Comparative Effectiveness Research Review Disposition of Comments Report

Research Review Title: *PCA3 Testing for the Diagnosis and Management of Prostate Cancer.*

Draft review available for public comment from May 7, 2012 to June 7, 2012.

Research Review Citation: Bradley LA, Palomaki G, Gutman S, Samson DJ, Aronson N. PCA3 Testing for the Diagnosis and Management of Prostate Cancer. Comparative Effectiveness Review No. 98. (Prepared by the Blue Cross and Blue Shield Technology Evaluation Center Evidence-based Practice Center under Contract No. 290-2007-10058-I). AHRQ Publication No. 13-EHC030-EF. Rockville, MD: Agency for Healthcare Research and Quality; April 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: http://effectivehealthcare.ahrq.gov/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productID=1449
Published Online: April 1, 2013

Commentator and Affiliation	Section	Comment	Response
KI # 1	Executive Summary ES-11	The term "area under the curve" (AUC) is used. An explanation of the meaning of the term is provided. Some explanation should be provided. I am not sure that the information provided on page 96 is sufficient: "area under the receiver operating characteristics (ROC) curve"	Footnote added on page ES-12 with a definition of AUC.
KI # 1	Executive Summary ES-14	The statement "The finding that the relative performance of PCA3 versus tPSA elevations is not dependent on biopsy history is a new observation that could impact future studies" appeared with no earlier explanation. The basis for this observation does appear much later in the document but should be clarified.	This statement (now p. ES-6) was based on results of analyses presented in two detailed paragraphs under *Key Questions 1 and 2: Initial and Repeat Biopsies* (pp. ES-4, 5). Briefly, regression analysis of AUC difference (PCA3 – tPSA) versus the proportion of study subjects having initial prostate biopsy was performed and the slope was not significant (p=0.97), indicating no relationship between biopsy status and AUC difference (for tPSA comparator only). The same regression analysis conducted in different datasets (e.g., including/ excluding 'grey zone' studies, stratified by assay type) consistently found no significant slope. In addition, very similar median AUC differences (PCA3 – tPSA) were found for studies enrolling all men having initial biopsy and studies enrolling all men having repeat biopsy.
TEP # 2	Executive Summary	The Executive Summary does justice to the report.	Comment acknowledged.
TEP # 4	Executive Summary ES-13, line 9	Clarify that the "consensus observed ROC curves" were calculated as median sensitivities and specificities across the studies. This could be clarified in the text or Figure ES 1 legend.	This has been clarified by adding a sentence to the ES-1 Figure legend.
TEP # 4	Executive Summary ES-13, line 10	It might be helpful to provide some brief details regarding the "modeled ability of PCA3 and tPSA".	A sentence has been added on modeling: "This modeling is based on fitting overlapping Gaussian distribution parameters to the observed summary ROC curves (Figure ES-1)." While the ES must be a stand-alone document, readers of the report will have access to a complete description of the methods used in Figure 9 and Appendix J.
PR # 8	Executive Summary ES-10, line 18	It is stated several times that this report is not about PSA and prostate cancer screening (page 7, line 4). I agree. The controversy surrounding PSA and prostate cancer screening is mentioned several times in the introductory materials. I think you must be careful in how much of the controversy you wish to discuss. "A recent evidence review found no reduction in prostate cancer-specific mortality based on tPSA screening and the test's low specificity has led to harms (page ES-10, line 18)."	We have considered all the comments on this point and reworded these sections in both the ES and the Introduction.

Commentator and Affiliation	Section	Comment	Response
TEP # 9	Executive Summary p. 10 of 174, line 35	p10 of 174, line 35: consider defining tPSA as total PSA; later on in the Exec Summary you use fPSA for free PSA, so put fPSA in a parenthesis in this line to take care of this.	We agree and have gone one step further and used %fPSA to avoid confusion with absolute values that appear in some studies.
TEP # 9	Executive Summary	If there is room, a very brief description of these other metrics.	Length is an issue, but we have added a brief paragraph about the comparators.
TEP # 9	Executive Summary	Some more description of the methods for Figure ES-1. Based on the description you give in the main report, this is some sort of a composite ("summary") ROC obtained by local summaries of TPRs for given FPRs. For descriptive purposes it is appropriate.	A sentence has been added on modeling: "This modeling is based on fitting overlapping Gaussian distribution parameters to the observed summary ROC curves (Figure ES-1)." While the ES must be a stand-alone document, readers of the report will have access to a complete description of the methods used in Figure 9 and Appendix J.
TEP # 9	Executive Summary p. 16, line 14	Perhaps add methods for the joint meta-analysis of multiple tests that have analyzed 2 or more tests in the same patients. The methods should allow comparison of sensitivities and specificities across the 2+ tests capitalizing on the within-study correlations.	Unfortunately, none of the studies provided either joint performance tables or access to original data. Thus, this is not possible.
TEP # 13	Executive Summary	The executive summary and introduction were very informative and appeared to be complete. I have included some suggestions for consideration below.	Suggestions were reviewed and responses found below.
KI # 1	Introduction	The key questions are appropriate and explicitly stated.	Comment acknowledged.
TEP # 2	Introduction	Very complete and insightful.	Comment acknowledged.
TEP # 3	Introduction	The introduction is well written and the information appears to be accurate and up to date. The key questions are well presented and the analytical frameworks are fairly clear.	Comment acknowledged.
TEP # 4	Introduction	The Introduction provides a nice summary of relevant background information. The target population and audience are explicitly defined. The key questions are appropriate and explicitly stated.	Comment acknowledged.
PR # 5	Introduction	Introduction clearly explains the context.	Comment acknowledged.
PR # 5	Introduction p. 2	I sense a bias against PSA-based screening in the tone of the introduction - the authors state that the harms are clear and benefits are uncertain, whereas, by my assessment of the data, there are clear harms and clear benefits. For example, on page 2, with regard to the 40% reduction in prostate cancer mortality in the past 2 decades, the authors assert that "While some association with increased tPSA screening is likely, the effect size is currently unknown. However, it is known that the low specificity of the test has subjected many men to unnecessary prostate biopsy (false positives) and to overdiagnosis of indolent cancer..." In my opinion the evidence for an effect of screening is as strong as the evidence for overdiagnosis.	This was certainly not our intent, but rather reflected the current published literature that we identified. PSA-based screening remains controversial, and both points of view were referenced.

