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)	What is a medical decision? A taxonomy based on physician statements in hospital encounters – a qualitative study
AUTHORS	Ofstad, Eirik; Frich, Jan; Schei, Edvin; Frankel, Richard; Gulbrandsen, Pål

VERSION 1 - REVIEW

REVIEWER	Christy Ledford Uniformed Services University of the Health Sciences, USA None declared The views expressed in this review are those of the reviewer and do not necessarily reflect the official policy or position of the Uniformed Services University, the Department of Defense, nor the U.S. Government.
REVIEW RETURNED	22-Oct-2015

GENERAL COMMENTS	In introduction, authors infer that the purpose of this paper is why decision making is not commonly a common collaborative practice? (p 3 ln 43) However, the paper does not answer that question; rather it provides a classification system for decisions in general. I recommend removing the inference. P 3, ln 56 is unclear. Are the authors arguing that descriptive, normative, and prescriptive are three mutually exclusive functions? Generally in social science, researchers classify normative in a strictly descriptive or prescriptive function depending on their ontological approach. I may be misunderstanding the sentence (not sure if you're using an Oxford comma or not). More clearly explain how your objective fills the gap pointed out in para 1, p 5. How will this objective include decision making in problem solving? Method This is an impressive data set. Are videotaped encounters part of the pre-intervention (communication skills course) portion of the dataset? Or are they from after the physicians completed the course? This study appears to be a phenomenological approach to clinical decision conversations. I expected to see more language about that approach. The Miller/Crabtree method fits within that qualitative methodology. The specialty information in the abstract is not in the text of the methods. Which specialties are included? That is, specify the types of encounters included. How many are med encounters vs peds encounters? Please describe how you balanced verbal/nonverbal cues in the coding construction. How did transcribed notes reflect nonverbal cues? Unclear when videos versus transcription was used in the process. What is the role of the SOAP note? Were these the actual notes the

	practicing clinician created? Or did the research team produce their own SOAP notes to structure coding? Why was 30 chosen as the subsample for the codebook construction set? How was saturation determined? What's the unit of analysis? Results I would like to see more parallel structure in the category descriptions. Perhaps a definition, example, nonexample pattern. As written, it is hard to follow the large categorical framework. Category 8 is especially hard to grasp the mutual exclusivity. Most likely, precautionary advice appears it could overlap with drug-related or treatment statements. Unclear why they are ordered as is. Provide some context for ordering, or state that it is not systematic. Discussion Unclear how this helps patients sum up appointments. How are patients able to identify these points that required a team of four docs to parse out? Describe how you foresee this scheme being used in practice and in research. For instance, do you see value in exploring demographic differences in the categories enacted? Or specialty differences? Why was it not tested in general practice? This may be a national/systematic difference that I don't perceive. Is general practice not like internal medicine outpatient clinic? Recommend including future research to include extending the coding application to primary care.

REVIEWER	Stephen G Henry University of California Davis Medical School USA I have previously co-authored a paper with one of the co-authors of this manuscript (RMF). I have had no involvement in the study described in this manuscript.
REVIEW RETURNED	26-Oct-2015

GENERAL COMMENTS	This paper describes development of a novel system for coding patient-physician interactions that aims to identify and characterize all decisions within a specific patient-physician interaction. Such a coding system would have important potential for furthering health communication research. I commend the authors for undertaking this project and for the amount of work they have put into this project. Including the coding system as an appendix was very helpful. I do note several important limitations and problems, however, which I shall detail below with the aim of providing authors advice for improving their coding system (DICTUM) and paper. 1. A major potential contribution of DICTUM is to define what actually constitutes a decision. I agree with the authors that traditional decision-making research has focused very narrowly on one-off major decisions (eg whether to have knee surgery or not) and that a broader concept of "decision" is much needed. Developing and justifying this broader concept of "decision" is at least as important conceptually, and may be of more interest to readers, than the specifics of the 10 content categories. The authors should spend more space in the manuscript defining and justifying how they identified / defined a "decision." In particular their rationale for including both classic "decisions" and "judgements" could be

clearer.

