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)	The Gut Feelings Questionnaire in daily practice: a feasibility study using a mixed-methods approach in three European countries
AUTHORS	Barais, Marie; van de Wiel, Margje; Groell, Nicolas; Dany, Antoine; Montier, Tristan; Van Royen, Paul; Stolper, Erik

VERSION 1 – REVIEW

REVIEWER	Sophie Turnbull University of Bristol
REVIEW RETURNED	04-May-2018

GENERAL COMMENTS	This paper explores several stages of the further development and validation of a questionnaire designed to capture GP gut feeling. GP gut feeling and the role it plays in decision making is an important research area and still largely unknown. The systematic attempt to capture the phenomenon is valid. I have a few reflections that I hope will be helpful to the authors: Overall: The need for this questionnaire over a well explained, low burden, binary yes/no question of whether the clinician has a gut feeling has not been covered. Previous studies have reported clinician gut feeling on this basis and I cannot see a convincing argument for why this questionnaire improves on this. Although, the authors are right in saying that previous studies have not defined what they mean by 'gut feeling' clearly to the study participants. Have the study team considered a comparison of the performance of their binary Qu 1 with the full questionnaire? We have found light touch approaches often work better when engaging GPs in research- I would want to see some justification for why this approach is more beneficial for capturing gut feeling. They would be more likely to report a yes/no answer after the consultation so would this be better at measuring the outcome than a detailed questionnaire completed at the end of the day? I can see value in the questionnaire providing more detail about how gut feeling may manifest i.e '…uneasy feeling because I am worried about the potentially unfavourable outcome'- perhaps that case could be stated more strongly. Have the authors also considered a comparison of the response to Qu1 and Qu11? As highlighted in the think aloud section of the paper, having the question about gut feeling at the end of the questionnaire risks the response capturing rational decision making rather than an intuitive gut feeling. How have you justified including the question at the end after finding this as a potential threat to validity? Introduction Why is a full questionnaire required?

	Methods L125 Please provide a definition of 'experienced GPs'- was it years since qualification? L178 Agreed- putting it last involves reasoning rather than intuition- what does the rest of the questionnaire therefore add on top of this? Is it to better understand the phenomenon- If so please cover this in the introduction? What is the purpose/context this questionnaire would be used? L191 Did you look at experience of GPs? Years since qualification were found to be associated with reporting gut feeling https://bmcfampract.biomedcentral.com/articles/10.1186/s12875-018-0716-7 Feasibility study 1: Methods: L204 How long after the 2 week period were they interviewed? L220-228 How did you decide that high scores on other questions would override the clinician's decision that it was not applicable? Results L 237 How do you know most consultations were captured? Was this done by comparing the number of gut feeling questionnaires to the number of consultations logged/ clinician reported? L 292 There are missing cases here- if 336 cases were analysable why are there only 319 reported for the gut feeling outcome? 77+242=319 24% gut feeling cases rather than 23% Feasibility 2: L 305 If it was the same procedure were they asked to use the GFQ it on every new consultation again? Because L 330 the GP indicates that they only use it on the first consultation of the day. L 314 Again when were the interviews conducted L 333 Issues with recall bias L 387 Depending on the reported gut feeling for the missing cases the discrepancy between reported gut feeling in feasibility 1 and 2 the Chi2 difference may reach the significance threshold of <0.05. Regardless there is an indication of a difference here whether it reaches the arbitrary threshold or not. Can you provide more information on where this difference has come from? Is it because the clinicians better understood the question in feasibility 2? L 444 There is the potential for serious issues around recall bias here. They maybe able to logically recall details of the case, but it is quite possible that the intuitive response that this questionnaire is aiming to capture is lost by this point. You are therefore at risk of measuring rationed decision making. L 447 But similar to a previous study finding 20% of clinicians reported gut feeling after seeing children with RTI and cough. https://bmcfampract.biomedcentral.com/articles/10.1186/s12875-018-0716-7 L445- Reference in a different format to the rest of the document

