the limitations of using such a generic approach to data collection. Clearly there are differences that exist across surgical procedures (e.g., equipment utilized, personnel involved, and medications/anesthesia required, etc.). Understanding what one might sacrifice in terms of specificity from using a generic data collection method is very important. The authors have also proposed a novel way of analyzing events during individual surgical cases. The analysis of the cumulative proportion of “glitches” across the duration of a case provides a unique insight into the timing of “glitches” during a procedure. Unfortunately, the authors neglect to analyze the effects that a specific event might have on other events later in the case. Previous research indicates that events that occur during the early part of a procedure can significantly impact events that occur later in the case. A deeper, more thorough analysis of the relationship among “glitches” would be even more enlightening. Finally, it appears that the main focus of this paper is to highlight the “glitch method” that the authors have developed. However, this “glitch method” is no different than other methods that already exist. Adding another term to the literature when other more established terms already exist (e.g., non-routine events, surgical flow disruptions) only muddies the waters and adds confusion to the literature. In the Introduction, the authors appear to downplay the prevalence and significance of previous research in an attempt to “sell” their “glitch method”. They also fail to integrate their results with the findings of previous studies, which also suggests that their goal is to portray their study as being completely unique when in fact it is an extension of a large body of research that has preceded it. For this study to be of major significance, it would need to have established what the authors state in the Introduction as a “systematic analysis of the magnitude of the impact of [glitches] on patient outcomes.” In summary, this manuscript corroborates findings in the existing literature regarding the impact of work system factors on performance and surgical care. The paper also provides some additional data concerning the reliability of observational methods commonly used in this type of research, as well as some novel approaches for analyzing data. A more thorough integration of this research within the context of previous research would help provide a more accurate portrayal of this study.

VERSION 1 – AUTHOR RESPONSE

Reviewer: Alexander F. Arriaga, Brigham and Women's Hospital, Ariadne Labs, Harvard School of Public Health

This is a well-designed study and well-written manuscript presented by authors who are widely-published leaders in the field of surgical safety.1-3 Their presentation of the “glitch method” helps us come to a better understanding of intraoperative process deviations based on data obtained through real-time measurement of operating room behavior. In a multi-center study of over 400 operations, the authors found that, on average, there are fourteen “glitches” per operation (a “glitch” is defined by the authors as a “deviation from the recognized process with the potential to reduce quality or speed, including interruptions, omissions and changes”). The majority of glitches occurred within the first half of the operation. Distractions were the most common glitches observed in all cases (Figure 2). These findings only emphasize the value of interventions such as safety pauses and checklists, which have been shown to reduce morbidity and mortality4-7 and are increasingly being used in the operating room. I have a few questions and comments for the authors:

1. It is interesting that the range of glitches for a given case was 0 to 83 per operation. Some of this range of course represents the length of the operation, but there may be lessons that can be gathered from this finding. In manuscript page 10, lines 42-55, the reader can appreciate that this range was 1-63 for elective orthopedic surgery, 1-35 for trauma orthopedics, 2-49 in elective vascular surgery, and 1-83 in elective plastic surgery. Hence, across specialties, some cases had less than 3 glitches total.

Do the authors have any speculation, other than operative duration or institution, on why some cases had so few glitches and others had a much greater quantity? The only comment I see relating to this is in manuscript page 12, lines 50-55: “The rate of distractions is the greatest in plastics which relates anecdotally to the discursive and fluid nature of the teams involved.” If the authors/field observers have any further insight on the range of glitches, it would be an interesting addition to either the discussion section of this work or the topic of a future manuscript.

RESPONSE: We agree that this would be an interesting topic for a future manuscript. In summary, the observers do feel that they did witness occasions in which some operations had some glitches which seemed to result in additional glitches.

a. Can the authors clarify the total range of glitches for a given case? In the abstract, the range is stated as 0-83. On page 10 of the manuscript, it states that “The average total glitch count per operation was 14, range 1-83.” I wasn’t sure if the range cited in page 10 was referring to a subset of the data.

RESPONSE: The first of these is correct. There were operations with 0 glitches so the range given (now on page 12) has been amended.

2. The authors observed that glitches seemed to vary more by institution than by specialty. They appropriately point out that this observation was limited by the fact that there were large differences in the case distribution in each site (Table 3) with only 1 site performing trauma orthopedics, one site performing plastic surgery, and one (or more[?]) site(s) performing vascular surgery (manuscript page 13, lines 7-11: “Elective orthopaedic and vascular surgical procedures were observed in multiple sites, providing an opportunity for inter-site rate comparison amongst teams performing the same types of surgery”).
a. Can the authors clarify how many sites had vascular surgical cases? (re: comparison of Table 3 and manuscript page 13, lines 7-11).