Commentator and Affiliation	Section	Comment	Response
PR # 5	Introduction	See for example, the work of the CISNET group, the ERSPC, etc. Stating benefits as speculative, but harms as fact is misleading. Using citations from the literature for harms and none for benefits demonstrates a bias for reporting harms rather than benefits.	CISNET and ERSPC and more recent publications have been reviewed and considered in the revision of text on this controversial point.
PR # 6	Introduction	Introduction is sufficient.	Comment acknowledged.
PR # 7	Introduction	The report assesses the utility of PCA3 in management of men that either are about to go to biopsy, have already had one negative biopsy and are headed for a second biopsy or are about to have a prostatectomy. For each setting, the goal set by the review authors is to see if the studies support short term assessments such as diagnostic accuracy either alone or as supplements to other available tools, intermediate outcomes that may help with management choices (e.g. degree of watchful waiting or choices of therapy in the context of patients already diagnosed) and impact on mortality and morbidity.	Correct and comment acknowledged.
PR # 8	Introduction	If you are going to quote the American Preventative Services Task Force from 2008, then I think you must also mention the results of the two randomized trials of prostate cancer screening, the PLCO and ERSPC trials. Updated results from the larger trial, the ERSPC trial, were published in NEJM, March 15 2012, showing a 21% relative risk reduction in prostate cancer-specific mortality at 11 years follow-up. Maybe it is enough to simply state that the use and effectiveness of prostate cancer screening with PSA remains controversial.	Now that the report is available, reviewers have asked that we include mention of the 2011 evidence report and 2012 USPSTF Recommendation. We have also reviewed the NEJM paper from March 15, 2012. We agree with your suggestion to make the point that this remains controversial.
PR # 8	Introduction p. 3, line 6	I would delete “infection” as a complication of prostatectomy. Though possible, it is so rare that mentioning it here in a discussion about the morbidity of surgery is irrelevant, in my opinion.	This has been removed.
PR # 8	Introduction p. 4, line 48	delete “with”	Done
PR # 8	Introduction p. 4	You mention all of the manipulations of PSA that can be performed to increase specificity and sensitivity under PCA3 comparators. However, you do not mention age-adjusted PSA. This may not be something that you compared PCA3 to directly, but I think it is worth mentioning somewhere in the manuscript. (Oesterling JE, Jacobsen SJ, Chute CG, Guess HA, Girman CJ, Panser LA, et al. Serum prostate-specific antigen in a community-based population of healthy men: Establishment of age-specific reference ranges. JAMA 1993; 270:860.) Age adjustment of PSA thresholds is almost universally performed by urologists when counseling patients about whether PSA results are abnormal and whether to proceed to biopsy. It is also part of the most recent PSA screening guidelines from the AUA.	Age-adjusted PSA was mentioned in only one study, but we agree that this needs to be included and have added a brief description and the Oesterling reference to the introductory section on total PSA.