2. Related to #1, the authors need to justify / explain why DICTUM dictates that only physician statements convey decisions. The title and intro are misleading. The definition of “decision” on page 7 requires that the statement be made by a medical expert.” A truly comprehensive coding system would drop this restriction. A more accurate title for the current coding system would be “a novel taxonomy for physician decisions during patient-physician encounters.” DICTUM makes no mention of patient statements. Suppose a physician asked, “Do you want to get an xray for your chest,” and the patient replied, “Yes, I’d like an x-ray to help diagnose my problems.” How is this patient statement not a decision? The discussion of patient-centered decision making in the introduction makes clear that patients drive decisions in some cases. Why omit such decisions from DICTUM? Physicians may voice most decisions, but by their nature decisions often emerge during patient-physician interactions, and a truly comprehensive coding system should reflect the fact that decisions are a product of patient-clinician interactions and may be voiced by either patient or physician. In the absence of very convincing justification, the authors should make clear throughout that DICTUM only relates to physician decisions and is not a truly comprehensive coding system of decisions that occur during patient-clinician interactions. Such a coding system would still be potentially useful but would not truly capture and characterize all decisions during patient-clinician interactions, as the authors seem to promise. Alternatively, the authors could expand DICTUM to include patient statements.

3. The introduction does not provide a sufficiently convincing rationale for why a descriptive coding system like DICTUM is needed or how it has potential to advance health communication research. The discussion of patient-centered decision making should be curtailed, because it is not directly relevant for why DICTUM is important. There is a very compelling and convincing case for why a decision coding system such as DICTUM is needed, but I’m not convinced by the authors’ current rationale.

4. The development and coding process for DICTUM is confusing and should be more clearly described. Authors should clarify the degree to which coding was based on video versus transcribed portions of video. The sampling process for DICTUM development should be more clearly explained. It seems that the authors used 30 internal medicine videos for the coding system development. If true, this is a limitation that should be justified. Shouldn’t a truly comprehensive system also draw on inpatient and surgical consultations? Especially since the authors clearly have videos involving many different types of physicians. Was the final coding system applied to all 300 visits, or will this be presented in a future paper?

5. The authors need to provide more detail and explanation for how they handled the issue of “unitizing reliability” in DICTUM. In other words, how did they define the unit of analysis for coding? What did they do, for example, when 1 coder identified a “decision” during a physician statement but the other coder did not? This is relevant because the measure of agreement the authors cite assumes that all reviewers code every “unit.” It may be acceptable to resolve disagreements about the unit of analysis prior to assessing reliability (ie Krippendorf’s alpha) but the authors should state this explicitly in

the manuscript. For further discussion of this issue, see:

Kravitz, R. L., R. A. Bell, et al. (2002). "Characterizing patient requests and physician responses in office practice." Health Services Research 37(1): 217-238.

6. Description of the 10 topical codes could be clearer. For one thing, some of the category labels are quite awkward. Based on the coding manual, “procedure-related” seems more accurate than “surgery-related.” Several other topic headings, including “legally related” and “contact-related” are awkward. For a second issue, the authors should include an explicit definition of each category in the manuscript text related to that topic, something along the lines of “This category includes decisions that involve…” The lack of clear category definitions in the manucript renders many of the topic descriptions confusing and unclear.

7. The coding system itself includes a detailed discussion of the temporality of decisions (past, present, future). This is an important aspect of decisions that is rarely discussed or even acknowledged, and is a strength of the coding system. Briefly discussing in the manuscript text how /why the authors decided to address temporality would likely be of interest to readers.

8. If the authors decide to revise this manuscript, its impact and clarity would be increased by being carefully edited to make sure that the concepts are clear and easy to understand for readers. Although the general thrust of the authors’ reasoning is clear, sentence clarity could be notably improved in many areas.

MINOR COMMENTS
9. Statements about the importance of adding a social psychologist to the team should be deleted. While helpful, inclusion of non-physicians in coding system development is not novel or worth mentioning (other than perhaps with a single sentence in the methods section).

10. The authors should provide a brief explanation of Krippendorf’s alpha. Many readers will not be familiar with this measure.

