Comparative Effectiveness Review Disposition of Comments Report

Research Review Title: *Use of Natriuretic Peptide Measurement in the Management of Heart Failure*

Draft review available for public comment from December 20, 2012 to January 24, 2013.

Research Review Citation: Balion C, Don-Wauchope A, Hill S, Santaguida PL, Booth R, Brown JA, Oremus M, Ali U, Bustamam A, Sohel N, McKelvie R, Raina P. Use of Natriuretic Peptide Measurement in the Management of Heart Failure. Comparative Effectiveness Review No. 126. (Prepared by the McMaster University Evidence-based Practice Center under Contract No. 290-2007-10060-I.) AHRQ Publication No. 13(14)-EHC118-EF. Rockville, MD: Agency for Healthcare Research and Quality; November 2013. www.effectivehealthcare.ahrq.gov/reports/final.cfm.

Comments to Research Review

The Effective Health Care (EHC) Program encourages the public to participate in the development of its research projects. Each comparative effectiveness research review is posted to the EHC Program Web site in draft form for public comment for a 4-week period. Comments can be submitted via the EHC Program Web site, mail or email. At the conclusion of the public comment period, authors use the commentators' submissions and comments to revise the draft comparative effectiveness research review.

Comments on draft reviews and the authors' responses to the comments are posted for public viewing on the EHC Program Web site approximately 3 months after the final research review is published. Comments are not edited for spelling, grammar, or other content errors. Each comment is listed with the name and affiliation of the commentator, if this information is provided. Commentators are not required to provide their names or affiliations in order to submit suggestions or comments.

The tables below include the responses by the authors of the review to each comment that was submitted for this draft review. The responses to comments in this disposition report are those of the authors, who are responsible for its contents, and do not necessarily represent the views of the Agency for Healthcare Research and Quality.

Source: www.effectivehealthcare.ahrq.gov/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productID=1754
Published Online: November 20, 2013

Commentator & Affiliation	Section	Comment	Response
Rev 1	General	Thank you for the opportunity to act as a reviewer for this outstanding document. In General, I found the text to be extremely well-written, clear, and with a consistent standard throughout with respect to language and definitions. The referencing was Generally extremely sound, with no obvious pivotal data missing.	We thank the reviewer for the positive feedback
Rev 1	General	I would ask the authors to consider the current state of both biomarkers relative to 1) consensus statements regarding their use, as well as 2) current clinical practice guidelines that consider their use. I suggest this as it would help provide less confusion about alignment with peer-reviewed statements, while allowing for equipoise relative to novel data.	Most Guidelines suggest that the value of the natriuretic peptides is in their ability to rule-out HF. Values above the decision points encourage further investigations (i.e. echocardiograph) rather than a diagnosis. This is reflected in our findings
Rev 1	disc	Diagnosis: thanks largely to a lack of industry focus on appropriate cut-off points for NT-proBNP, early and subsequent trials lacked a focal point for upper reference limits. The statement on P 24, line 39 regarding a lack of consensus for optimal cut-offs is actually not true—the NT-proBNP consensus panel endorsed an age stratified cut-off approach of 450 pg/mL/900 pg/mL and 1800 pg/mL for ages <50/50-75/>75 years of age, as studied in the International Collaborative of NT-proBNP study and validated in the recent BNP4Ever study. These thresholds are now internationally used, and it would be of great benefit—and particularly increase the impact of this document if some comment about these thresholds were made.	We have added a comment to refer to the ICON study which is relevant to KQ1. Several of the ICON papers are acknowledged in this review – Januzzi et al, 2006; Mohammad et al, 2010; Baggish et al, 2010. A statement regarding cutpoints has been added to the Overview for KQ1 Section of the full report. The BNP4Ever study was published after the search for this review was completed so was not mentioned in this report.
Rev 1	ES	Relative to the lack of cost-effectiveness commented upon on Page 36, line 23, bullet 7: there are prospective data (BASEL study, Mueller et al, NEJM, as well as IMPROVE-CHF, Moe et al, Circulation) as well as modeling data (PRIDE Study, Siebert et al Am Jour Cardiol) showing cost-effectiveness. Those data seem to have been overlooked.	Bullet # 7 has been removed The Discussion in the main document has also been adjusted to remove the same point. The scope of this review did not include cost-effectiveness data. The additional work required to extract this data was considered beyond the scope of the project.

Commentator & Affiliation	Section	Comment	Response
Rev 1	General	Regarding alignment with clinical practice guidelines, the upcoming ACC/AHA clinical practice guidelines for heart failure will be issuing a Class I, LOE A for BNP/NT-proBNP as diagnostic tools for heart failure, noting—as I strongly believe this document should—that these biomarkers are useful for diagnosis when taken into the context of the broad range of clinical variables gathered in patients with suspected heart failure, including history, physical examination, and other tests.	For KQ-1 (see ES page 21) we summarize the evidence as "BNP/NT-proBNP have good performance to rule out, but lesser performance to rule in, the diagnosis of HF compared to the reference standard of overall global assessment on the patient's record" For KQ-2 (See ES page 22) "Both BNP and NT-pro BNP have good diagnostic performance in primary care settings…" The guidelines you are referring to are not yet published. In the Introduction we have added a paragraph which includes points from the 2012 guidelines from the ECA and the Canada. We do not think this document supports the level of evidence you suggest the ACC/AHA guideline will allocate and would recommend that you provide the guideline panel with a copy of this report.
Rev 1	General	In point of fact, this document may wish to make the comment that like no other diagnostic test, the natriuretic peptides have been studied in great detail—in much greater depth than chest radiography for example—and at a certain point, the pluses and minuses on a biological or scientific level must be taken in the greater context, which is that randomized data (such as BASEL or IMPROVE CHF) do show utility.	The body of literature is certainly overwhelming but good quality studies that demonstrate clear patient outcome benefit are not that prominent. Simply adding another test because it has excellent rule out diagnostic accuracy does not make sense for all patients presenting with HF. A more weighted approach may be required.
Rev 1	General	Regarding outpatient diagnosis, the authors seemed to overlook the reference by Hildebrandt et al European Heart Journal, 2010, which strongly supported age-stratified reference limits for NT-proBNP.	We acknowledge this omission and have this review to the discussion.

