PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Quantifying uncertainty in intervention effectiveness with structured expert judgment: an application to obstetric fistula
AUTHORS	Colson, Abigail; Adhikari, Sweta; Sleemi, Ambereen; Laxminarayan, Ramanan

VERSION 1 - REVIEW

REVIEWER	Roger M. Cooke Resources for the Future, USA, Univ. Strathclyde, UK
REVIEW RETURNED	12-Jan-2015

GENERAL COMMENTS	Remarks on "Quantifying uncertainty in intervention effectiveness with structured expert judgment: an application to obstetric fistula" Abigail R. Colson, Sweta Adhikari, Ambereen Sleemi, Ramanan Laxminarayan Roger M. Cooke Resources for the Future, Univ. Strathclyde, TU Delft (ret) Jan. 12, 2015 Introduction Although not involved in this study, I know the authors well, have collaborated with them in the past, and hold their diligence and professionalism in highest regard. It is very gratifying to see how the authors have applied the Structured Expert Judgment (SEJ) method to a new problem area. In a blind referee process I would recuse myself on grounds of personal acquaintance; however, in an authored review this barrier is inoperative. Nonetheless, asking me to review this paper is a bit like asking the priest if you should go to mass. The authors make a good case for applying SEJ in this case, but I am unable to comment on the value of this research for fistula treatment. I can offer some clarification on the methodology. A final section gives links to recent applications of SEJ, where additional background information may be found. Methodological remarks Although "Classical Model" and "Cooke Method" both acronymize to CM, the former is the correct designation. "Classical" refers to the fact that experts assessing their uncertainty on potentially observable quantities are treated as classical statistical hypotheses and scored in terms of statistical accuracy and informativeness. The description of the method in the manuscript is appropriate for an audience more concerned with fistula than with SEJ. I would nonetheless place somewhat different inflections. Describing statistical accuracy as "90% of calibration question true

values fall within the 90% credible range and 50% of the true values fall within the 50% credible range" invites confusion. Scoring statistical accuracy depends critically on the number of observations at hand. If 80% of the true values fall within the 90% credible range, for 10 calibration variables, that might be very decent statistical accuracy. However, if this is the case for 100 calibration variables, it would point to very low statistical accuracy. Indeed, statistical accuracy measures the probability of falsely rejecting the hypothesis that an expert's probabilistic statements are statistically accurate. Suppose the probability of a value in the 90% credible range were really 90% for each of the (independent) true values. What is then the probability of seeing 8 or fewer of 10 true values falling in this range? An elementary calculation gives this probability as 0.2639. The probability of 6 or fewer of the 10 true values fell within the 90% credible range is 0.01280. Believing that the "6 out of 10" expert were statistically accurate would entail believing that an event with 1.28% probability had just occurred. If there were 100 calibration variables, then the probability of seeing 80% or fewer within the 90% credible range would be 0.001979. The statistical accuracy score in the case of "6 out of 10" (0.0128) is much higher than the statistical accuracy score of "80 out of 100" (0.001979). Judging statistical accuracy involves such calculations for all the probabilistic information given by the expert, not just the 90% credible range. It also involves properly accounting for the number of calibration variables.

Informativeness is roughly described as "a measure of how peaked an expert's uncertainty distributions are, with more peaked distributions indicating a narrower range of values in the credible range and thus less uncertainty". Somewhat more accurately, informativeness measures the degree to which an expert is able to concentrate high probability mass in a small interval. Still, much important detail is concealed in such descriptions. Why not measure informativeness as a standard deviation, or precision, as taught in all introductory statistics classes? There are two reasons. (1) Not all distributions are adequately characterized by their mean and standard deviation, and both these quantities are driven by the "tails" of the distribution. In SEJ we typically do not query experts about these tails, as judgments about the very unlikely events are notoriously poor. Instead we rescale the questions. Instead of asking for the number of failures in one meter of steel piping in one year, we would ask for the number of failures in 100 km of similar steel piping in 10 years. (2) The standard deviation has a physical dimension. If we convert a question from meters to kilometers, the standard deviation decreases by a factor 1000. When scoring informativeness over a set of calibration questions with different physical dimensions, it is imperative to have a "scale invariant" measure of informativeness. Information theory provides just such a measure, called the Shannon relative information. This measure forces us to face an important fact: there is no absolute measure of informativeness, we measure only the information in one distribution relative to another. Hence CM measures informativeness relative to an analyst-chosen background measure which applies to all experts in a panel. This entails that informativeness scores may be meaningfully compared within a panel but not between panels. An expert can give him/herself a high information score by choosing small credible intervals. However, that may make "surprise" values outside the 90% credible interval more likely, thus driving down the statistical accuracy score. The challenge is to find a combination rule which achieves both statistical accuracy and informativeness.