Commentator and Affiliation	Section	Comment	Response
PR # 8	Introduction	Additionally, typical clinical practice takes into account much more than a single PSA value when deciding whether to proceed with a prostate biopsy. Current AUA guidelines for PSA screening recommend another serum PSA when presented with an abnormal value. Additionally, other considerations such as age-adjusted reference ranges (see above), BPH and prostate size, previous PSA values (PSA velocity) if available, free PSA if available, family history, race, etc. are considered when considering prostate biopsy. This may or may not influence the sensitivity and specificity of prostate cancer screening with PSA.	Agreed. Some of these concepts were stated but others have been added and the section rewritten for clarity.
KI # 9	Introduction p. 10, line 7	Prostate cancer is the most common non-skin cancer in men.	Noted, and this has been added.
KI # 9	Introduction p. 10, line 14	The rationale for PSA-based screening was NOT reduced prevalence of the disease, that's not possible with a screening test. Screening a previously unscreened population will always lead to a higher prevalence of disease. The purpose is to reduce the prevalence of late stage disease or advanced prostate cancer.	You are correct that this was stated in a confusing manner. It has been modified to reduction of prevalence of advanced prostate cancer and of prostate cancer-related mortality based on early detection.
KI # 9	Introduction p. 11, line 39	What do the authors mean by "risk factors" in the prognostic comparators comment? Are they referring to African American race and family history?	Yes, the parentheses directly following the term 'risk factors' was (e.g., age, family history). We have added race to the parenthetical examples.
KI # 9	Introduction p. 21, line 56	The authors should discuss that PSA screening has come under fire with the recent final USPSTF D recommendation.	This has been updated and added to the ES and Introduction.
KI # 9	Introduction p. 22, line 35	The risk stratification of newly diagnosed prostate cancer isn't just to inform candidacy for active surveillance. It affects use of hormones in patients getting radiation therapy, can help make decisions about surgery vs radiation, etc.	The text has been revised to reflect this other aspect of risk stratification.
KI # 9	Introduction p. 22, line 52	There are well established treatments for all risk strata of prostate cancer.	This point is noted and the text has been modified.
KI # 9	Introduction p. 23, line 12	Irritative proctitis and rectal urgency are similar: rectal urgency is a symptom of irritative proctitis.	This is correct. The term irritative proctitis is used and rectal urgency has been removed from the text.
KI # 9	Introduction p. 23, line 20	Ryo and HIFU are being mostly used for whole gland treatment but newer studies are being conducted of focal treatment with these modalities.	This point is noted and focal treatment has been omitted.
KI # 9	Introduction p. 24, line 29	There is no evidence that "many" urologists use a cutoff for PSA of 2.5-3. Selected experts have proposed this cutoff, but it hasn't ever been shown to have been adopted widely	This does seem to be an open question and this will be clarified, along with discussion of age-related cutoffs.
KI # 9	Introduction p. 25, line 13	Do the authors mean "National Comprehensive Cancer Network?"	Indeed that was intended and this has been corrected.
KI # 9	Introduction p. 25, line 31	To clarify, PSA doubling time is used to evaluate patients after a prostatectomy when the only PSA increase should be from cancer.	This has been clarified in the text.
KI # 9	Introduction p. 26, line 25	The most commonly used risk stratification algorithm is from D'Amico. The Epstein criteria are for low risk cancers and they correlate with prostatectomy findings. The Epstein criteria are almost exclusively used for selection to active surveillance.	Agreed, and this has been clarified in the text.

Commentator and Affiliation	Section	Comment	Response
TEP # 11	Introduction	Clear & to the point. Adequate description of the context for the report.	Comment acknowledged.
PR # 12	Introduction	This is appropriate, but would benefit to results that have been presented at the AUA regarding the PIVOT trial, which shows that at 12 years radical prostatectomy does not improve survival of prostate cancer patients when compared to observation. This trial has been accepted for publication. The point of mentioning this is that even if PCA3 were better at identifying cancer, it is probably not useful unless it identifies high-grade cancers.	The PIVOT trial and its implications have been added to the discussion with regard to applicability.
PR # 12	Introduction	It would be useful to include more recent studies showing marginal utility of PSA screening in the introductory material, and the most recent recommendation of the US Preventive Services Task Force, which was not available at the time this report was prepared.	Agreed and this information has been added.
TEP # 13	ES10, line 20	Refers to a recent evidence review and gives 3 references. Only one of these references appears to be the actual review, and instead of citing the latest 2011 update of the USPSTF review on PSA screening, the 2008 update is cited.	Agree incorrect references entered. Language was clarified and references corrected.
TEP # 13	ES-10, line 23	Refers to the PCA3 tests as "noninvasive" and "from urine." While I understand what is meant here, it just struck me as a curious choice of wording given the requirement for attentive DRE, at least with the Progensa test. Would it be sufficient to simply say that the tests are done using a urine sample, or to qualify the noninvasivity with information on the prep?	Even though this test is not invasive, it is a reasonable point that it does require manipulation; the term “non-invasive” was removed.
TEP # 13	ES-10, line 30	Mentions FDA approval of PCA3 assay for repeat biopsies. It may be worth mentioning the warning about not using in men with atypical small acinar proliferation on most recent biopsy.	This information was added.
TEP # 13	ES-11, line 20	States that searches were run on MEDLINE, however, Appendix A refers to searches in PubMed...these are not exactly the same thing	Very true; this has been clarified as PubMed searches.
TEP # 13	ES-12, line 9	Not clear what is meant by "we anticipate a qualitative analysis."	Clarification added.
TEP # 13	ES-12, line 13	Describes PICOT, whereas other places in the report have PICOTS (including setting).	PICOTS is correct; change made.
TEP # 13	ES-14, line 31	Appears to contain the citations for 11 studies listed twice instead of just once.	Duplicates were removed.
TEP # 13	ES-14, line 47	May want to capitalize "insufficient" to match usage elsewhere in the report Introduction:	Strength of evidence has been capitalized and italicized throughout. Quality of individual studies as *good*, *fair* and *poor* has been italicized.
TEP # 13	p. 2, lines 29-39	May want to include something about risk and potential harms associated with biopsies.	Agree and this has been added.
TEP # 14	Introduction	Thorough, informative, sets forth relevant background and key issues surrounding PCA3 testing in the context of prostate cancer screening. Key questions clear, appropriate, clinically relevant with populations well defined.	Comment acknowledged.