11. I question the authors’ statement on page 16 that blames the ethics committee for the lack of patient involvement in coding system development. The authors surely could have included patients in the development process if they had so desired – these need not be the same patients as were video recorded. Failure to include patients on the research team is not a major limitaton; this limitation shoud be deleted.

12. Table 1 formatting should be improved.

13. I am not sure that the text ever references table 2. In addition the text does not clearly discuss the different subcategories and why the authors included them. Without Table 2 or a detailed review of the coding manual, readers might not realize that DICTUM includes subcategories. Authors should make sure that use of subcategories is discussed in the manuscript text.

14. In the coding manual, the authors give “I think we should get a liver biopsy” as an example of “information gathering. Why is this statement not coded as surgery-related, since it involes

	recommendation of invasive procedure?” The inclusion of such “judgements” (which are certainly not “decisions) in the coding system needs a stronger justification. What would happen if the patient replied, “I don’t want a liver biopsy.” ? Would that still be coded as a decision?

VERSION 1 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:
Reviewer: 1, Christy Ledford, Uniformed Services University of the Health Sciences, USA
- In introduction, authors infer that the purpose of this paper is why decision making is not commonly a common collaborative practice? (p 3 ln 43) However, the paper does not answer that question; rather it provides a classification system for decisions in general. I recommend removing the inference.

Response to comment: We are happy that both reviewers have endorsed the need for a taxonomy on medical decisions. Accordingly, we have toned down references to shared decision making and collaborative practice as reasons for our study. We have rewritten the first paragraph of the Introduction and as a result removed the final sentence of the first paragraph as reviewer 1 recommended.

- P 3, ln 56 is unclear. Are the authors arguing that descriptive, normative, and prescriptive are three mutually exclusive functions? Generally in social science, researchers classify normative in a strictly descriptive or prescriptive function depending on their ontological approach. I may be misunderstanding the sentence (not sure if you’re using an Oxford comma or not).

Response to comment: The distinction between descriptive, normative and prescriptive functions of decision-making is cited from the perhaps most influential book on medical decision making (Schwartz A, Bergus G. Medical decision making : A physician's guide. New York: Cambridge University Press; 2008.) The different functions are not necessarily mutually exclusive, but in this respect we argue that they provide a meaningful way of separating approaches to the field of decision-making and would like to keep the distinction/citation.

- More clearly explain how your objective fills the gap pointed out in para 1, p 5. How will this objective include decision making in problem solving?

Response to comment: Our objective was to identify and classify all medical decisions emerging in conversations between patients and physicians, and following the review of this paper we have specified these decisions to be based on physician statements in these encounters. In page 5 para 1 we first cite Deber who asserts a distinction between problem-solving and decision-making, before we cite authors who argue against such a distinction. One of the key findings in this study is that diagnostic decisions (in medical literature often referred to as making a diagnosis, making diagnostic judgements or clinical reasoning) demand similar cognitive decision-making efforts, as statements committing to action. Judgment and action decisions may both be rapid and intuitive – or they may both be more complex and require an analytic approach. And both always demand a choice between at least two alternatives.

Method
- This is an impressive data set. Are videotaped encounters part of the pre-intervention (communication skills course) portion of the dataset? Or are they from after the physicians completed the course?

Response to comment: Of the 380 encounters, 209 were recorded before and 171 after training. We have inserted the proportions in the Methods section. We assume that the medical decisions were not much changed by the training, partly because most of them are little influenced by patient input (we do not say that this should be so), and partly because the training was a general course in communication skills that did not leave much room for the demanding task of shared decision making.

- This study appears to be a phenomenological approach to clinical decision conversations. I expected to see more language about that approach. The Miller/Crabtree method fits within that qualitative methodology.

Response to comment: Our study is phenomenological in the sense that we made few presuppositions about what we would find and were therefore open to the emergence of categories that arose from our observations rather than categories that were pre-specified. Jeff Borkan's article on the immersion/crystallization method in Miller and Crabtree's book describes this approach and we have specified our reference with relevant page numbers in our citation.

- The specialty information in the abstract is not in the text of the methods. Which specialties are included? That is, specify the types of encounters included. How many are med encounters vs peds encounters?