RESPONSE: Data was omitted from Table 3 – Site E, Vascular surgery – and this has been

amended. More information has been included in the section on the relationship between glitch rate and hospital site (page 14).

b. On page 13 of the manuscript, the authors comment on the variation of glitch rate by institution. The passage (manuscript page 13, lines 5-29) reads as if they tested each pair of sites for differences in glitch rates but only reported the pairs that were statistically significant. They did provide a combined p-value of the relationship between glitch rate and institution ("There was a statistically significant differences [sic] in mean glitch rates per operation between the four sites [$p<0.001$]"), and they did point out that they did not adjust for multiple testing. However, it would be helpful to the reader for it to be stated that (if the authors tested it) that there were no significant differences between sites A and C, sites A and D, sites A and E, sites B and E, sites C and D, and sites C and E. If the authors did not test the pairs I listed above, then what made them test only the pairs that they reported? I do not think they need to remove the interesting and believable finding that there are differences in glitch rates by institution. A simple sentence would suffice clarifying that the other pairs were tested but not significant (if they were indeed tested).

RESPONSE: We appreciate the comment and have modified the text to make clear what tests were carried out, as suggested (see also our previous response).

3. Comment: I agree with the authors that the Hawthorne effect is not a large limitation to the study (manuscript page 16, lines 47-59: "…the observers quickly became well known to the theatre staff and as such became 'part of the furniture'…). Our research group has had similar anecdotal findings during our operating room field observation work.8-11

RESPONSE: It is nice to hear that other experienced groups share our findings!

In summary, this is a well-prepared manuscript describing a method to capture intraoperative process deviations though structured field observation. It was a multi-institutional study with a relatively large sample size for this type of work. It is worthy of publication.

References:
1. Catchpole K, Mishra A, Handa A, McCulloch P. Teamwork and error in the operating room: analysis of skills and roles. Ann Surg 2008;247:699-706.
2. Mishra A, Catchpole K, McCulloch P. The Oxford NOTECHS System: reliability and validity of a tool for measuring teamwork behaviour in the operating theatre. Qual Saf Health Care 2009;18:104-8.
3. McCulloch P, Catchpole K. A three-dimensional model of error and safety in surgical health care microsystems. Rationale, development and initial testing. BMC Surg;11:23.
4. Haynes AB, Weiser TG, Berry WR, et al. A surgical safety checklist to reduce morbidity and mortality in a global population. N Engl J Med 2009;360:491-9.
5. de Vries EN, Prins HA, Crolla RM, et al. Effect of a comprehensive surgical safety system on patient outcomes. N Engl J Med 2010;363:1928-37.
6. Neily J, Mills PD, Young-Xu Y, et al. Association between implementation of a medical team training program and surgical mortality. JAMA 2010;304:1693-700.
7. van Klei WA, Hoff RG, van Aarnhem EE, et al. Effects of the introduction of the WHO "Surgical Safety Checklist" on in-hospital mortality: a cohort study. Ann Surg 2012;255:44-9.
8. Hu YY, Arriaga AF, Roth EM, et al. Protecting patients from an unsafe system: the etiology and recovery of intraoperative deviations in care. Ann Surg 2012;256:203-10.
9. Hu YY, Peyre SE, Arriaga AF, Roth EM, Corso KA, Greenberg CC. War stories: a qualitative analysis of narrative teaching strategies in the operating room. Am J Surg 2011;203:63-8.
10. Hu YY, Arriaga AF, Peyre SE, Corso KA, Roth EM, Greenberg CC. Deconstructing intraoperative communication failures. J Surg Res 2012;177:37-42.
11. Hu YY, Peyre SE, Arriaga AF, et al. Postgame analysis: using video-based coaching for

continuous professional development. J Am Coll Surg 2011;214:115-24.