REVIEWER	Sandra Monteiro McMaster University, Canada
REVIEW RETURNED	27-May-2018

GENERAL COMMENTS	I reviewed a manuscript describing a study aimed at validating the gut feelings questionnaire for use in clinical practice. The authors conducted a 2-stage validation study, first to get feedback on the items and second to evaluate the construct of gut feelings. For the stated purpose the authors have done a good job. The PCA seems appropriate for evaluating a construct that is self-reported. But I feel the discussion could be strengthened. The

	authors report a lot of qualitative data but do not comment on it in the discussion. I feel without some elaboration on the potential uses of the GFQ there could be confusion regarding whether this is appropriate for use with novices compared to experts. Unless we know how the self-report of alarm is connected to patient outcomes or management decisions, I am concerned this study could be interpreted to suggest there is value in measuring gut feelings. While the authors are building on previous work to validate this questionnaire and have conducted their study well, I am not yet convinced that the questionnaire itself has value. what is the issue in medicine or medical education that measuring gut feelings would solve?

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Reviewer Name: Sophie Turnbull
Institution and Country: University of Bristol

This paper explores several stages of the further development and validation of a questionnaire designed to capture GP gut feeling. GP gut feeling and the role it plays in decision making is an important research area and still largely unknown. The systematic attempt to capture the phenomenon is valid. I have a few reflections that I hope will be helpful to the authors:
Overall:
The need for this questionnaire over a well explained, low burden, binary yes/no question of whether the clinician has a gut feeling has not been covered. Previous studies have reported clinician gut feeling on this basis and I cannot see a convincing argument for why this questionnaire improves on this. Although, the authors are right in saying that previous studies have not defined what they mean by 'gut feeling' clearly to the study participants. Have the study team considered a comparison of the performance of their binary Qu 1 with the full questionnaire? We have found light touch approaches often work better when engaging GPs in research- I would want to see some justification for why this approach is more beneficial for capturing gut feeling. They would be more likely to report a yes/no answer after the consultation so would this be better at measuring the outcome than a detailed questionnaire completed at the end of the day? I can see value in the questionnaire providing more detail about how gut feeling may manifest i.e '…uneasy feeling because I am worried about the potentially unfavourable outcome'- perhaps that case could be stated more strongly.

We thank you for this comment which gave us the opportunity to clarify what appeared to be unclear in the introduction. We added the following sentences when describing the gut feelings questionnaire: Line 104: "The objective of the questionnaire was to determine the presence, or absence, of gut feelings in GPs' diagnostic reasoning, at the end of a consultation based on a clear consistent definition of the concept in order to avoid response bias. This questionnaire measures not only whether a gut feeling is present, (i.e. not just by a yes or no response, as is mostly done in clinical studies about gut feelings [5–7]), but also differentiates between the sense of reassurance and the sense of alarm by more precise statements reflecting the outcomes of the diagnostic reasoning process. In validating the questionnaire, a Principal Component Analysis (PCA) showed one component explaining 70.2% of the total variance with the sense of alarm and the sense of reassurance as opposites."

We also discussed the added value of the gut feelings questionnaire in comparison to other measures of gut feelings in the discussion:
Line 503: “Several studies measured gut feelings with other definitions than the one we used here. For instance, Turnbull et al used in their questionnaire “my gut feeling is “something is wrong””: yes or no, whereas gut feelings were explained in the instruction booklet as “gut feeling that the child’s illness may be more serious than is superficially apparent” [7]. Several other studies measuring gut feelings do miss a detailed and accurate definition of the sense of alarm [18,19]. In the study measuring the predictive value of gut feelings for serious infections in children [18], a precise definition of what was considered as “gut feelings” or ‘instinct’ is not available. In another study regarding the recognition of sepsis in primary care [19], the authors did not give details of the concept or definition to which they were referring when using the expression “gut feeling”. In the questionnaire they used, one item was “How important were the following patient assessment aspects in the decision to refer?” “A gut feeling” was one of the possible choices. In our study we measured gut feelings more accurately. The concept of gut feelings in a closed question is not clear enough, and allows for differences in interpretation of different participants, especially within different languages and cultures. The sense of alarm and the sense of reassurance, as they were defined by Stolper et al., were considered, after linguistic validation procedures, as a transcultural concept validated in four languages [5,6]. Using such different definitions and measures of gut feelings in different contexts, it is not possible to compare the prevalence of the sense of alarm in different studies. The use of the GFQ is a uniform way of measuring the sense of alarm when diagnosing patients in primary care and to determine its prevalence.”