Response to comment: We have altered Table 1 to provide numbers and proportions of encounters from different specialties and think that table 1 has improved as a result. Thank you.

- Please describe how you balanced verbal/nonverbal cues in the coding construction. How did transcribed notes reflect nonverbal cues? Unclear when videos versus transcription was used in the process.

Response to comment: As our aim was to identify clinically relevant decisions appearing in conversations between patients and physicians, we focused on the content of the spoken words. We did not study non-verbal communication in particular and there was no balancing between verbal and non-verbal communication.

- What is the role of the SOAP note? Were these the actual notes the practicing clinician created? Or did the research team produce their own SOAP notes to structure coding?

Response to comment: We can see how this statement was unclear. We made the SOAP-notes ourselves in the analytic phase. We did not have access to the patients' medical records. We have incorporated this in the Methods section.

- Why was 30 chosen as the subsample for the codebook construction set? How was saturation determined?

Response to comment: For the purpose of a qualitative study the videotaped encounters was more than large enough to serve as a material depicting everyday clinical practice in hospitals. After having observed, discussed and analyzed statements from 30 encounters we agreed that we were able to identify clinically relevant decisions (as we have written in the Methods section). By formulating the sentence as "being able to identify decisions", we are hopefully more precise, than by agreeing to have reached a point of saturation. In a separate paper, we report the results from the coding of all 380 encounters, and we could not identify any decisions that were unclassifiable during this process.

- What's the unit of analysis?

Response to comment: Thank you for this question. We have now inserted that “statements conveying medical decisions” were our unit of analysis in the Methods section. In the search for statements that conveyed clinical decisions, we ran into discussions sometimes about what separated decisions during one talk turn of the physician. We describe this in detail in the DICTUM codebook which is offered to the reader as an online supplement:

“The identification of decisions is sometimes made easy by shifts in the dialogue. However, in many interchanges and longer monologues it is harder to decide what should count as a decision. The following general rules have been agreed upon and thoroughly tested:

Several codes possible per turn: More than one decision might be conveyed within one turn of speech. However, in order to be coded as separate decisions, they should cover different categories in at least one of the taxonomy’s two dimensions. The physician may do this either within:

- the same decision type (if the physician makes a decision about one drug and goes on to make a decision about another drug).
- different types (decision about a drug followed by a decision scheduling the next control).
- different temporality (see next paragraph)

One code per topic per turn: If the physician makes several decisions within the same topic, temporality and the same turn of speech it becomes more difficult to hold decisions apart, which is why only one code should be given.

Physicians do a lot of such “information-packaging” to their patients, for example; starting a drug, deciding upon dosage, intervals, informing about effects and side effects, checking for interactions etc. Our tests in the development phase have shown that it is feasible to code this sequence as one code. Detailed assessment of sequences like the one described above is more suited for sub analysis.

An exception to this rule (one code per turn per topic) is when the physician refers to decisions in different time dimensions, for example:

− Physician: “We decided to put you on a drug that you’ll have to take four times a day” (past).

− Patient: “I think I will forget if I have to do it that often, do I have to?”

− Physician: “Ok, I think it is OK to have a double the dose morning and evening, reducing the frequency to twice a day (present).

− Patient: “But I always seem to get stomach pains when I take any kind of medication.”

− Physician: “Well, stomach pain is a possible side effect of this drug. If you get severe pain you should stop taking the pills.” (future).”

A second challenge was what to do when decisions were repeated (which was a very frequent situation). We solved that by agreeing (excerpt from the codebook): “…events where the physician restates decisions that have already been made or conveyed in the consultation. We have chosen to code repetitions because they are so frequent, and because they help us complete the identification of all decisions.”

Results

- I would like to see more parallel structure in the category descriptions. Perhaps a definition, example, nonexample pattern. As written, it is hard to follow the large categorical framework. Category 8 is especially hard to grasp the mutual exclusivity. Most likely, precautionary advice appears it could overlap with drug-related or treatment statements. Unclear why they are ordered as is. Provide some context for ordering, or state that it is not systematic.

Response to comment: Thank you for this comment. We have incorporated examples of statements that exemplify statements that DICTUM would NOT identify as decisions in the Methods section. Concerning mutual exclusivity we have elaborated on this in the Discussion section. Regarding the order of the categories we have clarified this in the first paragraph of the Result section.