	The manuscript skirts one important feature. If an expert knows how (s)he will be scored, and wants to craft his/her response to maximize the expected score, what should (s)he do? A scoring rule is "strictly proper" if an expert maximizes his/her score per item by stating the probabilities (s)he actually believes. This is a very strong mathematical constraint and is massively violated by many naive scoring rules. CM scores experts not "per item" but "per set of items" and the mathematics in the latter case are somewhat different. However, CM combines the statistical accuracy and informativeness scores so as to become a strictly proper scoring rule in an appropriate long run sense. Links and details The following links provide freely downloadable information on SEJ. The Supplementary Online Material (SOM) with a recent article on messaging uncertainty in climate change gives background on the representation of uncertainty as subjective probability and the mathematics of CM. http://www.nature.com/nclimate/journal/v5/n1/extref/nclimate2466-s1.pdf SOM accompanying an article on the Asian carp invasion in Lake Erie gives mathematical background and elicitation details: http://onlinelibrary.wiley.com/doi/10.1111/cobi.12369/suppinfo The SOM accompanying an article on out-of-sample validation gives details on validation and cross validation: http://onlinelibrary.wiley.com/doi/10.1002/ieam.1559/abstract A special issue of Radiation Protection and Dosimetry contains several articles spun off a large SEJ study on the risks of nuclear power plants: http://rpd.oxfordjournals.org/search?tmonth=&pubdate_year=2000&submit=yes&submit=Search&submit=yes&andorexacttitle=and&format=standard&firstpage=&fmonth=&title=&tyear=&hits=10&titleabstract=&flag=&volume=90&sortspec=relevance&andorexacttitleabs=and&author2=&andorexactfulltext=and&author1=&fyear=&doi=&fulltext=&FIRSTINDEX=10

REVIEWER	Hassan Shaker Professor of Urology, Faculty of Medicine, Ain Shams University. Cairo, Egypt
REVIEW RETURNED	14-Mar-2015

GENERAL COMMENTS	Review of: BMJ manuscript Title: Quantifying uncertainty in intervention effectiveness with structured expert judgment: an application to obstetric fistula The objective of this manuscript is very intriguing because it addresses an area that seldom has been approached. Since expert opinion occupies the lowest rank in the modern era of evidence-based medicine, another more reliable method has to be used when more established evidence is not available to reach a conclusion. I am not aware of any other study in urology or female urology that

	has used the Cooke method to quantify the expert opinion. According to the authors, this method has been applied in the other non-medical areas. Thinking out of the norms in many instances may lead to a quantum leap of progress in some areas of science. Some points need to be addressed in the manuscript: A) Methodology: a. Kindly define “expert” in your study. b. The 90 and 50% credibility ranges need to be explained in a less confusing way since this is makes the crux of the methodology. c. Calibration questions are unrelated to areas explored. Although I am not familiar with the Cooke method, it seems to me that they have to be related. Most of the expertise of experts is concentrated in certain areas. For example, an expert fistula surgeon may have vast knowledge in surgery outcomes but not in epidemiology in contrast to epidemiologists or social workers concerned with fistula. B) Results: a. Why did the authors choose these scenarios in particular? How were these scenarios formulated? Could direct questions replace these scenarios? b. The authors stated, “The DMs indicated that experts were most certain about long-term outcomes when patients received treatment in high-income countries or high-volume fistula centers. Uncertainty increased when experts thought about outcomes following treatment in low-volume district hospitals or if patients did not receive treatment.” Could this be related to the choice of the experts? For example, if most of the experts are from high-income countries who occasionally worked in Africa in some campaigns or in high volume centers then this may yield this result. C) Discussion: a. It is not clear for me how is it plausible to compare the results of your study to that of the literature. Most of the studies published study certain cohorts of patients that share certain inclusion criteria. Your structured expert judgment is based on certain scenarios that address a very large spectrum of patients. Finally, although this study is addressing a very important problem that we face in our everyday practice, I still can’t see its practicality from the clinical point of view. In other words, as a clinician, I am not sure that the outcome of a study based on this system can influence my decision-making.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1: Roger Cooke