Have the authors also considered a comparison of the response to Qu1 and Qu11? As highlighted in the think aloud section of the paper, having the question about gut feeling at the end of the questionnaire risks the response capturing rational decision making rather than an intuitive gut feeling. How have you justified including the question at the end after finding this as a potential threat to validity?

Thank you for this comment. We explained that we started with item 1 to capture the presence of gut feelings immediately after the consultation. We emphasized now that some participating GPs mentioned they were not always able to answer item 1 in the beginning of the questionnaire. For this reason, we repeated item 1 as item 11. We only used item 11, when participants did not answer item 1.
To clarify this, we have added the following sentences in the results of the think aloud study:
Line 259: “We also proposed repeating this item at the end of the list for those participants who were not able to answer this question at the beginning. It was only for those participants who did not answer item 1 that we used item 11 as the indication of the presence or absence of a GF.”

In the results section of the second feasibility study we now better explained how to instruct participants:
Line 419: “In order to minimize rationalizations afterwards we also emphasized to immediately fill in the questionnaire to grasp GPs’ experience during the diagnostic process (preventing recall bias) for each patient that needs to be included in the study (preventing selection bias). The instructions should be embedded within the context and aim of any study [11]. (Text box 1). In this particular study we specified to fill in the questionnaire for the first consultation of the day with an adult patient, aged over 18 years, with a new reason for a consultation.”
In the text box we now phrased this instruction as follows:
“In order to avoid selection bias and to reflect your experience during the diagnostic process, we urgently ask you to fill in the questionnaire for each patient that needs to be included in the study directly after the consultation.”

And in the discussion section we also addressed this issue:

Line 463: “We started with the item asking for the presence of a gut feeling in diagnostic reasoning to capture their experience immediately after the consultation: the first item is now “Please indicate what kind of gut feeling you have at the end of the consultation”. We repeated this item at the end of the questionnaire for those participants who were not able to answer this question at the beginning. It was only for participants who did not answer item 1 that we used item 11 as the indication of the presence or absence of a GF. There might be a risk that the last group will also use their analytical reasoning in finding an answer to item 11 but, in any case, we reduced that risk by also putting the question at the top of the questionnaire. To minimize rationalizations afterwards we emphasized in the instructions to immediately fill in the questionnaire to better grasp GPs experience during the diagnostic process.”

Introduction
Why is a full questionnaire required?

As mentioned previously, we added in the introduction section the main objective of the questionnaire and its added value:
Line 104: “The objective of the questionnaire was to determine the presence, or absence, of gut feelings in GPs’ diagnostic reasoning, at the end of a consultation based on a clear consistent definition of the concept in order to avoid response bias. This questionnaire measures not only whether a gut feeling is present, (i.e. not just by a yes or no response, as is mostly done in clinical studies about gut feelings [5–7]), but also differentiates between the sense of reassurance and the sense of alarm by more precise statements reflecting the outcomes of the diagnostic reasoning process.”

Methods

L125 Please provide a definition of 'experienced GPs'- was it years since qualification?

Thank you for this comment, we added
Line 140: “Participants were 8 experienced GPs (7 female; average experience in GP practice was 18,6 years, ranging from 6 to 29 years), 8 first-year GP-trainees (5 female; average clinical experience before their traineeship was 24,5 months, ranging from 9 to 53 months) and 8 advanced medical students (7 female) doing their internship in general practice at Maastricht University.”

L178 Agreed- putting it last involves reasoning rather than intuition- what does the rest of the questionnaire therefore add on top of this? Is it to better understand the phenomenon- If so please cover this in the introduction? What is the purpose/context this questionnaire would be used?

Thank you for this other formulation of what was lacking in our introduction. We are grateful that we now have better explained in the introduction and discussion why we think the GFQ using more than 1 item is useful and has an added value. Please see our previous comments on this issue.

L191 Did you look at experience of GPs? Years since qualification were found to be associated with reporting gut feeling https://bmcfampract.biomedcentral.com/articles/10.1186/s12875-018-0716-7

Thank you for suggesting this article. The role of experience in experiencing gut feelings in diagnostic reasoning is indeed very interesting. We will examine the differences between the three experience groups in diagnostic reasoning and reporting gut feelings in a later phase. For the purpose of the present study we only analysed the think-aloud protocols of half of the participants in this study while they filled in the questionnaire. Our aim was to get detailed insight into how participants processed the questionnaire items in order to improve the formulation. This was an important first step in increasing future participants’ understanding of the questionnaire items in line with their objectives. We reached data saturation with half of the participants taking part in the original study.