Discussion

- Unclear how this helps patients sum up appointments. How are patients able to identify these points that required a team of four docs to parse out?

Response to comment: Thank you for this comment. We have elaborated on this in the Discussion.
Describe how you foresee this scheme being used in practice and in research. For instance, do you see value in exploring demographic differences in the categories enacted? Or specialty differences?
Response to comment: We have outlined some potential approaches with this taxonomy for practice and research and have due to reviewer 1's comment incorporated possibilities of exploring differences on specialty and demographic levels. We have explored specialty differences in the paper that describes all 380 encounters.
Why was it not tested in general practice? This may be a national/systematic difference that I don't perceive. Is general practice not like internal medicine outpatient clinic? Recommend including future research to include extending the coding application to primary care.
Response to comment: In Norway and most northern European health care systems internal medicine, both inpatient care and outpatient consultations take place within the hospital. We did not test the taxonomy in general practice, because our material was hospital encounters only. We are planning to test the taxonomy on encounters from general practice. It will be exciting to test its applicability and validity in this setting.

Reviewer: 2, Stephen G Henry, University of California Davis Medical School

This paper describes development of a novel system for coding patient-physician interactions that aims to identify and characterize all decisions within a specific patient-physician interaction. Such a coding system would have important potential for furthering health communication research. I commend the authors for undertaking this project and for the amount of work they have put into this project. Including the coding system as an appendix was very helpful. I do note several important limitations and problems, however, which I shall detail below with the aim of providing authors advice for improving their coding system (DICTUM) and paper.
1. A major potential contribution of DICTUM is to define what actually constitutes a decision. I agree with the authors that traditional decision-making research has focused very narrowly on one-off major decisions (eg whether to have knee surgery or not) and that a broader concept of "decision" is much needed. Developing and justifying this broader concept of "decision" is at least as important conceptually, and may be of more interest to readers, than the specifics of the 10 content categories. The authors should spend more space in the manuscript defining and justifying how they identified / defined a "decision." In particular their rationale for including both classic "decisions" and "judgements" could be clearer.

Response to comment: Thank you for this comment. We have tried to clarify and elaborate our rationale for including both action statements and judgement statements in the Methods section, Result section and Discussion of the manuscript.

2. Related to #1, the authors need to justify / explain why DICTUM dictates that only physician statements convey decisions. The title and intro are misleading. The definition of "decision" on page 7 requires that the statement be made by a medical expert." A truly comprehensive coding system would drop this restriction. A more accurate title for the current coding system would be "a novel taxonomy for physician decisions during patient-physician encounters." DICTUM makes no mention of patient statements. Suppose a physician asked, "Do you want to get an xray for your chest," and the patient replied, "Yes, I'd like an x-ray to help diagnose my problems." How is this patient statement not a decision? The discussion of patient-centered decision making in the introduction makes clear that patients drive decisions in some cases. Why omit such decisions from DICTUM? Physicians may voice most decisions, but by their nature decisions often emerge during patient-physician interactions, and a truly comprehensive coding system should reflect the fact that decisions are a product of

patient-clinician interactions and may be voiced by either patient or physician. In the absence of very convincing justification, the authors should make clear throughout that DICTUM only relates to physician decisions and is not a truly comprehensive coding system of decisions that occur during patient-clinician interactions. Such a coding system would still be potentially useful but would not truly capture and characterize all decisions during patient-clinician interactions, as the authors seem to promise. Alternatively, the authors could expand DICTUM to include patient statements.

Response to comment: Thank you for this comment. We agree with Reviewer 2 that framing the taxonomy as based on physician statements adds necessary precision. As a result of this comment we have altered the title to “What is a medical decision? A taxonomy based on physician statements in hospital encounters – a qualitative study” and have framed the manuscript accordingly.
With that said, patient input may affect physician decisions. E.g. Patient: “I want an x-ray of my chest” (no decision) Dr. “I think that is a good idea, I'll order an x-ray to be taken today” (decision).