Commentator & Affiliation	Section	Comment	Response
Rev 1	prognosis	No prognostic tool is perfect, and there are numerous studies of prognostic models such as the Seattle Heart Failure Model that are negative, yet we endorse and use such tools without controversy. I am unsure why we need more data about 'whether' natriuretic peptides are prognostic; they are. What are needed are data that establish how to utilize the prognostic information gained. In point of fact, again, the ACC/AHA clinical practice guidelines will be issuing a Class I, LOE A for BNP/NT-proBNP as prognostic tools.	We do agree with the reviewer that there are a number of prognostic models that were created to assess prognosis of patients with heart failure. This SHFM predicts mode of death given clinical, diagnostic and laboratory data for ambulatory patients (i.e. not in acute decompensation). This model does not include BNP/NTproBNP as a predictor and was therefore not evaluated in our systematic review. Our data would suggest that BNP/NTProBNP is a predictor of mortality and morbidity outcomes. However, the evidence for this is at the lowest level of validation. We do agree with the reviewer that we do need better approaches to utilize the prognostic model data. Our discussion suggests that there are higher level validation models required that include BNP/NTproBNP.
Rev 1	Methods/statistics	It is worth considering a commentary that the gross heterogeneity of the study designs has led to an incorrect notion that the approach does not work. In fact, there are two meta analyses (Felker et al, Am Heart Jour and Porapokkham Arch Int Med) that show a favorable 20-30% reduction in all cause mortality when pooling the very data you cite. A pooled primary-level analysis of all these trials is currently submitted for publication, which verifies this finding.	Additional text has been added in the Discussion and includes reference to the previous 2 systematic reviews on this topic. The decision to perform a meta-analysis is based on the assessment of clinical and methodological diversity. We found the diversity between studies to be too large to perform any meta-analyses. Further, studies with the smallest CI will have more weight in a pooled estimate. Sensitivity analysis performed in the Porapakkham's study found that removal of the TIME study (contributing 49.6% weight) gave an estimate that was non-significant demonstrating the weakness of the meta-analysis. Heterogeneity tests can be helpful but the interpretation of the data should not be solely based upon them.
Rev 1	General	For those studies that chose a low BNP or NT-proBNP target and achieved substantial lowering of the marker, a significant benefit on outcomes was observed. A good summary of the nuances of this important topic—which this document comes close to articulating but leaves the reader wanting slightly—can be found in our recent publication in Clinical Pharmaceuticals and Therapeutics (Motiwala and Januzzi, 2012). Thus, it is fair to say that there is considerable potential here, and multicenter trials that are based on these lessons learned are currently planned.	Yes, that is a great review and adds to the knowledge on BNP guided trials.

Commentator & Affiliation	Section	Comment	Response
Rev 1	General	In recognition of a good understanding of the data (several positive trials plus 2 meta analyses) the upcoming ACC/AHA guidelines are contemplating an elevation of biomarker guided care to a Class IIa recommendation with the usual caveats regarding the importance of taking the approach within the context of standard care. It is worth considering this fact in crafting the tone of your work.	Unpublished information is not included.
Rev 1	General	Please see my comments relative to equipoise with clinical practice guidelines. Upcoming guidelines will give a Class I, LOE A for diagnosis and prognosis, and a Class IIa, LOE B for management.	Unpublished information is not included.
Rev 2	General	The problem with this type of methodology is the exclusion of papers which contribute to the answers under investigation as they do not fit the narrow methodological criteria used.	We designed inclusion and exclusion criteria so that studies included in the review had population characteristics very similar to the populations that clinicians would face in practice in the ED, Urgent care or primary care settings.
Rev 2	Methods	Methods: Whilst exclusion of some data sets is reasonable on methodological grounds they can then be used for consistency checking of the findings. Some answer the questions posed.	We did not exclude data sets but studies. Good systematic review methodology implies that we do not assess information from excluded studies.
Rev 2	Discussion	Discussion/ Conclusion: There appear some significant gaps for reasons stated above	We have addressed these gaps to the best of our knowledge.
Rev 2	General	Clarity and Usability: Adequate but the there appears to be no discussion of the previous HTA by Mant et al	We have acknowledge this omission and have added the following to the discussion: “The results obtained from this review are in agreement with a recent systematic review using individual patient data meta-analysis where both BNP and NT-proBNP had high sensitivities (93%) for diagnosis of HF, when optimized for sensitivity [Mant et al, 2009]”
Rev 3	General	General Comments: The report summarises a substantial amount of work and gives a clear overview of the literature in the area. The authors should be congratulated on their ability to sort through the vast literature in the area and coherently summarise the results of this research.	We thank the reviewer for the positive feedback.
Rev 3	Intro	Introduction: The structured abstract and executive summary is clear and accurate summaries of the research contained within the report.	We thank the reviewer for the positive feedback.

Commentator & Affiliation	Section	Comment	Response
Rev 3	Methods	Methods: The inclusion/exclusion criteria are justifiable and have been consistently applied. The search strategies appear to be highly sensitive. The risk of bias tools have been used appropriately and adapted where necessary for the review. The statistical Methods used in the review were according to the most current standards, minimizing the risk of bias. The Methods are well described in the text of the report.	We thank the reviewer for the positive feedback.
Rev 3	Results	Results: Overall the results are well presented and appear to be an exhaustive description of the literature in the area. The analysis of the prognostic studies was particularly well done given the challenge of inconsistent reporting and lack of reporting standards in this area of research.	We thank the reviewer for the positive feedback.
Rev 3	Discussion	Discussion/ Conclusion: The findings have been clearly stated and an accurate summary of the results of the review. The research section is clear and provides direction on future research in the area.	We thank the reviewer for the positive feedback.
Rev 3	Discussion	The Discussion commencing on page 433 is an excellent summary of the literature covered by each of the key questions.	We thank the reviewer for the positive feedback.
Rev 3	disc	The decision to emphasize BMI and renal function as potential confounders of the diagnostic and prognostic accuracy of the natriuretic peptides is not well justified within the review. These factors appear to have been raised by an external committee but there is not an explanation why these factors, amongst many that could have been suggested as potential confounders, have been singled out.	The AHRQ report in 2006 identified both renal function and weight as potential physiological variables that modified BNP or NT-proBNP. These associations were confirmed as relevant during Discussions with experts during the design phase of the current review. Thus we purposefully included them as extractable data items for the review. This enables us to comment on the relevance of these items throughout the report. For both diagnosis and prognosis many studies did include these measurements and a fair number commented on them allowing us to make some comment on them. A recent review paper (Heart Fail Rev (2012) 17:81–96) discussed the topic in reasonable detail. To address your comments the text has been modified in the Discussion to state that these were identified from the 2006 report. This will hopefully clarify why they were emphasized.
Rev 3	General	Note that all references to pages are from the pdf version of the report rather than the page numbers printed on the page. The page numbers should be fixed for the final report, particularly as most people will access the report as a pdf.	This has been corrected, numbered according to guidelines.