In the discussion we addressed the possible effect of experience on diagnostic reasoning and the use of our questionnaire to study this relationship:
Line 532: “GPs’ experience is probably positively related to the prognosis of serious diseases: the greater the experience, the more likely it is that GPs’ predictions will be correct [7]. The GFQ can be used to study this relationship further. In the area of education, how both the sense of alarm and the sense of reassurance play a role in decision making should be addressed as an important non-analytical track of diagnostic reasoning, especially in general practice [2]. However, insight into the way gut feelings are used, and the role of experience, should be refined through further studies.”

Feasibility study 1:

Methods:

L204 How long after the 2 week period were they interviewed?

We added these details:
Line 185: “Most interviews were held within two weeks of the two-week period and were audio-recorded and transcribed verbatim. The interviews lasted between 3 and 18 minutes.”

L220-228 How did you decide that high scores on other questions would override the clinician's decision that it was not applicable?
Results

Thank you for this question. We have chosen to override the clinicians’ decision that it was not applicable as we valued the answers to the gut feelings items following the decisions rules which we added in the text:
Line 196: “The quantitative data, i.e. the answers to items 1-5 and 10 (Figure 1), were analysed with a Chi² test using specific criteria. A sense of alarm was considered as present when the answer to item 10 indicated a sense of alarm or when the answer to item 10 indicated that it was not applicable and at least one of the scores for items 2 to 5 was higher than 3/5. A sense of reassurance was considered as present when the answer to item 10 indicated a sense of reassurance or when the answer to item 10 indicated that it was not applicable and the score for item 1 was higher than 3/5. Gut feelings were considered absent when the answer to item 10 indicated that it was not applicable and none of the scores for items 2-5 was higher than 3/5 and the score for item 1 was lower than 4/5. These cut- off criteria were chosen in line with the study protocol of the study on the accuracy of the sense of alarm when faced with chest pain and dyspnoea [10].

L 237 How do you know most consultations were captured? Was this done by comparing the number of gut feeling questionnaires to the number of consultations logged/ clinician reported?

We do apologize for not including this important information: we did not include in the method section of the first feasibility study which cases we had asked our participants to include. This oversight made the method incomplete. We added this detail on line 177.
“They were asked to include only the first consultation of the day with an adult patient, aged over 18 years, with a new reason for a consultation”.

L 292 There are missing cases here- if 336 cases were analysable why are there only 319 reported for the gut feeling outcome? 77+242=319 24% gut feeling cases rather than 23%

Thank you for this comment. We have added the ones where no gut feeling was registered (17 or 5 %):

Line 329: “In total, 77 (23%) were concerned with a sense of alarm, and 242 (723%) with a sense of reassurance and there were 17 (5%) where no gut feeling was applicable.”

Feasibility 2:
L 305 If it was the same procedure were they asked to use the GFQ it on every new consultation again? Because L 330 the GP indicates that they only use it on the first consultation of the day.

We do agree with this comment, we made a mistake as we forgot to include that it was the first consultation of the day as in the first feasibility study. We corrected line 177: “They were asked to include only the first consultation of the day with an adult patient, aged over 18 years, with a new reason for a consultation”. Also, in feasibility study 2 we included the point that the same procedures were used as in the first feasibility study.
We also clarified this in the instructions (Text box 1): “In order to avoid selection bias and to reflect your experience during the diagnostic process, we urgently ask you to fill in the questionnaire for each patient that needs to be included in the study directly after the consultation.””

L 314 Again when were the interviews conducted

We added this detail:
Line 218: “Most interviews were held within two weeks of the two-week period and lasted 5-30 minutes.”

L 333 Issues with recall bias

In order to reduce recall bias – we better explained how to instruct participants before filling in the GFQ:
Line 419: “In order to minimize rationalizations afterwards we also emphasized to immediately fill in the questionnaire to grasp GPs’ experience during the diagnostic process (preventing recall bias) for each patient that needs to be included in the study (preventing selection bias). The instructions should be embedded within the context and aim of any study [11]. (Text box 1). In this particular study we specified to fill in the questionnaire for the first consultation of the day with an adult patient, aged over 18 years, with a new reason for a consultation.”