3. The introduction does not provide a sufficiently convincing rationale for why a descriptive coding system like DICTUM is needed or how it has potential to advance health communication research. The discussion of patient-centered decision making should be curtailed, because it is not directly relevant for why DICTUM is important. There is a very compelling and convincing case for why a decision coding system such as DICTUM is needed, but I’m not convinced by the authors’ current rationale.

Response to comment: We are very grateful for these comments. We admit that, as SDM and patient-centered care are hot topics and we also take much interest in them, we hoped that an inclusion of these perspectives in the introduction would increase reviewers’ and readers’ interest in our study. Both reviewers’ comments have convinced us that the lack of an extensive decision taxonomy is a sufficient reason for our study. Consequently, we have changed the Introduction and Discussion of the paper to tone down the reference to SDM and patient-centeredness. Having said that, we still think that this taxonomy helps us see the complexity of medical encounters and could be used in future studies of how patients are included in decision-making.

4. The development and coding process for DICTUM is confusing and should be more clearly described. Authors should clarify the degree to which coding was based on video versus transcribed portions of video. The sampling process for DICTUM development should be more clearly explained. It seems that the authors used 30 internal medicine videos for the coding system development. If true, this is a limitation that should be justified. Shouldn’t a truly comprehensive system also draw on inpatient and surgical consultations? Especially since the authors clearly have videos involving many different types of physicians. Was the final coding system applied to all 300 visits, or will this be presented in a future paper?

Response to comment: We decided to start with encounters from internal medicine for several reasons. It was the specialty with the largest amount of encounters (n=120) and physicians (n=20) and with several encounters from all three different clinical settings (ER, WR and OP). Modern medicine evolves continuously. New drugs are frequently introduced. Guidelines and recommendations concerning use of tests, criteria for diagnoses and treatments are frequently updated. Because of this, the first author’s background as a practicing resident internist was considered an advantage. To establish the taxonomy’s applicability and to evaluate inter-operator variability, we selected four sets of five videos from different settings and specialties, with variation in age and gender in both patients and physicians, in order to ensure a maximum variation. In this phase no new categories emerged and no categories were irrelevant for other specialties. In the analysis of the entire material, which is presented in a separate paper, the categories were applicable and highly relevant to all the 17 specialties, all three settings and all 380 encounters.

5. The authors need to provide more detail and explanation for how they handled the issue of "unitizing reliability" in DICTUM. In other words, how did they define the unit of analysis for coding? What did they do, for example, when 1 coder identified a "decision" during a physician statement but the other coder did not? This is relevant because the measure of agreement the authors cite assumes that all reviewers code every "unit." It may be acceptable to resolve disagreements about the unit of analysis prior to assessing reliability (ie Krippendorf's alpha) but the authors should state this explicitly in the manuscript. For further discussion of this issue, see:
Kravitz, R. L., R. A. Bell, et al. (2002). "Characterizing patient requests and physician responses in office practice." Health Services Research 37(1): 217-238.

Response to comment: The unit of analysis – which we now have inserted in the Methods section - was physician statements conveying clinical decisions. See more on this in the response to Reviewer 1's comment on unit of analysis.
All four coders watched the mentioned four sets of five videos from back to back. We needed a statistical method that allowed the comparison between several coders. Klaus Krippendorff's α-agreement for coding allows for the comparison of many observers, many nominal categories and missing values (Krippendorff 2004, pp. 230-236) As Krippendorf's α-agreement for multiple coders states; 'if one or more coders miss a code, this makes the alpha lower in value.' We have elaborated on this in the Methods section.
Getting four coders to a sufficient level of reliability trying to identify somewhere between 10-20 statements out of several hundred statements per encounter, was a challenge which we worked with for three rounds, before we decided to do a final round to assess reliability. Our result (alpha=0.79) was 0.01 below the threshold of reliability set by Krippendorff and we could be criticized for not repeating our IRR-testing until we got above the 0.8 mark. After some discussion we decided that doing a new IRR-test before applying the taxonomy and its codebook to our material, was not necessary. There were several reasons for this and the most important being that only two of the coders would proceed as coders in the coding analysis of the whole material. Investing further effort in getting all four of us up to a higher level of calibration seemed futile. We decided it was more important to focus on consistency of coding precision during the analysis of the entire material.