Commentator & Affiliation	Section	Comment	Response
Rev 3	Methods	Page 21: the summary of the Methods should include the HSROC method that was used for combining the results of diagnostic test accuracy studies that is described later in the report.	Analysis was redone using GLM mixed model approach to bivariate meta-analysis of sensitivity and specificity suggested by Chu and Cole (1). This approach corresponds to the empirical Bayes approach to fitting HSROC model (2). 1. Chu H, Cole SR (2006). Bivariate meta-analysis of sensitivity and specificity with sparse data: a Generalized linear mixed model approach. Journal of Clinical Epidemiology 59:1331-1332. 2. Macaskill P (2004). Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. Journal of Clinical Epidemiology 57:925-932.
Rev 3	results	Page 23 line 26: did the authors mean that the negative likelihood ratios were all greater than 0.20? It would be more consistent with the previous sentence if this were all less than 0.20.	Yes – this has been corrected
Rev 3	results	Page 33 – the reason why a meta-analysis was not conducted of BNP guided therapy should be included in the executive summary.	The ES has been revised and includes the reason why no meta-analysis was conducted.
Rev 3	Methods	Page 157 and similar tables throughout the report. It is not clear whether yes indicates a low or a high risk of bias. I suggest that the legend should be changed to High, Low or Unclear risk of bias.	The legend has been added to all ROB tables throughout the report. The specification of the questions was detailed in the Methods section (chapter 2). However, we have modified the title of each risk of bias figure to indicate this more clearly. We agree with the reviewer and have modified the figure titles to indicate that yes means low risk of bias, no means high risk of bias and unclear indicates that we were unable to ascertain risk of bias for that criterion.
Rev 3	results	Page 163 and similar tables throughout the report. It is not clear what are the meanings of the A, C and D legend that is used in this table.	Corrected in report
Rev 3	results	Page 166 and several other tables in this section. It is not immediately clear that the HR refers to the cut-point referred to under BNP levels. Can this be made clearer?	When a HR is reported with additional information such as, “per unit of change” or “per unit of standard deviation”, it indicates that the BNP/NTproBNP variable was used in the model as a continuous variable. As such, the threshold listed does not apply. If the BNP/NTproBNP was used as a dichotomized variable, then the threshold applies.

Commentator & Affiliation	Section	Comment	Response
Rev 3	General	Page 237 – line 33 “most of the study designs” – “the” is missing	Corrected in report
Rev 3	General	Page 238 – the notes that appear at the bottom of figure KQ3-3 did not appear in previous tables and are helpful.	Corrected in report
Rev 3	results	Page 249 – several of the studies are described as a RCTs. However, these are not RCTs of this clinical question but are Generally sub-studies of RCTs conducted for another purpose. This needs to be made clearer in the description of the studies throughout the report.	We have reviewed the study design classification and clarified this issue in the “study characteristics section and the summary tables. We have changed the text as noted: “Two articles were randomized controlled trials (RCTs) of BNP-guided therapies versus non-BNP-guided therapies. Four articles were secondary analyses of data initially collected in RCTs; however, the secondary analyses did not account for the groups to which participants were randomized.” RCTs have been assessed throughout the report.
Rev 3	Can’t find	Page 274 line 22– there is a HR reported with no unit .	This has been addressed.
Rev 3	Can’t find	Page 297 – the units used for calculation of the HR require clarification in rows 2, 3 and 4.	This has been addressed.
Rev 3	results	Page 381 2nd para. To know if MR-proADM is a stronger predictor than NT-proBNP, would need to add NT-proBNP to the model and then show that MR-proADM adds to the prognostic value. The results as stated do not show that either of the biomarkers is a stronger predictor than the other.	We have removed the sentence and modified the text preceding this to clarify the findings from this study.
Rev 3	General	Page 433 line 36 – the word of is missing	Corrected in report
Rev 3	results	Page 434 – the key question should be repeated.	This has been addressed in the report.
Rev 3	Can’t find	Page 439 line 29 and 30 – the last sentence of this paragraph is not clear.	This has been addressed in the report.
Rev 3	Can’t find	Page 440 line 6 – prediction is misspelt	Corrected in report
Rev 3	disc	Page 440 1st para – the list of the common factors in the prediction of cardiovascular disease outcome are the factors from the Framingham risk equation, and are the factors that increase the risk of developing cardiovascular disease. This seems quite different to the subject of this section, which is the prognosis of patients with stable heart failure.	We have added some further clarification as follows: “as these have been shown to be associated with mortality from cardiovascular disease and should thus be accounted for in all-cause mortality and cardiovascular specific mortality assessment.”

Commentator & Affiliation	Section	Comment	Response
General	disc	Page 444 – it is not clear how prognostic studies would be employed in randomised controlled trials. What would be the intervention and what outcomes would be measured?	A recent overview of General approaches to prognostic studies states the following: "Data from randomized trials of treatment can also be used to study prognosis. When the treatment is ineffective (relative risk=1.0), the intervention and comparison group can simply be combined to study baseline prognosis. If the treatment is effective the groups can be combined, but the treatment variable should then be included as a separate predictor in the multivariable model. Here treatments are studied on their independent predictive effect and not on their therapeutic or preventive effects. However, prognostic models obtained from randomized trial data may have restricted Generalizability because of strict eligibility criteria for the trial, low recruitment levels, or large numbers refusing consent." Moons 2009
Rev 3	Formatting	Page 451 line 20 – the word either is inappropriate as there are 3 markers	This has been corrected in the text to read "the other markers"
Rev 3	disc	Page 453 2nd para – it is not established from the evidence presented in this report that renal function and BMI are confounders for NT-proBNP. There is a consistent relationship, but this establishes correlation not confounding.	The choice of these specific covariates identified as confounders is based on consultation with the expert panel and findings from our previous review on this topic. These factors were singled out for assessing risk of bias for confounding. In our judgment we had to establish the most important confounders but could not be all inclusive all possible confounders.
Rev 3	General	The established risk factors are for the prediction of the development of cardiovascular disease. It needs to be clear whether any prognostic model is for this purpose or for a more General prognosis.	The studies in this systematic review looked at prognostic factors rather than risk factors. The prognostic models in this review attempted to evaluate the predictive strength of BNP/NTproBNP with respect to mortality outcomes (all cause and cardiovascular) and morbidity outcomes (hospitalizations, etc) and composite outcomes (mortality
Rev 3	Formatting	Page 455 – line 37 and line 40. There is an incomplete sentence in these lines.	This has been addressed in the report.
Rev 3	Can't find	Page 456 – line22. A number is missing in this sentence.	This has been addressed in the report.