Although not involved in this study, I know the authors well, have collaborated with them in the past, and hold their diligence and professionalism in highest regard. It is very gratifying to see how the authors have applied the Structured Expert Judgment (SEJ) method to a new problem area. In a blind referee process I would recuse myself on grounds of personal acquaintance; however, in an authored review this barrier is inoperative. Nonetheless, asking me to review this paper is a bit like asking the priest if you should go to mass. The authors make a good case for applying SEJ in this case, but I am unable to comment on the value of this research for fistula treatment. I can offer some clarification on

the methodology. A final section gives links to recent applications of SEJ, where additional background information may be found.

Methodological remarks
Although "Classical Model" and "Cooke Method" both acronymize to CM, the former is the correct designation. "Classical" refers to the fact that experts assessing their uncertainty on potentially observable quantities are treated as classical statistical hypotheses and scored in terms of statistical accuracy and informativeness. The description of the method in the manuscript is appropriate for an audience more concerned with fistula than with SEJ. I would nonetheless place somewhat different inflections.

Response:
We have changed all references to the “Cooke method” to the “classical model.”

Describing statistical accuracy as "90% of calibration question true values fall within the 90% credible range and 50% of the true values fall within the 50% credible range" invites confusion. Scoring statistical accuracy depends critically on the number of observations at hand. If 80% of the true values fall within the 90% credible range, for 10 calibration variables, that might be very decent statistical accuracy. However, if this is the case for 100 calibration variables, it would point to very low statistical accuracy. Indeed, statistical accuracy measures the probability of falsely rejecting the hypothesis that an expert's probabilistic statements are statistically accurate. Suppose the probability of a value in the 90% credible range were really 90% for each of the (independent) true values. What is then the probability of seeing 8 or fewer of 10 true values falling in this range? An elementary calculation gives this probability as 0.2639. The probability of 6 or fewer of the 10 true values fell within the 90% credible range is 0.01280. Believing that the "6 out of 10" expert were statistically accurate would entail believing that an event with 1.28% probability had just occurred. If there were 100 calibration variables, then the probability of seeing 80% or fewer within the 90% credible range would be 0.001979. The statistical accuracy score in the case of "6 out of 10" (0.0128) is much higher than the statistical accuracy score of "80 out of 100" (0.001979). Judging statistical accuracy involves such calculations for all the probabilistic information given by the expert, not just the 90% credible range. It also involves properly accounting for the number of calibration variables.

Informativeness is roughly described as "a measure of how peaked an expert’s uncertainty distributions are, with more peaked distributions indicating a narrower range of values in the credible range and thus less uncertainty". Somewhat more accurately, informativeness measures the degree to which an expert is able to concentrate high probability mass in a small interval. Still, much important detail is concealed in such descriptions. Why not measure informativeness as a standard deviation, or precision, as taught in all introductory statistics classes? There are two reasons. (1) Not all distributions are adequately characterized by their mean and standard deviation, and both these quantities are driven by the "tails" of the distribution. In SEJ we typically do not query experts about these tails, as judgments about the very unlikely events are notoriously poor. Instead we rescale the questions. Instead of asking for the number of failures in one meter of steel piping in one year, we would ask for the number of failures in 100 km of similar steel piping in 10 years. (2) The standard deviation has a physical dimension. If we convert a question from meters to kilometers, the standard deviation decreases by a factor 1000. When scoring informativeness over a set of calibration questions with different physical dimensions, it is imperative to have a "scale invariant" measure of informativeness. Information theory provides just such a measure, called the Shannon relative information. This measure forces us to face an important fact: there is no absolute measure of informativeness, we measure only the information in one distribution relative to another. Hence CM measures informativeness relative to an analyst-chosen background measure which applies to all experts in a panel. This entails that informativeness scores may be meaningfully compared within a panel but not between panels. An expert can give him/herself a high information score by choosing

small credible intervals. However, that may make "surprise" values outside the 90% credible interval more likely, thus driving down the statistical accuracy score. The challenge is to find a combination rule which achieves both statistical accuracy and informativeness.