6. Description of the 10 topical codes could be clearer. For one thing, some of the category labels are quite awkward. Based on the coding manual, "procedure-related" seems more accurate than "surgery-related." Several other topic headings, including "legally related" and "contact-related" are awkward. For a second issue, the authors should include an explicit definition of each category in the manuscript text related to that topic, something along the lines of "This category includes decisions that involve…" The lack of clear category definitions in the manuscript renders many of the topic descriptions confusing and unclear.

Response to comment: Thank you for these comments. Category 5 is a category that we have struggled with naming. We called it "Intervention" first and then "Non-pharmaceutical intervention", but were advised to change it as it might be associated with alternative medicine. Surgery-related is not precise enough either, and we have agreed on altering it to "therapeutic procedure-related" (as opposed to diagnostic procedure-related, which is incorporated in category 1, ref reviewer 2's final comment) - and have adjusted the text in the Result section, Table 2 and taxonomy accordingly.
Concerning contact-related; this category has not been up for discussion, as it says something about the future contact with the health care services. We could call it admit/discharge/follow-up/referral, but we argue that contact-related sums it up sufficiently.
Concerning legally related; in a government funded health care system, this category makes sense and we can see that in an insurance-based health care system it does not describe what we mean sufficiently. We have extended the category name to "Legally and insurance-related" and have adjusted the text in the Result section, Table 2 and taxonomy accordingly.
Secondly, upon Reviewer 2's request – we have included a sentence at the start of each

paragraph/category in the Results section stating what is included by each decision category.

7. The coding system itself includes a detailed discussion of the temporality of decisions (past, present, future). This is an important aspect of decisions that is rarely discussed or even acknowledged, and is a strength of the coding system. Briefly discussing in the manuscript text how /why the authors decided to address temporality would likely be of interest to readers.

Response to comment: Thanks for this comment. We have already published a paper on the temporality of decisions (Ofstad EH, Frich JC, Schei E, Frankel RM, Gulbrandsen P. Temporal characteristics of decisions in hospital encounters: A threshold for shared decision making? A qualitative study. Patient Educ Couns. 2014 Nov;97(2):216-22.). The complexity (both a topical and a temporal dimension) of the classification system led us to go for two separate papers, because of the word limit restriction of most journals. We referred to this previous paper in the Methods section.

8. If the authors decide to revise this manuscript, its impact and clarity would be increased by being carefully edited to make sure that the concepts are clear and easy to understand for readers. Although the general thrust of the authors' reasoning is clear, sentence clarity could be notably improved in many areas.

Response to comment: Thank you for your comment. We have made efforts to increase the clarity of sentences throughout the manuscript. All alterations done in the revised submission are marked with blue letters throughout the manuscript.

MINOR COMMENTS

9. Statements about the importance of adding a social psychologist to the team should be deleted. While helpful, inclusion of non-physicians in coding system development is not novel or worth mentioning (other than perhaps with a single sentence in the methods section).

Response to comment: We have kept the sentence in the Methods section where we report the inclusion of Richard Frankel. We have deleted the sentence implying it might have strengthened our study to include him from the Strengths and limitations heading and from the Discussion section, as Reviewer 2 requests.

10. The authors should provide a brief explanation of Krippendorf's alpha. Many readers will not be familiar with this measure.

Response to comment: We have elaborated on why we chose Krippendorf's alpha for the reliability assessment in the Method section.

11. I question the authors' statement on page 16 that blames the ethics committee for the lack of patient involvement in coding system development. The authors surely could have included patients in the development process if they had so desired – these need not be the same patients as were video recorded. Failure to include patients on the research team is not a major limitaton; this limitation shoud be deleted.

Response to comment: We have deleted the sentence implying it might have limited our study not to have had patient input from the Strengths and limitations heading and from the Discussion section, as Reviewer 2 requests.

12. Table 1 formatting should be improved.

Response to comment: We agree. Thank you. The first bar has been deleted, the long bulks of text describing the specialties/subspecialties lumped into the categories Internal Medicine and Surgical disciplines, has been moved outside the table using stars. And on Reviewer 1's request we have listed the proportion of the different specialties by number of encounters, and not number of physicians, per specialty category.