Commentator & Affiliation	Section	Comment	Response
Rev 4 AE	General	The best review I have ever read! If this is going to be published, I would suggest submission to Clinical Chemistry (disclosure – I am a Associate Editor). 2. Not sure if these papers would be worth including: a. St. Peter JV, Hartley GG, Murakami MM, Apple FS. (BNP) and N-terminal pro-BNP in obese patients without heart failure: relationship to body mass Index and gastric bypass surgery. Clin Chem 2006; 52: 680-685 ; Published February 23, 2006. doi: 10.1373/clinchem.2005062562. b. Apple FS, Murakami MM, Pearce LA, Herzog CA. Prognostic value of high sensitivity C-reactive protein, N-terminal proBNP, and cardiac troponin T and I in end stage renal disease for subsequent death over two years. Clin Chem 2004; 50: 2279-85. c. Peacock WF, De Marco T, Fonarow GC, Diercks D, Wynne J, Apple FS, Wu AHB, for the ADHERE scientific advisory committee study group. Cardiac troponin and heart failure outcome in acute heart failure. New Eng J Med 2008; 358: 2117-26.	1- Thank you. Submission to a journal or journals will happen and we will keep your suggestion in mind. 2- a and b- Not a heart failure population and thus excluded. c- This report did not include NP data and was thus not found in the search.
Rev 4 AE	General	What about studies that address the optimal ordering times of BNP or NT-proBNP during the course of an admission; such as a) an admit value, b) a dry weight value, c) a predischarge value to assist in patient management and risk outcomes assessment post discharge as well as likelihood of readmission both short and long term post discharge.	We examined the first analysis, taken at presentation or as soon after as possible.
Rev 5	General	The manuscript assumes that there is no difference between assay performance or cutoff values that are necessary for any of the natriuretic peptide assays. That is not the case. For NT-proBNP since there is only one manufacturer, Roche, which leases to other companies, the values are mandated to be similar. However, for BNP there is significant diversity. Many of the companies decided to harmonize their assays at a value of 100 ng/ml after the initial BNP Breathing not Properly trial but there are substantial discrepancies both above and below that value. It is not clear that was recognized or taken into the account in the analysis.	As the reviewer states, results from NT-pro-BNP assays are, by design, similar which simplifies the comparison of performance between assays and across studies. BNP assays, on the other hand, are not harmonized. We attempted to address this issue by comparing the various BNP assays at the lowest, optimal and 100 ng/mL cutpoints. Comparison at the various cutpoints we feel would give the reader an understanding of the performance of BNP, since the 100 ng/mL would allow readers to directly compare at a specific value as well as assess the overall performance using the optimal and lowest cutpoints.

Commentator & Affiliation	Section	Comment	Response
Rev 5	Methods	When one looks at overall values one could, depending upon the relative proportions of various subgroups in a population, be misled. Thus, the failure to find in one study that there are differences related to age and gender in the appropriate cutoff values does not eliminate that as a possibility. Indeed, it is clear, since the number chosen both in your review and in others for BNP was of 100ng/ml which is substantially below that which might be found in elderly individuals, men and women that this is highly likely. This ought to be made clear in the interest of not inadvertently leading to the mistriage of patients who are elderly.	The aim of our study was not to recommend the optimal cutpoints that should be used in clinical practice. Because studies use various cutpoints and, as you have stated, diverse populations, we chose the most commonly used cutpoints to assess the diagnostic performance. We did attempt to address the determinants affecting BNP/NT-proBNP performance, including age.
Rev 5	General/Methods	No attempt is made to segregate studies into those that determined cutoff values post hoc versus those that tested pre hoc values. This is obviously critical and makes it very difficult to harmonize studies because most studies optimize their own local results. This suggest that their extrapolation to other circumstances may be less than ideal. This is a major limitation that ought to be clearly articulated. It is from the theoretic point of view related to the measurement biologic variation that it is only applicable to normals because in an abnormal population, there is pathophysiology that alters the values in a nonbiological or at least a pathobiological manner in addition to the normal biologic variation. This distinction is not made very clearly if at all and is key to the science. There would be substantial objection to the term biological variation in heart failure patients or even suggesting those data could be reliable. It should be obvious that the veracity of those data would be highly dependent on the population studied and the intrinsic stability of their heart failure.	When examining BNP, we analyzed the data, using minimum, manufacturers' suggested and optimum cut points. While the optimum (and perhaps the minimum) cut point can be argued to be post-hoc, the manufacturer's suggested cut point for each assay is pre-hoc.
Rev 5	General	Methodological issues not considered in results or disc and need to be.	We have addressed several methodological issues in our results and Discussion sections. We would be happy to speak to any additional methodological issues the reviewer would like us to address; we just need more specific comments here.
Rev 5	General	Not adequately nuanced to take into account some of the ambiguities that were ignored as indicated.	We are unclear as to the context of this comment.