Commentator and Affiliation	Section	Comment	Response
TEP # 14	Introduction	An important potential benefit of the use of PCA3 testing would be a reduction in unnecessary prostate biopsies in patients with elevated PSA levels, hence the FDA approved indication for patients undergoing rebiopsy. However, based on the systematic review, any such benefit presently associated with PCA3 testing seems as if it may owe less to the performance characteristics of PCA3 testing itself or our knowledge of them, than to the poor performance of PSA. Rather, there appears to be an inherent benefit in reducing prostate biopsies based on PSA elevation generally because of a limited likelihood of clinical benefits from the procedure, together with a concomitant risk of harms from biopsies and the potential for overtreatment of the underlying disease.	We agree with this assessment.
KI #1	Methods	This appears to be an excellent study which has been conducted with sound methodology. The outcome measures are appropriate. The target audience and population is explicitly stated.	Comment acknowledged.
KI # 1	Methods	This study doesn't have exclusions and exclusion critieria.	Inclusion and exclusion criteria can be found on pp.14-15 under the heading of Study Selection.
TEP # 2	Methods	The inclusion and exclusion criteria are very appropriate considering the complexity of this literature. The requirement for "matched" patients is justified.	Comment acknowledged.
TEP # 3	Methods	The literature (including grey literature) search strategies are well described, as is the data extraction methodology. Individual Study Quality Assessment methods are also well described. The outcomes measured appear to be appropriate, as is their definitions and diagnostic criteria. Statistical analysis methodology seems appropriate.	Comment acknowledged.
TEP # 4	Methods	The methods are well-described and appear to be appropriate. More specifically, the search strategies are explicitly stated and logical. Appropriate definitions of the outcome measures are used.	Comments acknowledged.
TEP # 4	p. 18, lines 30-35	p. 18, lines 30-35: The numbered list has incorrect numbering.	This has been corrected.
PR # 5	Methods	The methods are appropriate.	Comment acknowledged.
PR # 6	Methods	Target populations are well defined and key questions are explicitly stated.	Comment acknowledged.
PR # 6	Methods	Substantial amount of excluded studies: What type of design flaw lead to judgment that design is invalid in the excluded studies?	The most common reasons for a 'design flaw' exclusion were studies that looked at PCA3 and one or more comparators in an appropriate population but were not matched studies (i.e., biomarkers were not tested in the same population of men), or matched or unmatched studies that looked at comparators but did not include relevant data on PCA3.

Commentator and Affiliation	Section	Comment	Response
PR # 6	p. 147 (?)	Which methodology was used to "account" for verification bias?	We attempted to account for verification bias by creating overlapping Gaussian curves that fit both the observed ROC curve for tPSA and also the 'shape' of the reported tPSA distributions in both men with and without a positive biopsy. In order to account for the bias, the means of the tPSA distributions needed to be lowered, and the expected change in variance estimated. The expected modifications were created by modeling an unbiased population distribution and then assume uptake of biopsy was correlated with the extent of tPSA elevation. This is explained in more detail in Appendix J
PR # 6	p. 13, line 3	It is not clear how exactly "modeling was used to account for potential impact of verification bias"	We attempted to account for verification bias by creating overlapping Gaussian curves that fit both the observed ROC curve for tPSA and also the 'shape' of the reported tPSA distributions in men with and without a positive biopsy. In order to account for the bias, the means of the tPSA distributions needed to be lowered, and the expected change in variance estimated. The expected modifications were created by modeling an unbiased population distribution and then assume uptake of biopsy was correlated with the extent of tPSA elevation. This is explained in more detail in Appendix J
PR # 7	Methods	The authors did an extensive search on Medline and other databases and looked at grey literature and narrowed the list of references to those that could potentially help in answering these questions. The search strategies are explicit and I cannot judge if excluding data from non-English sources is a reasonable approach except that the report is trying to address a US population and there are a number of studies from outside the US in this summary. I am a statistician and cannot comment on the completeness of the studies used in the report.	Though we did exclude a small number of articles that were not available in English, we reviewed articles reporting studies from the US, Europe, Asia and Africa that were published in English.