The manuscript skirts one important feature. If an expert knows how (s)he will be scored, and wants to craft his/her response to maximize the expected score, what should (s)he do? A scoring rule is "strictly proper" if an expert maximizes his/her score per item by stating the probabilities (s)he actually believes. This is a very strong mathematical constraint and is massively violated by many naive scoring rules. CM scores experts not "per item" but "per set of items" and the mathematics in the latter case are somewhat different. However, CM combines the statistical accuracy and informativeness scores so as to become a strictly proper scoring rule in an appropriate long run sense.

Response:
We thank the reviewer for these comments, but feel the level of detail described here is beyond the scope of the article. We have slightly revised the definition of statistical accuracy so that it now reads (page 8, lines 5-7):
"That is, as the number of calibration questions increases, the frequency of capturing the true values within the 90% credible range approaches 90%. Similarly, the frequency of capturing true values within the 50% credible range approaches 50%."

We have also added the following text and references to the discussion of expert scoring in the methods section (page 8, lines 11-12). We can add additional detail in the online supplementary material to the paper, if desired.
"A more detailed description of these two scores and the expert scoring procedure is available elsewhere.[1,14,16–19]"

References:
1 Cooke RM. Experts in Uncertainty: Opinion and Subjective Probability in Science. New York: Oxford University Press 1991. doi:10.1016/0040-1625(93)90030-B

14 Cooke RM, Goossens LLHJ. TU Delft expert judgment data base. Reliab Eng Syst Saf 2008;93:657–74. doi:10.1016/j.ress.2007.03.005

16 Wittmann ME, Cooke RM, Rothlisberger JD, et al. Using structured expert judgment to assess invasive species prevention: Asian carp and the Mississippi - Great Lakes hydrologic connection. Environ Sci Technol 2014;48:2150–6. doi:10.1021/es4043098

17 Cooke RM, Wittmann ME, Lodge DM, et al. Out-of-sample validation for structured expert judgment of Asian carp establishment in Lake Erie. Integr Environ Assess Manag 2014;10:522–8. doi:10.1002/ieam.1559

18 Cooke RM. Messaging climate change uncertainty. Nat Clim Chang 2015;5:8–10. doi:10.1038/nclimate2466

19 Wittmann ME, Cooke RM, Rothlisberger JD, et al. Use of structured expert judgment to forecast invasions by bighead and silver carp in lake erie. Conserv Biol 2015;29:187–97. doi:10.1111/cobi.12369

Links and details
The following links provide freely downloadable information on SEJ.

The Supplementary Online Material (SOM) with a recent article on messaging uncertainty in climate change gives background on the representation of uncertainty as subjective probability and the mathematics of CM.
http://www.nature.com/nclimate/journal/v5/n1/extref/nclimate2466-s1.pdf

SOM accompanying an article on the Asian carp invasion in Lake Erie gives mathematical background and elicitation details:
http://onlinelibrary.wiley.com/doi/10.1111/cobi.12369/suppinfo

The SOM accompanying an article on out-of-sample validation gives details on validation and cross validation:
http://onlinelibrary.wiley.com/doi/10.1002/ieam.1559/abstract

A special issue of Radiation Protection and Dosimetry contains several articles spun off a large SEJ study on the risks of nuclear power plants:
http://rpd.oxfordjournals.org/search?tmonth=&pubdate_year=2000&submit=yes&submit=Search&submit=yes&andorexacttitle=and&format=standard&firstpage=&fmonth=&title=&tyear=&hits=10&titleabstract=&flag=&volume=90&sortspec=relevance&andorexacttitleabs=and&author2=&andorexactfulltext=and&author1=&fyear=&doi=&fulltext=&FIRSTINDEX=10

Reviewer 2: Hassan Shaker

The objective of this manuscript is very intriguing because it addresses an area that seldom has been approached. Since expert opinion occupies the lowest rank in the modern era of evidence-based medicine, another more reliable method has to be used when more established evidence is not available to reach a conclusion.

I am not aware of any other study in urology or female urology that has used the Cooke method to quantify the expert opinion. According to the authors, this method has been applied in the other non-medical areas. Thinking out of the norms in many instances may lead to a quantum leap of progress in some areas of science.

Some points need to be addressed in the manuscript:

A) Methodology:
a. Kindly define “expert” in your study.

Response:
Thank you for your comments and suggestions for our paper. Page 9, lines 1-7 define “expert” as used by our study. No strict inclusion/exclusion criteria (e.g., an expert must have practiced surgery for a certain number of years) was used, except experts must practice in low- and mid-income countries. All of the experts have performed over 500 surgical cases and are active in the training of other surgeons.

b. The 90 and 50% credibility ranges need to be explained in a less confusing way since this is makes the crux of the methodology.