Commentator & Affiliation	Section	Comment	Response
Rev 5	General	To assess the diagnostic accuracy of B-type natriuretic peptide (BNP) and N-terminal proBNP (NT-proBNP) for detecting heart failure (HF).For this, it is key that heart failure itself is defined. Far too often this is defined by echocardiography in studies. Imaging may be inferior to BNP in detecting clinical heart failure. Left atrial size of evidence of venous congestion are superior to LVEF in detecting dysfunction.	We agree that heart failure is not well defined. We used the definition of HF as described by each author for his/her study.
Rev 6	General	To determine whether BNP and NT-proBNP are independent predictors of mortality and morbidity in HF and whether they add to the predictive value of other markers; main concern here is the minimal data-set applied. If key data such as heart rhythm, body mass and renal function are not included then the model is not clinically valid. Atrial fibrillation has not been dealt with adequately.	Atrial fibrillation has been evaluated where it was reported. It is found in the diagnostic, prognostic and therapeutic sections. However it was not widely reported and the influence on BNP results directly seems less consistent than the other factors identified. AF is seen as predictor of poor prognosis and thus it should be included in models and multivariable analysis. We have listed all the occasions where AF was included in the studies that we reviewed in the tables. AF did not stand out as an item that should be individual set apart even though there were a number of studies that found univariate association and a smaller number that found multivariate association. We feel that we have reported this fairly from our findings in this review.
Rev 6	General	To ascertain whether treatment guided by BNP or NT-proBNP improves outcomes in HF; Should state compared to what? Slovenly care or going for the guidelines. Also, which components of care change - mostly diuretics I think.	Revised the objective by adding compared to usual care (as per key question wording).
Rev 6	General	This is a very thorough/exhaustive review. The authors are to be congratulated. However, the Introduction fails to tackle adequately the complexity of diagnosis of heart failure and in particular the superiority of measure of atrial compared to ventricular structure and function in predicting outcome.	We agree that the diagnosis of HF is challenging. The aim of this review was to examine the diagnostic performance of the natriuretic peptides against the reference criteria provided by each author, not to evaluate the performance of other diagnostic protocols. The Introduction has been modified to include a paragraph that reinforces the challenge of the diagnosis HF. The review did not consider atrial and ventricular structural changes. It would probably incorrect to include commentary on this aspect in the Introduction or elsewhere unless we felt that the reviewed literature confirmed this and we would then include it in the results and Discussion.

Commentator & Affiliation	Section	Comment	Response
Rev 6	General	The Methods appear robust including the description of methodology for diagnosing heart failure. However, how good are cardiologists?	It is true that the accuracy of diagnosis is dependent on the skill of the cardiologist, however this is true both in the studies as well as in routine clinical care. Therefore, we believe our results should be applicable to real-work clinical diagnosis. Unfortunately the literature leans heavily on clinical judgment with all its failings.
Rev 6	General	I would like the authors have to ranked the quality of papers depending on the completeness of relevant data for studies of BNP (age, sex, symptoms, LVEF, left atrial size, peripehral oedema, heart rhythm, BMI, renal function).	Ranking papers across multiple quality variables is always challenging. To rank them according to one criteria – completeness of relevant data – would not be correct in context of an overall evaluation of quality. Quality assessment (Method section) provides an assessment of bias. Simple ranking of papers by completeness of data included does not adequately address bias and therefore study quality.
Rev 6	results	Key messages are somewhat lost. Should be more emphasis on the key importance of heart rhythm along with adiposity and renal function (page 78/79) - in addition to any difference in LV function.	The text has been modified.
Rev 6	General	Greater care needs to be applied to assessing reports on multi-marker data. So often data are simply split by medians into 4 groups (with two markers). This is very crude and misleading. More reservations should be placed on the robustness of multi-marker data.	We believe we have addressed this point within the future recommendations section, suggesting the need for internal and external validation.
Rev 6	General	I think the incremental power of NT-proBNP is not emphasised enough and yet, perhaps because the relationship between prognosis and NT-proBNP is smoothly incremental - statistical power may be stronger than practical value. It is quite worrying when CART analyses are applied how clinically poor even NT-proBNP really is. There should be more requests for NRI, IDI and CART analyses.	We agree with the reviewer that there should be more requests for NRI, IDI computations. The future recommendations section has been modified to emphasize the need for appropriate computations.
Rev 6	General	Failure to recognise the importance of atrial fibrillation (or I missed it in the 404 main pages)	We agree with the reviewer that atrial fibrillation may be an important subgroup to identify. However, if the study authors did not identify this or stratify the analyses for this group then we are unable to evaluate this.
Rev 6	General	I think it would be useful to summarise the important points in free, reference-light, text.	We have attempted to add some practical statements in the Discussion.

Commentator & Affiliation	Section	Comment	Response
Rev 6	General	Overall, excellent but requires revision in places, particularly on the issues of atrial fibrillation, lack of a diagnostic 'gold-standard' for heart failure (maybe NT-proBNP is a vital component of the gold-standard against which echo or clinical diagnosis should be judged) and lack of evidence that it is a practically useful prognostic tool (OK I admit low levels indicate a great outcome - but high levels are not so good at discriminating in analyses I have seen).	The standard for HF diagnosis has been revised in the Introduction and discussed under comment 90. The practical utility of NP in prognosis is low and hopefully the revised text demonstrates this better. We have also enhanced the Discussion on 'gold-standard' for heart failure.
Rev 7	Intro	General Comments: All of the key questions are appropriate and explicitly	We thank the reviewer for the positive feedback.
Rev 7	Intro	In question #6, this question could have been further divided whether BNP/NT-proBNP guided management is useful in the ER setting, or in hospitalized patients or in the community.	Our proposal was to evaluated guided therapy in patients diagnosed with HF in a chronic care setting.
Rev 7	Intro	Introduction: This section is well written and contains all necessary information.	We thank the reviewer for the positive feedback.
Rev 7	Intro	One missing area is cost effectiveness which were addressed in several trials but not explicitly covered.	Cost-effectiveness was beyond the scope of this already large review.
Rev 7	Intro/disc	Consider including the heart failure guidelines in 2012 from the European Society of Cardiology and the Canadian Cardiovascular Society.	A paragraph has been included in the Introduction and further reference has been made to these 2 documents.
Rev 7	General	Search strategy and study selections were comprehensive. There should have been "spot checks" for data extraction with investigators making spot inquiries as he/she reviews the report.	Our method sections describes the following: "During the course of writing the report, investigators reviewed the extracted information for accuracy and made corrections as necessary." We believe this would be consistent with spot checks.
Rev 7	Intro	The definitions for the outcome measures are appropriate.	We thank the reviewer for the positive feedback.
Rev 7	General	It is unclear why meta-analyses were only limited to KQ1 and 2.	The choice for undertaking a meta-analysis must consider sources of heterogeneity. In the context of prognosis and therapy, clinical heterogeneity (which considers the population, method of BNP use, the outcomes and the comparison group) were considered. If it did not seem appropriate to pool the individual study findings for these reasons, meta-analysis was not undertaken. This has now been more clearly stated in several areas of the report.
Rev 7	Methods	In your literature search, you have 35 reports from "grey" area such as conference reports. How did you weigh this and how did you choose the 35 reports among others?	The text in chapter 3 indicates the following: "three gray literature sources: regulatory agency websites, clinical trial databases, and conference sources." These sources were included and screened as per other citations obtained from bibliographic databases.