Commentator and Affiliation	Section	Comment	Response
PR # 7	Methods	I believe the inclusion and exclusion criteria for the articles selected are reasonable. In order to facilitate a comparison of a new test associated with managing patients with a potential for prostate cancer means only including studies with direct comparisons to more traditional tests such as tPSA and its variants. Because the inclusion and exclusion criteria within each article varied so much, it was hard to understand the value of each metric when going across studies. For example, not all the studies used the same PCA3 assay, some articles used different cutoffs for the PCA3 assay and others used different ranges of tPSA for eligibility. This is not the fault of the authors but is hard on the reader.	You are correct that there were differences in key variables across studies, but some could be assessed to some extent. 76% of studies used Gen-Probe reagents and 59% specified PROGENSA. One analysis looked at studies that used 35 as a cutoff; another used the ROC curves to assess performance, which eliminated the issue of cutoff. tPSA cutoffs did vary (mainly between 2 and 4 ng/mL), but those that focused on the 'grey zone' were noted and stratified. Initial vs repeat biopsy was accounted for. More studies are needed to address impact of race, family history, DRE results and other variables.
PR # 7	Methods	Using summary data is always a weakness when a meta-analysis is performed.	Agreed, comment acknowledged.
PR # 7	Methods	There did not appear to be pre-specified statistical analysis plans in support of each of the three questions raised. I believe the authors recognized the weakness of combining such disparate studies in their interpretations and perhaps that drove analysis choices.	For all three KQs, we outline an overall study plan that focused on results in matched populations. For KQ1/ KQ2, we expected to find matched analyses, but none were reported by individual studies; matched analysis could be derived in only one small study. For KQ3, we outlined a qualitative analysis and this was done as planned.
PR # 7	Methods	Working with summary data is challenging if one wants to determine if PCA3 is a logical supplemental piece of information on top of other data the physician may have. Nomograms are one way of combining data but if the article did not do it, the authors could not take the data in the article and apply the rules.	Comment acknowledged.
PR # 7	p. 37	I felt that many of the analyses were designed to compare PCA3 to tPSA and other analytes as if PCA3 was being considered as a replacement. But many of the studies were evaluating men already scheduled for biopsy or prostatectomy and perhaps the more logical analyses would have been to see if PCA3 could improve the prediction of cancer when used to augment tPSA. As noted by the authors, when the range of tPSA is restricted by the way patients were selected for the study, the diagnostic value of tPSA could be underestimated and this in turn could impact the relative comparisons of tPSA and PCA3.	This review was designed to summarize the literature with regards to the key questions. The reviewer brings up an interesting point that was part of the key questions 1 and 2. Unfortunately, only a small number of studies address this point and we were unable to summarize the performance of an externally validated model with/without PCA3. However, we have summarized the data showing the two markers are relatively independent. The point you raised is also noted and has also been made in the Gaps in Knowledge section.

Commentator and Affiliation	Section	Comment	Response
PR # 7	p. 52	There are comments made indicating that because physicians typically do not use PSA density in determining a decision to send a patient for biopsy, that this measure is not impacted by verification bias. It is likely that tPSA and PSA density are correlated because tPSA is used in the calculation of PSA density and therefore if verification bias impacts the distribution of tPSA it likely impacts the distribution of PSA density.	There was a statement that verification bias might be less of an issue for PSAD because it is not routinely used in all men with a tPSA/DRE positive result and would, therefore, be less strongly associated with biopsy uptake. However, the reviewer is correct that the tPSA verification bias can influence other PSA related markers. This now reads: verification bias "..would be less likely to have been an issue for the other comparators (e.g., %fPSA, PSA density), but the extent of this bias is likely related to the correlation between that marker and tPSA measurements. In addition, this correlation may be low because these comparators were not routinely used in all men with a tPSA/DRE positive result and may, therefore, not be strongly associated with biopsy uptake." (p. 38)
PR # 7	Methods	In the assessment of diagnostic accuracy, we are told that the distributions of PCA3 and PSA in these studies are skewed. I agree. However, I found the resulting discussions of z-scores confusing. Z-scores are regularly used in the assessments of anthropometric measurements and are highly dependent on what population is being studied. Use of Z-scores relies on an assumption of the data being approximately normally distributed. I think the discussions of these analyses need to be clarified. If the same study is used and the data is not transformed, a correlation of z-scores and the original measurements is the same. It is clear that a log transformation was used but I couldn't follow the rest.	Using z-score (or some other normalizing factor) is necessary when comparing the separation of +/- biopsy measurements for PCA3 and another comparator. After log transformations, the PCA3 difference was expressed as a measure of the pooled log SD. A similar analysis was done for the comparators (if data were available). These z-score summaries can then be used as a common unit to compare the separation provided by PCA3 and the comparator.
PR # 7	Methods	All meta-analyses that rely on published reports can have bias because of publication bias. About half the studies were either sponsored by one company or the investigators stated they had a conflict of interest. I wonder if the company is aware of other studies that were not published because of negative findings. Having access to line data would have been quite useful, even if only for half the articles. There are key measures such as age and ethnicity that impact risk of prostate cancer and accounting for these in an analysis would have improved power and interpretability. Many of the studies did not record ethnicity at all and that is clearly a shortcoming.	We agree that examining publication bias is an important component of the review process. The only analysis that had sufficient numbers of studies to have some power to identify bias was the AUC results for PCA3 and tPSA. Figure 7 shows that there was no obvious publication bias present. Other potential confounders could not be examined because there were insufficient data reported (either at the study or sub-analysis level).
PR # 8	Methods	All clearly explained and appropriate	Comment acknowledged.
KI # 9	Methods	Inclusion and exclusion criteria for submitted articles seem appropriate	Comment acknowledged.