Response:
Thank you for this feedback. This description (page 7, lines 13-18) has been revised and clarified.

c. Calibration questions are unrelated to areas explored. Although I am not familiar with the Cooke method, it seems to me that they have to be related. Most of the expertise of experts is concentrated in certain areas. For example, an expert fistula surgeon may have vast knowledge in surgery outcomes but not in epidemiology in contrast to epidemiologists or social workers concerned with fistula.

Response:
The rationale for choosing epidemiology-focused calibration questions is explained on page 8, lines 19-21:
"Calibration questions were not designed to identify the expert most skilled at performing fistula surgery, but rather the experts able to best think about average outcomes and the surrounding uncertainty for a generic set of fistula patients."

Furthermore, as the experts work at high-volume fistula centers which are involved in all aspects of fistula care, they do have knowledge beyond best surgical practice. None of the expert participants felt the calibration questions were beyond the scope of their knowledge.

B) Results:
a. Why did the authors choose these scenarios in particular? How were these scenarios formulated? Could direct questions replace these scenarios?

Response:
The following information on the scenarios and their reason for inclusion has been added to page 9, lines 14-18:
"Variables of interest questions focused on five scenarios (Table 1), which were constructed by the study team in collaboration with another expert in obstetric surgery. Scenarios were chosen to include a variety of factors that could impact the likelihood of successful surgical repair and were written to be specific enough that the experts' uncertainty distributions would reflect only uncertainty in outcomes and not confusion over the clinical presentation of a given case."

b. The authors stated, "The DMs indicated that experts were most certain about long-term outcomes when patients received treatment in high-income countries or high-volume fistula centers. Uncertainty increased when experts thought about outcomes following treatment in low-volume district hospitals or if patients did not receive treatment."
Could this be related to the choice of the experts? For example, if most of the experts are from high-income countries who occasionally worked in Africa in some campaigns or in high volume centers then this may yield this result.

Response:
All of the experts are based in low- and mid-income countries. The "Participants" section of the abstract and page 9, line 7 have been amended to reflect that.

C) Discussion:
a. It is not clear for me how is it plausible to compare the results of your study to that of the literature. Most of the studies published study certain cohorts of patients that share certain inclusion criteria. Your structured expert judgment is based on certain scenarios that address a very large spectrum of patients.

Response:
Although we think a general discussion of how our results compare to other existing studies is useful, we agree that our results cannot be directly compared to existing studies. We have moved the paragraph discussing caveats for such comparisons earlier in the discussion and expanded its

content (page 14, lines 4-16). The text now reads:
"Estimates of disability and mortality following fistula repair from our study are not directly comparable to the existing literature for several reasons. First, our study focused on five specific scenarios of fistula rather than all cases generally. These scenarios may not represent the full range of cases seen in an observational study. However, the specific cases presenting under our scenarios will vary somewhat. Experts were asked to fold these expected variations into their uncertainty ranges. In an observational study, however, some patients may be dropped under the study's inclusion criteria. Second, all of our cases focus on the first attempt to repair fistula. Surgery outcomes worsen if patients have a history of previously unsuccessful repair, and observational studies based on patients at a single fistula center may include a high proportion of patients referred from other facilities after a previously failed repair attempt.[11] Third, continence may improve time, and few studies look at long-term repair outcomes, due to the challenge of off-site follow-up.[5,11] Thus, existing studies may under-report the actual rate of no incontinence following fistula repair."

Finally, although this study is addressing a very important problem that we face in our everyday practice, I still can't see its practicality from the clinical point of view. In other words, as a clinician, I am not sure that the outcome of a study based on this system can influence my decision-making.

Response:
Page 13, line 18 through page 14, line 2 has been added to clarify the practical implications of the expert elicitation exercise:
"This information can inform future fistula treatment programs. Helping district hospitals identify and treat relatively simple cases like this can improve the quality of fistula care in a region. For more complicated cases, though, where median rates of disability are higher (e.g., Scenario 1) or the uncertainty about outcomes is much larger (e.g., Scenarios 3-5), the best treatment strategy may be referring patients to a high-volume center for the first repair. The difference between expected outcomes following treatment in a high-volume specialty hospital rather than a low-volume district hospital confirms the valuable role specialty care needs to continue to play in a comprehensive fistula treatment program."