L 387 Depending on the reported gut feeling for the missing cases the discrepancy between reported gut feeling in feasibility 1 and 2 the Chi2 difference may reach the significance threshold of <0.05. Regardless there is an indication of a difference here whether it reaches the arbitrary threshold or not. Can you provide more information on where this difference has come from? Is it because the clinicians better understood the question in feasibility 2?

Thank you for this comment. The missing cases were cases where data were missing and could not, therefore, be interpreted. We added these sentences in the discussion:
Line 474: “The prevalence of the sense of alarm seemed to be higher in the second feasibility study than in the first one (23% vs 31%), but statistically there is no difference. Both studies took place in winter, with the same incidence of diseases. It can be an accidental finding which is confirmed by the fact that there is no statistical difference. Further studies with the GFQ in clinical practice are needed to examine the prevalence of gut feelings in general practice and its predictive validity in different contexts.”

L 444 There is the potential for serious issues around recall bias here. They maybe able to logically recall details of the case, but it is quite possible that the intuitive response that this questionnaire is aiming to capture is lost by this point. You are therefore at risk of measuring rationed decision making.

We added the following sentences in the instructions:
Line 419: “In order to minimize rationalizations afterwards we also emphasized to immediately fill in the questionnaire to grasp GPs’ experience during the diagnostic process (preventing recall bias) for each patient that needs to be included in the study (preventing selection bias). The instructions should be embedded within the context and aim of any study [11]. (Text box 1). In this particular study we specified to fill in the questionnaire for the first consultation of the day with an adult patient, aged over 18 years, with a new reason for a consultation.”

In the discussion:
Line 486: “Quite a number of the GPs failed to follow the full instructions given prior to the feasibility studies. It did not always appear to be feasible to fill in a questionnaire right after a consultation. These GPs mentioned, however, that when responding to all the items, they were able to recapitulate the information regarding the patients involved without any problems. None of them mentioned that it could have induced a recall effect. The prevalence of the sense of alarm, of 23% and 31% found here, was higher than the prevalence of 8.2% found in an earlier study [19]. The difference with our study might be explained by a selection bias in terms of patients. We have highlighted this point for attention in the instructions (See Text box 1).”

In the text box 1: “In order to avoid selection bias and to reflect your experience during the diagnostic process, we urgently ask you to fill in the questionnaire for each patient that needs to be included in the study directly after the consultation. Please, read the questionnaire, so we can discuss any questions you might have.
”
L 447 But similar to a previous study finding 20% of clinicians reported gut feeling after seeing children with RTI and cough. https://bmcfampract.biomedcentral.com/articles/10.1186/s12875-018-0716-7

Thank you very much for this suggestion that we integrated in the following paragraph in the discussion:
Line 515: “In our study we measured gut feelings more accurately. The concept of gut feelings in a closed question is not clear enough, and allows for differences in interpretation of different participants, especially within different languages and cultures. The sense of alarm and the sense of reassurance, as they were defined by Stolper et al., were considered, after linguistic validation procedures, as a transcultural concept validated in four languages [5,6].
Using such different definitions and measures of gut feelings in different contexts, it is not possible to compare the prevalence of the sense of alarm in different studies. The use of the GFQ is a uniform way of measuring the sense of alarm when diagnosing patients in primary care and to determine its prevalence.”

L445- Reference in a different format to the rest of the document
Thank you for this remark; we have modified the references to match the format of the rest of the document.

Reviewer : 2
Reviewer Name: Sandra Monteiro
Institution and Country: McMaster University, Canada
Please state any competing interests or state ‘None declared’: None declared

Please leave your comments for the authors below

I reviewed a manuscript describing a study aimed at validating the gut feelings questionnaire for use in clinical practice. The authors conducted a 2-stage validation study, first to get feedback on the items and second to evaluate the construct of gut feelings.
For the stated purpose the authors have done a good job. The PCA seems appropriate for evaluating a construct that is self-reported. But I feel the discussion could be strengthened. The authors report a lot of qualitative data but do not comment on it in the discussion. I feel without some elaboration on the potential uses of the GFQ there could be confusion regarding whether this is appropriate for use with novices compared to experts. Unless we know how the self-report of alarm is connected to patient outcomes or management decisions, I am concerned this study could be interpreted to suggest there is value in measuring gut feelings. While the authors are building on previous work to validate this questionnaire and have conducted their study well, I am not yet convinced that the questionnaire itself has value.

what is the issue in medicine or medical education that measuring gut feelings would solve?