13. I am not sure that the text ever references table 2. In addition the text does not clearly discuss the different subcategories and why the authors included them. Without Table 2 or a detailed review of the coding manual, readers might not realize that DICTUM includes subcategories. Authors should make sure that use of subcategories is discussed in the manuscript text.

Response to comment: The Results section starts out by referring to Table 2. Thank you for the comment about subcategories, we have added a column outlining the different subcategories in Table 2.

14. In the coding manual, the authors give "I think we should get a liver biopsy" as an example of "information gathering. Why is this statement not coded as surgery-related, since it involes recommendation of invasive procedure?" The inclusion of such "judgements" (which are certainly not "decisions) in the coding system needs a stronger justification. What would happen if the patient replied, "I don't want a liver biopsy." ? Would that still be coded as a decision?

Response to comment: Thank you for these thoughtful questions. First, the decision to get a liver biopsy would be coded as "Gathering additional information", because the purpose – no matter how invasive the test – is to gather additional information about what is the cause of the patient's liver malfunction. Because of reviewer 2's comments we have clarified what is comprised by category 5 (surgery-related) by accepting reviewer 2's suggestion to call it procedure-related – and relevant to the question in hand, we have renamed it "therapeutic procedure-related", more specifically. Second, due to Reviewer 2's comment about whether the taxonomy comprises both patient and physician decisions, we concur with Reviewer 2 that framing our taxonomy as based on physician statements to be a more precise presentation. So if the patient says "I don't want a liver biopsy" we do not code this as a decision. E.g.: Dr. "You should take this liver biopsy." (decision). Patient: "No, I don't want to." (no decision). "OK, so we'll postpone it for now and see how it goes" (new decision). Third, is a recommendation of an invasive procedure a judgment statement? Or an action statement? This could be discussed, but it is not really the point. Every time a physician makes a judgment statement, the physician has options which call for a decision. If recommending a liver biopsy is a judgment statement, a relevant alternative will always be to not recommend it. It is an invasive procedure, not without potential complications, such a judgment statement is made as a result of a decision-making process, in this case going on inside the physician's mind. When the physician's conclusion is expressed to the patient in form of a statement, the taxonomy codes it as a decision.

VERSION 2 – REVIEW

REVIEWER	Christy Ledford Uniformed Services University of the Health Sciences, USA
REVIEW RETURNED	21-Dec-2015

GENERAL COMMENTS	Thank you to the authors for considering and addressing our reviewer recommendations. The manuscript is much improved. I have only minor comments moving forward. Would like to see a stronger application statement to end the abstract. The last sentence sounds nebulous for all of the hard work

	you've put into this project. One category's name does not appear grammatically correct. It should be "legal and insurance-related." Para at p15 ln 41, beginning "frequently used clinical guidelines" and the next para appear out of place. And even may be just the kernel of an idea. These paragraphs need a transition of more development to understand how the authors are proposing that the taxonomy will be used in guideline decisions and conversations. Include a comment in limitations about the Krippendorf. That could be stronger.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 1
Christy Ledford
Uniformed Services University of the Health Sciences, USA

Would like to see a stronger application statement to end the abstract. The last sentence sounds nebulous for all of the hard work you've put into this project.

Response to comment: We have altered the final sentence of the abstract to provide a stronger application statement.

One category's name does not appear grammatically correct. It should be "legal and insurance-related."

Response to comment: We have altered the name of the category as recommended by the reviewer in both the manuscript, the tables and in the appended codebook.

Para at p15 ln 41, beginning "frequently used clinical guidelines" and the next para appear out of place.
And even may be just the kernel of an idea. These paragraphs need a transition of more development to understand how the authors are proposing that the taxonomy will be used in guideline decisions and conversations.

Response to comment: We have deleted the two paragraphs concerning guidelines. They were written when the manuscript was framed more towards shared decision-making and patient involvement.

Include a comment in limitations about the Krippendorf. That could be stronger

Response to comment: We have included a statement about Krippendorff in the limitations paragraph in the Discussion.