Reviewer: Dr Tobias Gauss, Consultant, Surgical Critical Care Unit, Department of Anaesthesia and Critical Care Hôpital Beaujon HUPNVS, APHP, Clichy, France

1) It appears not clear which interventions have been included and which not, see comment 7) and 10).
2) see indicated in comment 2) and 3) and 5)
3) see comments 6) and 9)
4) with regard to the consort checklist only the items that have been mentioned before

The authors have attempted to develop a method to screen, quantify and categorize events, so called glitches that disturb the surgical workflow and could ultimately lead to adverse events and complications. This effort is pertinent and welcome, since to my knowledge, no such tool exists so far. Although there is an abundant literature on the subject, it appears necessary to homogenize the description of perioperative events. The question is pertinent and the methodology appears adapted.

Questions and comment in details:

1) P6, L 14, I suggest to start with a description of the development of the method and validation. This appears easier to follow for the reader.

RESPONSE: We accept the argument. As suggested, the sections have been reordered, and we agree that this has improved clarity and flow.

2) P6, L53, the description of the development and validation of the method appears quite short. Although the reader can of course refer to the description in Morgan et al 2011, a resume of the main elements appears necessary: how long was the validation period, how many interventions, which specialty, how were the categories chosen, etc.
It would also be interesting what the baseline glitch rate, as well as the baseline complication rate was.

RESPONSE: As requested, the main elements of the method have now been included, along with the baseline glitch rate. We do not have data on the baseline complication rate at this time as clinical data will be analysed later

3) P6, L20, How many observers were there in total? Were the observers identical to the researchers that conceived the categories? If so they might be biased towards their own methodology. They same logic would apply if the researches also conducted the initial validation of the methodology described in the paper Morgan et al 2011.
This seems to be the case P7, L8, as the observers also conducted the final review of the categorization of all glitches. A bias towards the conceived categories cannot be excluded. Was any independent person involved in the observation as well in the final review? This appears of importance to me, as this may question the external validity of the method.

RESPONSE: The point about bias is an important one. However it is important to distinguish between the glitch method per-se and the classification of glitches. The observational method had already been established in our earlier work referenced in the paper by Morgan et al 2011. The novel classification of the observed events was developed by the original team members via discussion and consensus as described, and we do not feel that, conceptually, such consensus can be described as

bias.

Further information has been added to the text on pages 6 and 7 to clarify this point, as follows:
“Prior to the final analysis, the glitch data was reviewed jointly by the observers (LM, MH, SP, ER).”
“Four of the six observers (MH, SP, ER and LM) were involved in the creation of the method, the remainder (LB and JM) were introduced to the categorisation at a later date.”

4) P6, L 38:
The authors state that technical skills and WHO checklist adherence were collected, but these data are not clearly provided later in the result section. Is non-compliance with the WHO-Checklist also a glitch?

RESPONSE: A line has been added to the relevant paragraph, now on page 7, to clarify this point:
“Non-compliance with the WHO surgical safety checklist was not considered within the glitch scale.”

5) P6, L58 and table 1:
It appears to me that the glitch categories do not include glitches such as deviation from planned surgical strategy or any unexpected clinical events (allergy, unexpected bleeding, etc) or unexpected surgical difficulty, but maybe do I not understand the meaning of all categories.

RESPONSE: We thank the reviewer for pointing out the deficiency in our explanation of the categories. We have revised Table 1 to show how these two types of events would be categorised, and to demonstrate that the categories deal with all types of process deviations including the important ones mentioned by the reviewer.

Was rescheduling considered as a glitch?
Do the glitches only apply to intraoperative glitches or was the whole perioperative period included? In my opinion perioperative (pre-, intra- and postoperative) organisational difficulties and communication breakdowns can be very disturbing impacts on the workflow and patient safety.
Were glitches with regard to the anaesthetic management taken into account?

RESPONSE: Further information has been added to the text on page 7 as follows to clarify this point:
“Any process disruption which occurred in the pre- or post- operative phase were not included in this method as it was thought that the collection of these events would not be as reliable as those collected in the intra-operative period.”

Personally I would suggest that not all suggested glitch-categories carry the same amount of risk for potential harm. Did the authors attempt or think about rating the categories and glitches?

RESPONSE: Sentence added to discussion section (page 16):
“We consider that some glitch categories may correlate with patient harm events more than others, however we did not test this hypothesis in this study. However the consequent harm from any glitch depends so critically on other factors including other glitches and team resilience that teasing out the relationships between glitch type and harm is likely to be very complex”

6) P7, L3 Does the indicated kappa take into account that there were several raters involved? A kappa of 0,7 appears low under the assumption that the researchers themselves have used it and perform a joint review in the case of disagreement. It would be interesting to dispose of the kappa of the prospective data presented in this paper.