Commentator and Affiliation	Section	Comment	Response
KI # 9	p. 12, line 7	What type of "regression analyses" were done?	The type of regression is whatever was chosen by the authors. In most instances it was a logistic regression, but the form used for the predictors varied (some used categorical some used continuous measurements). The details of the analysis reported are found in the body of the review.
KI # 9	p. 12, line 8	What type of "qualitative analysis" was considered? Did the authors preplan any qualitative methods?	This was included in the Executive Summary but missed in the narrative. This has been corrected.
Groskopf/ Gen-Probe # 10	Methods	A total of 17 studies were used to perform comparative analysis of PCA3 vs. serum PSA for predicting biopsy outcome (Table 9, page 35). Five of these studies were performed using a sample processing and/or assay procedure different than PROGENSA® PCA3: Cao (Ref 68), Hessels (74), Mearini (75), Ouyang (78) and Rigau (80). Different methodologies can yield different results, and there is no information on analytical performance or assay robustness for these research-level PCA3 tests. Gen-Probe strongly recommends either excluding or separately reporting data on other PCA3 assays so reviewers can clearly see the performance data that are specific to PROGENSA® PCA3, the only FDA-approved PCA3 assay.	We have stratified the results into PROGENSA and non-PROGENSA test methodologies and found little difference. The results are summarized in the discussion regarding AUC differences (p. 47).
Groskopf/ Gen-Probe # 10	Methods	The AHRQ analysts found that all published studies describing PCA3 clinical performance were of poor quality. There were various reasons for the analysts' conclusion, but one of the key criteria was bias: The AHRQ analysts used serum PSA as the comparator method, and most study subjects already had elevated PSA levels. All studies included men already scheduled for prostate biopsy. We acknowledge these potential sources of bias. However, to avoid verification bias, some patients would have had to undergo biopsy regardless of serum PSA level or clinician judgment (i.e., men not recommended for biopsy). This is not generally feasible or practical, especially for repeat biopsy studies since there are no clear criteria to guide repeat biopsy decisions. Due to published study design, the AHRQ analysis might be better described as an assessment of the incremental improvement in predictive accuracy when PCA3 is used in addition to PSA for men undergoing prostate biopsy. A true comparative effectiveness study can only be done in the absence of pre-screening with PSA. To date, such a study has not been performed.	We agree that the key criteria for calling all individual studies addressing KQ 1 and KQ 2 *poor* quality were the source of data (opportunistic cohorts) and the potential for bias, though this was addressed to some extent through stratification and modeling. While important and needed, we note that conducting the ideal study is challenging. This study was assigned as a comparative effectiveness review, with the caveats that you have noted. Certainly, this data could also be used to assess incremental improvement in performance if PCA3 were used in combination with PSA.
TEP # 11	Methods	My personal preference would be to describe/list the methods for the various syntheses you did in this section. Instead you describe the methods in the respective parts of the results. (In the end, it works, apparently either way is OK.)	We find it more approachable to have the methods located close to the results. Had the results become repetitive with multiple explanations of the methods, we would have chosen to move a more complete description into the Methods section.
PR # 12	Methods	Inclusion and exclusion criteria are appropriate, and the search strategies and methods for comparison are well-chosen and appropriate.	Comment acknowledged.

Commentator and Affiliation	Section	Comment	Response
TEP # 13	Methods	The Methods section was very clear and all measures and outcomes appear to be appropriate.	Comments acknowledged.
TEP # 13	p. 33, line 57	See previous comment regarding MEDLINE/PubMed provided for the executive summary section.	This has been corrected to PubMed.
TEP # 13	p. 34, line 8	As above	This has been corrected to PubMed.
TEP # 14	Methods	Overall methodology appropriate and justifiable. Creative use of statistical analysis and modeling to attempt to answer key questions in the face of limited data. This is appropriate and a positive feature as long as strengths and weaknesses of approach well described, and conclusions appropriately weighted as they were in this systematic review.	Comment acknowledged and appreciated.
TEP # 14	Methods	In conclusion, the review is comprehensive, thoughtful, and fair. The strengths and weaknesses of the review were well described. The authors were creative in applying statistical analyses and modeling techniques in an attempt to deal with an incomplete data set that did not allow for independent analysis of the first 2 key questions. Unfortunately, an absence of high quality studies assessing the use of many diagnostic tests is common, and this type of improvisation and the development of other such pragmatic approaches to data evaluation can help inform clinical practice and direct future research.	We agree with your stated concerns about the quality of studies and the incomplete descriptions of data.
KI # 1	Results	The details present in the results are appropriate. See attachments for details.	Comment acknowledged; attachments reviewed.
TEP # 2	Results	The detail in the results section is appropriate considering the critique this report will receive. I think the amount of material and clarity of the analysis that appears in the appendices is extraordinary. This was a very nice balancing act.	Comment acknowledged.
TEP # 3	Results	The results are presented in great (perhaps excessive) detail. Study characteristics are clearly described. I'm not familiar with any studies that were overlooked.	Comment acknowledged.
TEP # 4	Results	An appropriate amount of detail is provided. The key messages are explicit and applicable. As noted in a few specific comments below, stronger rationale could be provided for a few analyses and a couple additional analyses may be helpful.	Comment acknowledged.