Thank you for this comment. We included in the introduction and in the discussion several sentences on the added value of the gut feelings questionnaire to follow your recommendations.
In the introduction we have added the following sentences:
Line 104: “The objective of the questionnaire was to determine the presence, or absence, of gut feelings in GPs’ diagnostic reasoning, at the end of a consultation based on a clear consistent definition of the concept in order to avoid response bias. This questionnaire measures not only whether a gut feeling is present, (i.e. not just by a yes or no response, as is mostly done in clinical studies about gut feelings [5–7]), but also differentiates between the sense of reassurance and the sense of alarm by more precise statements reflecting the outcomes of the diagnostic reasoning process. In validating the questionnaire, a Principal Component Analysis (PCA) showed one component explaining 70.2% of the total variance with the sense of alarm and the sense of reassurance as opposites.”

In the discussion we have added the following new paragraphs to discuss this issue:
Line 503: “Several studies measured gut feelings with other definitions than the one we used here. For instance, Turnbull et al used in their questionnaire “my gut feeling is “something is wrong””: yes or no, whereas gut feelings were explained in the instruction booklet as “gut feeling that the child’s illness may be more serious than is superficially apparent” [7]. Several other studies measuring gut feelings do miss a detailed and accurate definition of the sense of alarm [18,19]. In the study measuring the predictive value of gut feelings for serious infections in children [18], a precise definition of what was considered as “gut feelings” or ‘instinct’ is not available. In another study regarding the recognition of sepsis in primary care [19], the authors did not give details of the concept or definition to which they were referring when using the expression “gut feeling”. In the questionnaire they used, one item was “How important were the following patient assessment aspects in the decision to refer?” “A gut feeling” was one of the possible choices. In our study we measured gut feelings more accurately. The concept of gut feelings in a closed question is not clear enough, and allows for differences in interpretation of different participants, especially within different languages and cultures. The sense of alarm and the sense of reassurance, as they were defined by Stolper et al., were considered, after linguistic validation procedures, as a transcultural concept validated in four languages [5,6].
Using such different definitions and measures of gut feelings in different contexts, it is not possible to compare the prevalence of the sense of alarm in different studies. The use of the GFQ is a uniform way of measuring the sense of alarm when diagnosing patients in primary care and to determine its prevalence.”

VERSION 2 – REVIEW

REVIEWER	Sophie Turnbull University of Bristol, UK
REVIEW RETURNED	27-Jul-2018

GENERAL COMMENTS	This revision is much clearer, but I think the introduction could be made clearer. Particularly paragraph 2. I found myself having to read it several times to understand what the authors were trying to convey. Second para of the intro from L100 is unclear- What do the focus groups have to do with the questionnaire? Did it inform it? Please clearly express what the paragraph is conveying in the first line. Something like: A survey was created to capture gut feeling…..Definitions of were informed by a focus group study….. L106- I'm not clear on why the questionnaire reduces response bias? I think what it does that sets it apart from the yes/no is to provide detail about the reasons why the GP had a gut feeling. You do sort of cover this L108-110- I think it could be stated more clearly. The way it has been covered in the discussion L523-526 is much clearer. In terms of your added section L511 you discuss the definitions of the other two papers on gut feelings, but do not discuss Van der Bruel's paper. They did provide a precise definition- it would be worth discussing what value your paper adds over this. Text box: I don't know if I have understood your response correctly, but it sounds as it you have retrospectively altered the wording of the instructions you gave your participants(below)? I can't see text box 1 so I don't know if this is the case. Perhaps you mean future studies should include the following instructions? "In order to avoid selection bias and to reflect your experience during the diagnostic process, we urgently ask you to fill in the questionnaire for each patient that needs to be included in the study directly after the consultation." L532: The referenced paper did not find 'GPs' experience is probably positively related to the prognosis of serious diseases: the greater the experience, the more likely it is that GPs' predictions will be correct'. It found that clinicians are more likely to report having a gut feeling. Also that gut feeling did not predict poorer prognosis. There was no link made between experience and predicting prognosis

VERSION 2 – AUTHOR RESPONSE

Reviewer Name: Sophie Turnbull

Institution and Country: University of Bristol, UK

This revision is much clearer, but I think the introduction could be made clearer. Particularly paragraph 2. I found myself having to read it several times to understand what the authors were trying to convey.