RESPONSE: Yes this kappa result is between the four observers. It was on a sub-set of the glitches (94) and not the whole sample. The test was done on initial results before any discussion of

disagreements. We have the data to calculate kappa scores for the entire study population, but are not able to analyse this within the response time required by the journal. We hope to include this in a later publication once clinical outcomes are available.

7) P7, L37
Why has visceral surgery not been included? Visceral surgery is one of the most common types of surgery. Including visceral surgery would probably increase the external validity.

In this context, with regard to table 3, orthopaedics were observed in all five centres, but trauma, plastic and vascular, only at one site each. Why could not more observations be made at the other sites?
40% of interventions have been observed at one site with a disproportionate amount of observations made in orthopaedics. This could be a source of bias.
(Minor remark, the numbers on duration is a bit cumbersome to read.

RESPONSE: The larger programme, the Safer Delivery of Surgical Services (S3), of which this study formed part was based originally in Elective Orthopaedics and only subsequently was expanded to include other specialities. We agree that including a wider variety of specialities would have provided greater external validity. We have used the system in a small number of visceral surgery cases with apparently consistent results.

8) P7, L39
It would be interesting to know more about the kind of surgery, the surgical difficulty, the experience of the operators, the percentage of emergency surgery and to dispose of some clinical data, such as median ASA.

RESPONSE: The Orthopaedic surgery cases comprised mainly hip and knee replacements, arthroscopies and cruciate ligament repairs. The Trauma surgery varied widely but the predominant injuries were fractured neck of femur, tibial fractures and wrist fractures. These were the only emergency cases in the study. The vascular surgery cases comprised aortic aneurysm repairs, femoro-popliteal and femoro-distal bypasses and varicose vein surgery. The experience of the operators varied widely, but the majority of the cases were performed by experienced Consultants except in the case of Trauma surgery. We hope to publish more clinical data in a later paper as indicated above, but feel that it is unlikely to have a major bearing on the properties of the glitch method itself.

9) P9, L8, if I understand it correctly, no power analysis to determine a required number of observations for an estimated precision is provided, which is all the more striking, since the authors dispose of preliminary data.

RESPONSE: This is correct. We didn't carry out such an analysis as our previous experience with precursor methods[4,18] suggested our sample size should be more than adequate.

10) P 9, L 19, 429 operations over 16 months does not appear a lot to me. Why not more over a shorter period of time? Who made choice over which

RESPONSE: This number of operations is small for a study of clinical outcome, but very large for a study of operating theatre process in which the entire operation was observed by two team members. The observations alone comprise over 1,000 person-hours. We identified suitable operating theatres and lists for study and observed an entire list on each occasion that we attended. We attended as often as logistics allowed, resulting in a convenience sample of between 15 and 25% of the cases performed by the teams being observed over the time period of the study. As mentioned previously,

the results displayed here are part of the Safer Delivery of Surgical Services (S3) study which attempts to provide evidence for different intra-operative process improvement techniques. As the S3 study was set up as a controlled pre-post designed study we elected to present all control but only pre-intervention active glitch data. We felt that this was important as it was hypothesised that the intervention may prove to have an effect on the glitch rate and characteristics. This therefore reduces our total sample of observed operations. It is also important to state that within the 16 month period, the intervention was deployed reducing the amount of time to observe at each operative site.

11) P 10, L 45 and further on,
The glitch rate per hour and specialty is an interesting concept and appears intuitive and useful as indicator. Nevertheless it would be interesting to see a link to the level of competence of the operators, skill of team, level of teamwork, surgical and clinical difficulty, etc.

RESPONSE: We thank the reviewer for their comments. We collected information on which team members were present in theatre and we intend to discuss the potential influence of team stability on theatre performance in a future paper. With regards to the skills of the operator, the majority of the operations were performed by consultants or senior fellows. Due to difficulty in assessing 'training' of surgical teams we felt that it would be impracticable to reliably rate the lead operators' experience. We also felt that this puts a potentially unbalanced emphasis on the training of the surgical operator and we consider that if this information was to be collected it would be necessary to collect this information on the nursing and anaesthetic staff too. With regards to rating operative difficulty, we did attempt to collect information on the perceived difficulty of the operative procedure but did not find a sufficiently systematic method to achieve this goal. We did collect information on the 'level of teamwork' on a modified Oxford NOTECHS. We are about to submit this work for publication and within that paper we describe the relationship between glitch and Oxford NOTECHS II.

