In your revision, we would like you to pay particular attention to the following points:  
  
1.  Many of the comments from the reviewer and Associate Editor are significant and will require quite a bit of work for you to address, so do be aware that this request for a revision does not mean that the paper will eventually be accepted, and it is likely that additional revisions will be needed.    
  
2. I think it would be much more useful to readers if the conceptual model focused more on confidence and accuracy.  There is a box for confidence, but I don't see accuracy mentioned at all.  As is, it is more of a high level model of diagnosis, but models of the diagnostic process already exist.  Try to clarify how your scoping review contributes to the model and try to say more about confidence and accuracy.  This could be done by editing both the figure and the legend.   
  
3. Please shorten the  what is known/what this study adds/and how the study might affect research... to a total of about six sentences.

Associate Editor(s)' Comments to Author (if any):  
  
Associate Editor  
Comments to the Author:  
Thank you for submitting your article to BMJ Quality and Safety. We apologize for the time under review as we had some difficulty in obtaining suitable external reviewers. This happens at times and is not a reflection on the quality of your article. Below, please find comments from one external reviewer and one associate editor. You will see that the reviewers felt that your article has some strengths, but has some areas that must be addressed before it can be considered suitable for publication  
  
The study deals with an important topic, especially with the advent of artificial intelligence based approaches to diagnostic decision support. A scoping review is appropriate given the heterogeneity of the literature. The article has a number of strengths - it is organized and well written, and the conceptual model helps bring clarity to the current state and future research questions in this area. That said, I have several concerns as follows:  
1) It is not clear how the authors synthesized the data from the primary studies. What data was extracted and how were the results organized into the themes described in the results section?  
2) The primary and secondary research questions are well defined, but the results and discussion do not clearly link back to these questions. Since there is no quality assessment of the studies, it is hard to know how much weight to give to the specific studies that are emphasized in the results. I suggest reformatting the results section to correspond to the initial research questions.  
3) The included studies are a mix of studies that used simulated clinical scenarios (eg, vignettes) and those that used real patients. A key piece of data that is missing is what gold standard was used to determine diagnostic accuracy. This needs to be reported for each study and will likely be different between the two types of studies, because otherwise it is difficult to assess the accuracy of calibration and confidence.  
4) Were participants in the studies allowed to use any diagnostic decision support tools? This will also be important to acknowledge as it could impact both confidence and calibration.  
5) I think the article is best categorized as a scoping review rather than as a systematic review, and would edit the title to reflect that.  
  
Reviewer(s)' Comments to Author (if any):  
  
Reviewer: 1  
  
Comments to the Author  
This is a brave attempt to tackle an area which consists of many disparate studies of varying quality. For this reason, I think that some assessment of study quality is needed. It is important not to take findings at face value, giving them equal weight, and drawing conclusions from studies irrespective of their quality.  
  
I have some concerns regarding the exhaustiveness and rationale of the search strategy:  
- I note that Web of Science and Embase were not searched. Although there is some overlap with Scopus and Medline respectively, it is only partial.  
- Inclusion/exclusion criteria: Was there a requirement that studies measured the relationship between confidence and accuracy? For example, if a diagnostic study measured both confidence and accuracy - but not their relationship - was it excluded?  
- Although the inclusion criteria mention ‘any medical discipline’ and psychiatry is not specifically excluded, the review seems to have excluded psychodiagnosis (i.e., diagnosis of mental health conditions by psychiatrists and clinical psychologists). This requires some justification, especially since nursing has been included, which is not a medical specialty. It would also be useful to compare the review findings with this thorough meta-analysis of Miller et al. (2015) “A Meta-Analysis of Confidence and Judgment Accuracy in Clinical Decision Making”, which concentrated on mental health professionals and covered a wider range of clinical tasks beyond diagnosis.  
  
Research synthesis: Very little detail is provided about the methodology followed to extract information from the papers. Was this done by one author (which would be a limitation) or more? And if >1, was the level of agreement measured? Were any numbers extracted, e.g., correlations between confidence and accuracy?  
  
It is unclear how calibration was measured in the different studies and how over- and under-confidence was determined. What about studies that used verbal expressions of confidence (i.e., not scales with numbers)? What about studies that used 0-7 or 0-9 etc. scales to measure confidence? Did the authors do any calculations on the confidence and accuracy relationship in each study?  
  
Table 1. What does it mean that participant experience levels are fixed across participants? It would also be useful to note whether participants were purposefully selected for experience (to allow a better distribution of experience in the sample and test its relationship with accuracy, confidence, etc.) or were recruited without considering their experience (and so experience should be used only in exploratory analyses, strictly speaking).  
  
One important distinction that the review should address is the timing of confidence measurement in the diagnostic process. The authors allude to this towards the end but it would be useful to identify when confidence was measured in the studies – e.g., at the start, during, or at the end. Initial confidence can determine the outcome of the diagnostic process and, although it tends to correlate with final confidence, it is a more useful variable to measure.  
  
The authors rightly argue for more realistic designs and high fidelity simulations - I would draw their attention to the work of Kostopoulou and colleagues that used actors as patients in simulated consultations and measured confidence and accuracy (BJGP 2017).  
  
The authors also make a good argument for studying the confidence-accuracy relationship in team environments. This is indeed an understudied area, as it is rather difficult to study cognitive processes where several people interact, and it would indeed be challenging to isolate and study confidence calibration. It is nevertheless important, since communication within teams is highly influenced by confidence of individual members.