Commentator & Affiliation	Section	Comment	Response
Rev 7	Intro	Results: The amount of information is appropriate and the key messages are explicit.	We thank the reviewer for the positive feedback.
Rev 7	Methods	Race was limited to mostly African Americans and White comparisons. Notably absent were East Asians and South Asians.	For KQ3, KQ4 and KQ5 we did extract racial profile in studies where this was noted. We noted that this information was not consistently reported or not at all in the majority of studies. We have noted this in the future research recommendations. Race was checked and information added to the study description section of KQ1, KQ2, KQ6 and KQ7.
Rev 7	ES	On KQ3 on page ES-11, the sample size are confusing. Do they refer to combined studies (seems to few) or refer to number of studies? Similarly, for morbidity outcomes on ES-12, did the sample size refer to the number of studies?	This has been addressed in the report.
Rev 7	ES	On ES-13, citations on added value of BNP are missing.	The references have been added.
Rev 7	Intro	The flow diagram on pg 74/1018 is particularly helpful.	We thank the reviewer for the positive feedback.
Rev 7	ES	In Table B on page ES-16, the study PROTECT is missing.	The Protect study was not included as it was unclear whether the reported of death as cardiovascular death included all deaths.
Rev 7	results	On page 76/1018, this section is supposed to compare BNP and NT-proBNP, yet table KQ-1 included only BNP studies. Should probably refer to KQ-1-8.	This has been addressed in the report.
Rev 7	results	Regarding research gaps, several studies have addressed cost effectiveness. Furthermore, at this stage of development, it is unlikely that more research as suggested under research gaps will be conducted.	This bullet has been removed. The summary of research gaps is now consistent between the ES and the main document.
Rev 7	disc	In this section, some sections contain citations and some do not. Should probably be consistent.	We have attempted to address this where possible.
Rev 7	ES	The executive summary is well organized but could have been abbreviated. Main points were discussed. Nesiridie (hBNP) was not discussed but it was probably not the objective of the review.	Nesiritide was not part of this review. The executive summary has been reviewed and abbreviated.
Rev 8	General	This report addresses seven key questions regarding the use of (NT-pro-) BNP to diagnose heart failure and to predict outcomes on people already having heart failure. Although the report is very elaborate and includes many outcomes and information, it does not follow the most up to date Methods and misses context. As I am most familiar with Methods for diagnostic accuracy reviews, this is what I comment on most.	We thank the reviewer for these comments. We have addressed the discrepancies in Methods for the diagnostic section. We do not believe your comment applies to the other sections. One of the peer reviewers for the prognosis section, an epidemiologist who is expert in addressing questions of prognosis indicates that are Methods are sound in this area. We have added sections in the Introduction and Discussion to address context.

Commentator & Affiliation	Section	Comment	Response
Rev 8	results	Although the audience and the target population are explicitly defined, I miss information about the clinical context in which the test will be used (see also my comments about the Introduction). Furthermore, the results could have been explained better to improve readability.	The clinical context has been described as recommended in current clinical practice guidelines. In addition to this we have tried to organize the results in a way that makes clinical sense. Each section has been revised to account for the review feedback and hopefully there is improvement in the readability.
Rev 8	General	Please use consistent letter typing and fonts for the headings. For example, the primary heading on page 345 has a smaller font than the subheadings below. When scrolling through the text and looking for this particular section, it would have been easier if it was the other way around.	AHRQ publishing guidelines were followed.
Rev 8	Introduction	The first concern is the lack of context. What is the place or should be the place of (NT-pro-) BNP measurement in practice? When the authors discuss the results, they state that these tests are good to rule out disease, but not good to rule in disease. But is ruling out the purpose the test is used for? If that is the case, then that should be stated up front. What will happen to test positive patients and test negative patients? Will they be treated or referred for further diagnosis, resp. sent home or referred for further diagnosis? Will people only be measured once, at diagnosis? Or will they be measured again once in the X months/years when they are labelled as having HF? These questions will not affect the outcomes, but may help in interpreting the results.	The place of BNP in the diagnosis of HF is to both rule in and rule out. Test negative patients will likely be given a Dx of “not heart failure” whereas those with a positive test will likely be further evaluated We have described the clinical context in light of the clinical practice guidelines. Unfortunately the clinical practice guidelines do not always follow the best available evidence in their formation. The algorithms for use of NP in practice are described in both the Canadian and European guideline. We have not reproduced the diagrams in the report. Serial testing is eluded to in the CPGs but is not given an evaluation. We have tried to revise the Discussion to make the clinical interpretation more straight forward.
Rev 8	Introduction	The authors describe an analytical framework on page 5, but this framework is difficult to understand without any explanation. For example, there goes an arrow from General population via KQ5 directly to mortality etc. But even if there is a direct relationship between the levels of BNP, that does not mean that measuring BNP will indeed affect mortality positively or negatively. Or is that not the point of the analytical framework?	Text to describe the analytic framework has been added.

Commentator & Affiliation	Section	Comment	Response
Rev 8	General/Introduction	I do not understand the difference between KQ3 and KQ4. I am not an expert in prognostic modelling, but would an 'independent predictor' not always add information? Or the other way around?	The difference between KQ3 and KQ4 refers to computational or study design methods that are used to determine and test the value of the BNP/NTproBNP as a factor. Independence in multivariable modeling would suggest that there would be added information from the use of the BNP/NTproBNP result. However many of the multi-variable models do not include the typical factors used in prognostic modeling and thus when tested in a more complete model or with alternative statistical approaches there is not always statistically significant incremental change in the prognostic model. The key is that in KQ4 is the that BNP/NTproBNP is evaluated for its "additional" or incremental value.
Rev 8	Introduction/General	I would like to see some more explanation to key question 6. What does BNP-guided therapy mean and what do these RCTs look like? Are patients randomized to be tested before treatment and when BNP is positive, then the patients are treated and otherwise not? OR is this more a monitoring question? It is completely unclear to me what the clinical situation is that is assessed here.	Additional background on HF therapy has been added to the Introduction as well as a description of an RCT for BNP-guided therapy.
Rev 8	General/ Methods	The authors choose to do only meta-analysis for the accuracy questions. Why for the accuracy questions alone?	The choice for undertaking a meta-analysis must consider sources of heterogeneity. In the context of prognosis and therapy, clinical heterogeneity (which considers the population, method of BNP use, the outcomes and the comparison group) were considered. If it did not seem appropriate to pool the individual study findings for these reasons, meta-analysis was not undertaken. In our judgment the clinical heterogeneity precluded summary estimate computations.