Second para of the intro from L100 is unclear- What do the focus groups have to do with the questionnaire? Did it inform it? Please clearly express what the paragraph is conveying in the first line. Something like: A survey was created to capture gut feeling…..Definitions of were informed by a focus group study…..

Thank you for this comment. We rephrased the second paragraph and stated more clearly the role of the first qualitative step.

Line 91: “In studies earlier done, tThe sense of alarm and the sense of reassurance were defined following a qualitative analysis of the text of several focus groups on the topic and a subsequent Delphi consensus procedure [1,3]. The items of a Dutch Gut Feelings Questionnaire (GFQ) were based on these definition criteria.”

L106- I’m not clear on why the questionnaire reduces response bias? I think what it does that sets it apart from the yes/no is to provide detail about the reasons why the GP had a gut feeling. You do sort of cover this L108-110- I think it could be stated more clearly. The way it has been covered in the discussion L523-526 is much clearer.

We modified this part as following:

Line 96:” The objectives of the questionnaire were to determine the presence, or absence, of gut feelings in GPs’ diagnostic reasoning at the end of a consultation based on a clear consistent definition of the concept in order.”

In terms of your added section L511 you discuss the definitions of the other two papers on gut feelings, but do not discuss Van der Bruel’s paper. They did provide a precise definition- it would be worth discussing what value your paper adds over this.

Thank you for this comment, we modified the paragraph and added discussion on the definition on gut feeling used by Van Den Bruel in her study:

Line 495: “Several other studies measuring gut feelings do miss a detailed and accurate definition of the sense of alarm [5,6]. In a study regarding the recognition of sepsis in primary care [6], the authors did not give details of the concept or definition to which they were referring when using the expression “gut feeling”. In the questionnaire they used, one item was “How important were the following patient assessment aspects in the decision to refer?” “A gut feeling” was one of the possible choices. In our study we measured gut feelings more accurately. The concept of gut feelings in a closed question is not clear enough, and allows for differences in interpretation of different participants, especially within different languages and cultures. The sense of alarm and the sense of reassurance, as they were defined by Stolper et al., were considered, after linguistic validation procedures, as a transcultural concept validated in four languages [1,8].

In a study measuring the predictive value of gut feelings for serious infections in children [5], the gut feeling was defined as “an intuitive feeling that something was wrong even if the clinician was unsure why”. The word “intuitive” could be a source of misunderstanding as it covers several concepts in cognitive science which sometimes overlap because of different levels of abstraction [17-19]. Using only this term in research into diagnostic reasoning might be a source of confusion for the participating GPs. We wanted to avoid this possible bias by using well-defined descriptions of the intuitive sense of alarm and of reassurance in our questionnaire.”

Text box: I don’t know if I have understood your response correctly, but it sounds as it you have retrospectively altered the wording of the instructions you gave your participants(below)? I can’t see text box 1 so I don’t know if this is the case. Perhaps you mean future studies should include the following instructions?

“In order to avoid selection bias and to reflect your experience during the diagnostic process, we urgently ask you to fill in the questionnaire for each patient that needs to be included in the study directly after the consultation.”

We do agree with this remark and added the instructions were modified for future studies:

Line 420: “The instructions should be embedded within the context and aim of any future study [11] (Text box 1).”

Line 489: “We have highlighted this point for attention in the instructions for future studies”.

L532: The referenced paper did not find ‘GPs’ experience is probably positively related to the prognosis of serious diseases: the greater the experience, the more likely it is that GPs’ predictions will be correct’. It found that clinicians are more likely to report having a gut feeling. Also that gut feeling did not predict poorer prognosis. There was no link made between experience and predicting prognosis

Thank you for these precisions, we modified the sentences as following:

Line 523: “Experienced GPs’ were more likely to report having a gut feeling [7]. The GFQ is a useful tool for eliciting reflection on diagnostic processes between experienced GPs and trainees. There is some evidence that the more experience a GP has, the more accurate his/her gut feelings is related to the diagnosis cancer [20]. The GFQ can be used to study this relationship further.”