12) It seems that the authors do not present any data that could link the glitches to measurable and pertinent clinical events that may have an impact on quality of care, patient safety or even complications. This was of course not the scope of the paper, but if this were the case this would be a pity given the important effort they have made. The work would be all the more interesting.
Overall the questions is pertinent, the work and contribution to the field potentially useful, but a test of the external validity of the tool in other institutions appears appropriate before putting it into use.

RESPONSE: We accept the need for further work to study the relationship between glitch rates clinical outcomes, as noted above. We hope to be in a position to provide this information in a subsequent paper.

Reviewer: Douglas A. Wiegmann, PhD, Department of Industrial and Systems Engineering University of Wisconsin-Madison USA

What is the outcome of interest?
Update the references.
Over extend findings/implications

Authors fail to integrate findings to other studies Authors downplay previous research to over sell their methods/finding.

The goal of this research project was to develop a methodology for identifying "technical process imperfections during surgical procedures" in order to assess their impact on patient outcomes and evaluate the effectiveness of interventions.

There are several positive aspects of this study. The first being the authors' attempt to assess inter-rater reliability of the observational approach utilized. The authors were able to demonstrate acceptable levels of reliability across different types of raters (human factors vs. medical) and raters standing in different locations. This is somewhat conflicting with the results of previous research that indicates that an observer's location in the operating room impacts the types of events that they observe. Research also suggests that observers with medical backgrounds tend to observe more clinically related events than human factors observers (who tend to observe more work system factors). Unfortunately, the authors do not address their findings in light of the pervious research.

RESPONSE: We are glad to note the reviewers' positive comments about our reliability findings. Although we have stressed the agreement figures, where disagreement existed we saw trends which reproduce the findings this reviewer mentions. However we note that others have also found no significant differences between clinical and non-clinical observers in using this type of tool[1]. Agreement could only be evaluated where both observers noted a glitch and like other groups [2] we found that quite a large percentage of events were only noted by one observer. We have referenced and alluded to these earlier findings. We did not keep records of where observers stood so cannot comment on this as an influence on the observations.

Another positive aspect of this research project is that the authors were able to create a generic system for documenting "glitches" across procedures. Previous methodologies have been procedure specific and researchers have had to significantly modify existing data collection tools to accommodate the nuances associated with individual surgical procedures. The ability for both researchers and practitioners to have access to a generic "off the shelf" data collection tool could be of great benefit for those who do not have the resources to develop their own data collection tool. However, the authors fail to discuss the limitations of using such a generic approach to data collection. Clearly there are differences that exist across surgical procedures (e.g., equipment utilized, personnel involved, and medications/anesthesia required, etc.). Understanding what one might sacrifice in terms of specificity from using a generic data collection method is very important.

RESPONSE: We are grateful for this positive comment on our work but accept the criticism of our discussion of the trade-offs necessary to achieve a generically useful tool. We have added the following to the discussion section to correct this deficiency:
"This novel method builds upon previous experience and has resulted in a tool which is transferrable between surgical disciplines. We consider that the method has been shown to be sufficiently robust to be of use in the assessment of most intra-operative settings. However differences in personnel, procedures and equipment in different types of surgery are likely to result in systematic differences in median baseline glitch rates. We therefore suggest that the principal use of the method should be to follow change within a team in response to influences such as stressors or training, rather than comparisons between operation types."

The authors have also proposed a novel way of analyzing events during individual surgical cases. The analysis of the cumulative proportion of "glitches" across the duration of a case provides a unique insight into the timing of "glitches" during a procedure. Unfortunately, the authors neglect to analyze the effects that a specific event might have on other events later in the case. Previous research indicates that events that occur during the early part of a procedure can significantly impact events that occur later in the case. A deeper, more thorough analysis of the relationship among "glitches" would be even more enlightening.

RESPONSE: We agree that there is much work to be done to analyse the relationships between glitches and subsequent events, and most importantly operation outcome. Whilst our data may allow us to attempt such analysis at a later date, it would be too extensive to include in this paper, whose

principal objectives are to describe the properties of the method and make some simple observations about what it tells us about performance variability in surgery.

Finally, it appears that the main focus of this paper is to highlight the “glitch method” that the authors have developed. However, this “glitch method” is no different than other methods that already exist. Adding another term to the literature when other more established terms already exist (e.g., non-routine events, surgical flow disruptions) only muddies the waters and adds confusion to the literature. In the Introduction, the authors appear to downplay the prevalence and significance of previous research in an attempt to “sell” their “glitch method”. They also fail to integrate their results with the findings of previous studies, which also suggests that their goal is to portray their study as being completely unique when in fact it is an extension of a large body of research that has preceded it. For this study to be of major significance, it would need to have established what the authors state in the Introduction as a “systematic analysis of the magnitude of the impact of [glitches] on patient outcomes.”