Commentator & Affiliation	Section	Comment	Response
Rev 8	Methods	I think separately meta-analyzing sensitivity and specificity and positive and negative likelihood ratios is not correct. For example, the Cochrane Collaboration (Generally seen as leading in the field of meta-analysis), currently recommends the HSROC and bivariate meta-analytic models. See chapter 10 of the Cochrane DTA reviews handbook, on http://srdta.cochrane.org. The point estimates may not change by using these more advanced Methods, but the confidence intervals may. Furthermore, these Methods are based on the assumption that sensitivity and specificity are correlated with each other and should for that reason not be pooled separately, which makes conceptually sense.	Analysis was redone using GLM mixed model approach to bivariate meta-analysis of sensitivity and specificity suggested by Chu and Cole (1). This approach corresponds to the empirical Bayes approach to fitting HSROC model (2). 1. Chu H, Cole SR (2006). Bivariate meta-analysis of sensitivity and specificity with sparse data: a Generalized linear mixed model approach. Journal of Clinical Epidemiology 59:1331-1332. 2. Macaskill P (2004). Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. Journal of Clinical Epidemiology 57:925-932.
Rev 8	General/Methods	Pooling likelihood ratios is Generally discouraged, as they can result in 'strange' outcomes. See Zwinderman and Bossuyt, Stat Med, 2008.	Analysis was redone using GLM mixed model approach to bivariate meta-analysis of sensitivity and specificity suggested by Chu and Cole (1). This approach corresponds to the empirical Bayes approach to fitting HSROC model (2). 1. Chu H, Cole SR (2006). Bivariate meta-analysis of sensitivity and specificity with sparse data: a Generalized linear mixed model approach. Journal of Clinical Epidemiology 59:1331-1332. 2. Macaskill P (2004). Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. Journal of Clinical Epidemiology 57:925-932.
Rev 8	General/Methods	The authors state that they developed their own scale for cross-sectional studies. Does this include the diagnostic cross-sectional studies?	No - we used accepted tools This sentence has been removed. The QUADAS 2 and the Hayden Index are not specific to study design and can accommodate cross-sectional instruments.

Commentator & Affiliation	Section	Comment	Response
Rev 8	Methods	For the accuracy studies, the authors modified the QUADAS-2 tool. I do not agree with all modifications. In the first place, the authors state that pre-specification of threshold was not relevant (misspelled on page E-2 as "threshed"), because there are many factors involved with the choice of the threshold and with the relation between accuracy and threshold. I do think that prespecification of threshold is important, because by allowing yourself to select the most optimal cutpoint is asking for overestimation of accuracy (see Leeflang ClinChem 2008 and Ewald, JClinEpi 2006). I suspect that quite a few of the included studies did a post-hoc selection of cut-off value, and even this review reports the results for optimal cut-off value separately from the other results. I think the review would improve from an explanation of how cut-off values are used or selected in practice and make a choice according to that. For example, if in practice, people use the cut-off recommended by the manufacturer, then the results should be limited to (or at least focus on) these cut-offs.	We have included the QUADAS question on threshold in the assessment of Risk of Bias.
Rev 8	Methods	The authors state that the overall risk of bias for key question 1 is okay (that all of the domains show low risk of bias), but when I check the tables with the individual studies assessments, there are quite some crosses and question marks, especially for the patient inclusion domain. Therefore, I think the statement that all domains show low risk of bias is overly positive.	We have reworded the document on a per item basis based on the added Question from QUADAS II.
Rev 8	Methods	The applicability question for the patient-domain was interpreted (see page E-4) in terms of exclusion of certain patient groups. But isn't this one of the signalling questions for risk of bias in that domain (Did the study avoid inappropriate exclusions?)	The signaling question asks about "inappropriate" exclusions. In a study designed to have wide applicability for all patients presenting to the ED and primary care, it is appropriate to exclude studies that consider only patients with a specific diagnosis

Commentator & Affiliation	Section	Comment	Response
Rev 8	Methods	The authors removed the signalling question about time interval from theQUADAS-2 tool, with the comment that this was done because they only included studies that used sampling on the same moment anyway. This was however not stated under the inclusion criteria (page 11), moreover, there the authors say: "KQ1 to KQ7: No restriction on inclusion of articles based on length of followup". Which may indicate that studies have been included in which the final diagnosis was made much sooner or earlier than the BNP was done.	The length of follow up was unrestricted – there was not limit on how long after the index test the file was viewed for adjudication, but there was a restriction on the time between the index and reference test. Results –page 71 states that the time between the reference and the index test was considered in the QUADAS The QUADAS help table for reviewers states that the interval between the index and reference test should be less than 2-3 days
Rev 8	General	The authors state that they do not restrict on including a particular reference standard. As the reference standard is really key in determining the accuracy of a test, and differences in reference standards may cause differences in accuracy, I think the authors should address this as one of the factors that may influence accuracy.	The heterogeneity of the reference standards used is addressed in the discussion and we agree with the reviewer that this may influence accuracy.
Rev 8	General/ Methods	The operationalization of GRADE should be better explained (for diagnostic – prognostic and therapeutic questions separately) and perhaps re-considered. 9a. For example, directness for diagnostic questions is considered to be no problem, because most clinicians understand sensitivity and specificity. In the first place, I think many clinicians and methodologists will disagree with this statement: we see over and over again that clinicians do have problem in interpreting diagnostic accuracy measures. Moreover, this is not at all what this GRADE-item refers to. Directness has to do with translation from the retrieved evidence to practice. For diagnostic questions this means either 'can the sensitivity and specificity estimates found in these studies be directly used in practice?' or 'do these sensitivity and specificity estimates tell me that the patient will benefit from this test?' If a translation from accuracy measures to patient important outcomes is desired, then there is almost always indirectness.	A component of directness is how well the populations in the studies correspond to the population that is likely to be seen by clinicians using this data – Can the sensitivity and specificity derived from this study be used in my practice? We designed inclusion and exclusion criteria so that studies included in the review had population characteristics very similar to the populations that clinicians would face in practice in the ED, Urgent care or primary care settings. See reworded section in page 72 (results)