Commentator and Affiliation	Section	Comment	Response
TEP # 4	p. 29, last paragraph	Only 2 of the 11 studies have no men with an initial biopsy. Therefore, it would be informative to calculate the slope between proportion of men with an initial biopsy and AUC difference if analysis is restricted to the 9 studies with men receiving initial biopsy.	In the 9 remaining studies, 1 enrolled 100% initial biopsy men and the remaining 8 reported on mixtures of about 51-82% initial biopsies. It is not clear to us what the hypothesis would be for regressing only those studies that have some proportion of initial biopsies, since the 8 mixed studies are also 18-48% repeats. This would only be useful if the mixed studies provided separate results for initial and repeat biopsy enrollees. Three studies added for the final draft did provide stratified results, and no relationship was observed. Also, the same regression analysis conducted in different datasets (e.g., only 0% and 100% initial biopsies, including/excluding 'grey zone' studies, stratified by assay type) consistently found no significant slope. Results of additional analyses have been added to the text.
TEP # 4	p. 30, line 37	Stronger rationale should be provided for considering a constant specificity of 50%.	For the purposes of the summary, 50% specificity was chosen as it is in the region where the greatest separation between two tests might occur. At very low, or high specificities, the two tests will be most likely similar.
TEP # 4	p. 30, line 38	Only 2 of the 10 studies have no men with an initial biopsy. Therefore, it would be informative to calculate the slope between proportion of men with an initial biopsy and sensitivity difference if analysis is restricted to the 8 studies with men receiving initial biopsy.	The requested analysis has been performed for 15 studies for which difference in sensitivity at a constant specificity of 50% could be computed, and can be found in Figure 6 of the final draft. Again, the slope is not significant (p=0.79) when 4 'grey zone' studies were included. Excluding the 4 'grey zone' studies moved the slope closer to zero.
TEP # 4	p. 31, lines 33-40	Stronger evidence could be provided to justify that combining results from studies of initial biopsies, repeat biopsies, and mixtures of initial and repeat biopsies will not impact the comparison of PCA3 with tPSA elevations.	Additional recent studies have been included in the final draft analyses, and the conclusion remains unchanged. In addition, two of the new studies provided ROC curves for initial and repeat biopsies, and an analysis comparing the differences in sensitivity (PCA3 – tPSA) at a fixed specificity regressed against % initial biopsy has been added. No significant relationship was seen. However, the limitations of study quality and strength of evidence have been discussed, along with gaps in evidence that need to be addressed.
TEP # 4	p. 36, line 11	Re-word "in order to strengthen the potential bias"	This should have read "in order to strengthen the analysis of PCA3 and tPSA elevations", and has been corrected.

Commentator and Affiliation	Section	Comment	Response
TEP # 4	p. 37 (Table 10)	It is not clear why several of the values in the "Initial Bx" column differ from the values in the "Initial Biopsy" column in Table 7.	These typographical errors have been corrected.
TEP # 4	p. 39, line 37	Clarify that the "summary estimate of 0.118" is the median difference	That is correct, and this estimate will be reviewed as part of the reanalysis of data with additional studies from the undated search.
TEP # 4	p. 39, line 43	Clarify that the "consensus estimate" is the median difference	That is correct.
TEP # 4	p. 42, line 20	Replace "range" with "ranging"	The source of this comment was not found. All instances of "range" were reviewed.
TEP # 4	p. 42, lines 27-29	Stronger justification should be provided for selecting the "middle table".	The paragraph containing this phrase was intended to describe a method for estimating a difference in sensitivity between PCA3 and tPSA in the absence of 2x2 matched tables. This paragraph was deleted in the next draft and replaced with a discussion of heterogeneity as a way to assess consistency in estimates (page 49). This change from quantitative to qualitative assessment of statistical significance of heterogeneity is based on the lack of relevant data in nearly all studies to perform reliable quantitative analyses.
TEP # 4	p. 42, line 32 Table 10	Provide explanation why the two studies did not produce a reliable estimate of significance.	This preliminary approach was attempted to estimate a difference in sensitivity between PCA3 and tPSA in the absence of 2x2 matched tables, but could not be estimated in some studies (i.e., Aubin and Ochiai). This approach was replaced in the next draft with a discussion of heterogeneity as a way to assess consistency in estimates (page 49). This change from quantitative to qualitative assessment of statistical significance of heterogeneity is based on the lack of relevant data in nearly all studies to perform reliable quantitative analyses. Results of the Aubin and Ochiai studies are found in Table 12 in the revised draft.
PR # 5	Results	The results are reasonably clear.	Comment acknowledged.
PR # 6	Results	Studies are clearly described.	Comment acknowledged.
PR # 7	Results	I think the key message is very clear. All the studies appeared to address the issue of diagnostic accuracy in one or more of the contexts raised in the three questions posed. However none of the articles could address all the elements in any one of the three questions posed.	Agreed – comment acknowledged.