RESPONSE: We are disappointed at the reviewer’s interpretation of our motives, which is entirely incorrect. We have great respect for the work of previous authors who have investigated this difficult field, including Dr Wiegmann himself. We apologise if we did not refer sufficiently to previous work in the Introduction and Discussion, and we have attempted to correct this in the revised MS. We accept that it is important to put our work in context, but would argue (as do the other reviewers) that our method does have some properties which are an advance on previously published methods. We agree that the final arbiter of significance will be demonstration of relevance to patient outcomes, but have explained that we will present this in a later paper once follow-up data collection and analysis are complete.

In summary, this manuscript corroborates findings in the existing literature regarding the impact of work system factors on performance and surgical care. The paper also provides some additional data concerning the reliability of observational methods commonly used in this type of research, as well as some novel approaches for analyzing data. A more thorough integration of this research within the context of previous research would help provide a more accurate portrayal of this study.

1. Russ, S., et al., Observational teamwork assessment for surgery: feasibility of clinical and nonclinical assessor calibration with short-term training. Annals of surgery, 2012. 255(4): p. 804-809.
2. Schraagen, J.M., et al., Assessing and improving teamwork in cardiac surgery. Quality and Safety in Health Care, 2010. 19(6): p. e29.

VERSION 2 – REVIEW

REVIEWER	Tobias Gauss Service Anesthésie et Réanimation Hôpital Beaujon HUPNVS, APHP, Clichy, France
REVIEW RETURNED	02-Oct-2013

GENERAL COMMENTS	The authors have addressed all major comments and suggestions from the first review. One question that remains and that should in my opinion be specified: glitches that concerned the anaesthetic management were they part of the observation? It seems not. If so why? Anaesthesia is an inherent element of the medico-surgical process. Why not include it into the glitch matrix? This should be commented in the discussion.

REVIEWER	Douglas A. Wiegmann Department of Industrial and Systems Engineering University of Wisconsin-Madison U.S.A.
REVIEW RETURNED	14-Oct-2013

GENERAL COMMENTS	The authors have done a fine job revising this manuscript. They have addressed the majority of my concerns with the original version of this paper. One question that still remains however is the lack of distinction between “glitches” (as conceptualized by the authors) and other types of work system factors that have been found previously to influence performance in the operating room (e.g., surgical flow disruptions or non-routine events). Perhaps this issue is beyond the scope of the study, but if this area of research is to move forward, a general consensus regarding methodology and terminology needs to be derived. This paper clearly helps facilitate the former.

VERSION 2 – AUTHOR RESPONSE

Reviewer: Dr Tobias Gauss, Consultant, Surgical Critical Care Unit, Department of Anaesthesia and Critical Care Hôpital Beaujon HUPNVS, APHP, Clichy, France

The authors have addressed all major comments and suggestions from the first review.

One question that remains and that should in my opinion be specified: glitches that concerned the anaesthetic management were they part of the observation?
It seems not. If so why? Anaesthesia is an inherent element of the medico-surgical process. Why not include it into the glitch matrix? This should be commented in the discussion.

RESPONSE: Thank you for making this distinction. We hope we have clarified that anaesthetic activity was considered an important element, so was part of the observations. See Comment 1 in the revised paper (page 7 of 20).

Reviewer: Douglas A. Wiegmann, PhD, Department of Industrial and Systems Engineering University of Wisconsin-Madison USA

The authors have done a fine job revising this manuscript. They have addressed the majority of my concerns with the original version of this paper. One question that still remains however is the lack of distinction between “glitches” (as conceptualized by the authors) and other types of work system factors that have been found previously to influence performance in the operating room (e.g., surgical flow disruptions or non-routine events). Perhaps this issue is beyond the scope of the study, but if this area of research is to move forward, a general consensus regarding methodology and terminology needs to be derived. This paper clearly helps facilitate the former.

RESPONSE: Thank you for acknowledging the work to improve the quality of the paper, and your help in developing it. We hope we have clarified the differences between glitches and other types of work system factors, and agree the need for a general consensus. See Comment 2 in the revised paper (page 16 of 20).