Commentator & Affiliation	Section	Comment	Response
Rev 8	Methods	How is consistency interpreted for diagnostic studies? The authors state that 'the direction of estimates is consistent', but what do they mean with that? What is the direction of sensitivity estimates, for example? In intervention research it is easier to say that all studies show benefit (if OR is above or below 1, depending on the outcome), for example. But sensitivity and specificity estimates lay between 0% and 100%.	We have removed the reference to direction with regard to sensitivity and specificity.
Rev 8	General	Why not use the method from Deeks 2005 to assess publication bias in diagnostic studies?	Testing for publication bias was redone using Deeks' method (1). Deeks JJ. Macaskill P and Irwig Les. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. Journal of Clinical Epidemiology, Volume 58, Issue 9, September 2005, Pages 882-893.
Rev 8	Methods	Factors affecting sensitivity and specificity: is it possible to include these in a meta-regression model, to assess the effect of these factors on accuracy estimates?	Not possible at this time
Rev 8	results	Please check the tables. Sometimes sensitivity and specificity are reported, but not LRs, while it is easy to calculate LRs from sensitivity and specificity. Also sometimes the study population consists of patients with HF, but only specificity estimates are reported, while I would expect that sensitivity is calculated in the diseased (i.e. patients with HF) and specificity in the non-diseased.	We reported only those values that were reported by the author and did not calculate LR+ or LR – if not reported. We presented the data present in the publications and did not attempt to estimate which was not present in the publications
Rev 8	results	Figures H27 and H28 represent the log(DOR). Perhaps the DOR is easier to understand and interpret for the average reader. I also think that these figures can be removed, as they are not discussed in the text and they don't add information.	These figures have been moved to the appendices.
Rev 8	results	ES-6: the reporting is not very helpful here. The authors report as a finding that the sensitivity increases and specificity decreases when a lower cutpoint is used. This is logical and is the basis of the ROC curve, but I can imagine that some readers not immediately know that. But I am not sure if it really is a finding. What I would have found more interesting here, is a report of the range of cutpoints found in the studies. Or whether most studies used the same cutpoint or all different cutpoints. In the main report, these results are reported in a more helpful way.	This comment is confusing. The same text is repeated in both the ES and in the results section

Commentator & Affiliation	Section	Comment	Response
Rev 8	ES	ES-9: very detailed section about the effect of age on accuracy estimate, only based on two studies. Also, nowhere in the report is a real focus on AUCs and here the disc is focused around AUCs. Also, this section is longer and more detailed than the same section in the main report (page 92). This is a bit strange. Please revise and shorten.	This has been revised and shortened. The ES has been revised to make it more concise and to report the different sections more consistently.
Rev 8	General	I find the prognostic sections very hard to read, with many details of which I am not always sure if they are relevant. The tables contain all necessary information, but a concise and at-a-glance summary table would be more helpful for each of the key questions. Such an overall table should be limited to hazard ratios and other variables in the models and perhaps one or two key components of the population; preferably on line for each study. Also, a forest plot might have added value here, showing all HRs plus their confidence intervals (without pooling).	We have moved the current tables to the appendices. We agree with the reviewer that more concise tables may be easier to assess than the current tables. However, when you have over 200 studies, it is still onerous to place this in a single table. Our time constraints do not permit revision of the current tables, but we will develop these for subsequent publications.
Rev 8	disc/conclusion	Overall, I miss references to clinical relevance for each of the key questions. Not sure what the implications are of all these findings. I did find the future research sections clear.	We have included this in the Discussion and conclusion for each section.
Rev 8	results	Shouldn't patient important outcomes be mentioned under 'research gaps' for the diagnostic studies? Even if the BNP would be a perfect diagnostic, does that mean that the patient would benefit from using BNP? This looks like what KQ6 assesses, but then at diagnosis (and not in already diagnosed HF patients)?	Many (including reviewer #1) would suggest that this work has already been done and does not need to be stated as future research recommendation. Does BNP add "incremental value" To rule out HF – BNP does add incremental value.

Commentator & Affiliation	Section	Comment	Response
Rev 8	General/disc	Most of the studies included will have had the aim of 'finding the best predicting model'. These studies may not have had the aim to assess the independent predictive value of BNP. Also, all variables in models like these add to the total model and confounding is usually not an issues in these models (because the authors may not be interested in the value of one predictor, but in the best combination of several predictors). In this report, the authors are interested in the (added) value of BNP alone, or as independent predictor. This makes interpretation of the results and conclusions difficult. I would therefore like to encourage the authors to go beyond a summary of the crude results and to provide a bit more explanation to these results. If I would have a HF patient with elevated BNP, what does this mean for the patient? If I know the age and sex of a person, do I really need to measure BNP on top of that? These are the things that I actually want to know.	All eligible studies for KQ3 in essence evaluated "predictor" finding studies (showing that BNP/NTproBNP is an independent predictor). KQ4 addresses and evaluates those studies that included computational techniques to assess "added value" in comparison to other important prognostic factors. The use of prediction models requires the input of a number of variables into the model. This is not a simple task especially considering the log-linear relationship of NP's to outcomes. We have suggested that tools are developed and evaluated before the NP's are widely used in clinical practice for prognostic purposes.
Rev 8	General	The report is reasonably well structured and organized. My personal preference would only be to move the very elaborate tables under results to an appendix and replace them by tables that are more compact. I would also like to see more figures throughout the results section, rather than in the appendix. But this is only personal preference.	We have moved the tables and reduced the information in the prognosis tables to a single line per study. The original tables and the forest plots can be found in the appendixes.
Rev 8	General	General Comments: The research questions are clearly stated and the clinical rationale clearly described. The target population and key elements of each question are explicitly defined.	We thank the reviewer for the positive feedback
Rev 8	Intro	Introduction: The introduction is well written and concise.	We thank the reviewer for the positive feedback
Rev 9	General	For KQ 1 and 2, I am unclear from the disc, the authors thoughts on the clinical implications of the results. Given a conclusion that BNP, NT-proBNP have good diagnostic performance and useful tools for ruling out HF, to what extent will their use alter the current diagnostic pathway for the diagnosis of HF (described in the report as symptoms, signs, followed by CXR, echocardiagram)? Do the sensitivity estimates reported suggest symptomatic patients who test negative on BNP do not need further testing with other tests e.g. CXR to rule out HF? If so, what are the clinical implications? If some further testing will be routine e.g. CXR, is there evidence to define the incremental diagnostic value of BNP and NT-proBNP compared to these tests?	Yes - the implication of a negative result of a good "rule-out" test is that further specific testing (e.g. Echocardiography) is unlikely to be helpful. General diagnostic tests, likely ordered a the same time as the BNP (e.g. Chest x-ray) will continue to be ordered

Commentator & Affiliation	Section	Comment	Response
Rev 8	Intro	Clarity and Usability: Well structured and organized.	We thank the reviewer for the positive feedback