Commentator and Affiliation	Section	Comment	Response
PR # 7	Results	Because of the way patients were selected, notably based on PSA and possibly abnormal DRE, seeing if the AUC (PSA) values were lower than anticipated might keep a reader from over-interpreting the fact that 13 of 15 studies showed AUC (PCA3) is greater than the value for PSA. As a statistician, I cannot comment on whether key studies were omitted.	We agree. For summary data, we generally exclude the 'grey zone' studies (Table 13, Figure 11). When there are too few data to exclude these grey zone studies, we highlight them in the table and include a caution about over-interpretation.
PR # 8	Results	Very detailed, very comprehensive	Comment acknowledged
PR # 8	p. 21, Table 1	The pooled patient population is very diverse (page 21, table 1), not representing a typical screening population. For instance, the largest study (Aubin et al, reference 66) is from a dutasteride chemoprevention trial – with strict PSA criteria and negative biopsy required for entry.	We agree that the study inclusion criteria were broad. This was to ensure the largest number of studies reviewed. Of interest, the results from Aubin fall with the summary of other study findings.
PR # 8	p. 45, Figure 11	PCA3 is better than PSA at predicting prostate cancer. Is there any clinical relevance? This cannot be answered by this review, and that is discussed in the manuscript.	Agreed – comment acknowledged.
PR # 8	Results	The manuscript discusses the available data regarding PSA velocity and other PSA manipulations, and determines that there is insufficient data to make a comparison to PCA3. So the clinical use of PCA3 remains unclear.	With regard to comparison of diagnostic accuracy of these comparators with PCA3, that is correct.
PR # 8	Results	Your results do not necessarily support the FDA approved indication for this test. Additionally, the cost of performing PCA3 testing will have to be considered when deciding how it should be used in clinical practice, especially in relation to PSA.	A study providing further data relevant for the FDA review has not yet been published, and we do not know whether the inclusion criteria for this CER would be met. The FDA summary did not provide a matched study for inclusion in this review. In addition, comparison to the FDA process is outside the scope of this review**.** We agree that cost will need to be considered at some point, but usually such analyses are performed when data are available to support a clinical utility claim. In addition, a cost effectiveness or other cost analysis was outside the scope of this review.
KI # 9	ES-14, line 31	Duplicate reference callout.	Noted and corrected.
KI # 9	p. 71, line 18	The authors state that the risk stratification schemes are illegitimate as they are not correlated with clinical outcomes. The most famous of these, the D'Amico classification, was derived from correlation with clinical outcomes. Same with the designation for clinically "insignificant" cancers based on an examination of tumors found to be small at prostatectomy and for which clinical outcomes were consistent with overtreated, indolent disease.	The risk classification criteria in the identified studies that included testing with PCA3 could not be assessed because there was short term follow-up only in two studies using questionably validated surrogate markers, and no continuing follow-up to important intermediate or long-term clinical outcomes stated for this review. So clinical utility could not be addressed. D'Amico was not addressed in any of these studies, but has been added.

Commentator and Affiliation	Section	Comment	Response
Groskopf/ Gen-Probe # 10	Results	We strongly encourage AHRQ to include data from the US pivotal clinical study and EDRN validation in its final report, and to delay publishing a final report until the findings from these critical studies have been incorporated…Without this essential and timely data, patients and providers will not have the information they need to adequately assess the benefit of PCA3 testing when making biopsy decisions. AHRQ's procedures support the use of 'grey literature', including abstracts/conference proceedings and regulatory documents.	The initial and updated searches did include identification of abstracts; abstracts were followed up to determine if full reports had been published and provided insight into types of data that might become available in the future (relevant to gaps in knowledge). A modified funnel plot was used to investigate potential publication bias; abstracts identified from 2 years of highly selected US meetings were unlikely to provide an unbiased sample. The protocol notes that when published as a meeting abstract only, abstracts would be excluded from data extraction and analysis. Abstracts often represent early analyses of data that may contain errors and/or analyses may change/evolve prior to or following manuscript submission and peer review. An exception would not be possible, based on these methodological concerns, and that objectivity would require equal consideration of many other previously identified abstracts.
Groskopf/ Gen-Probe # 10	Results	Haese et al (Ref 73), one of the largest prospective multi-center repeat biopsy studies to date, was excluded from the analysis due to 'duplicate data' (Page 20). Instead, a secondary analysis performed using the Haese data (Ankerst, Ref 51) was used for the report. Ankerst et al utilized statistical methodology to show that PCA3 increases predictive accuracy when incorporated into an existing risk calculator. Primary data from Haese and Deras (71) studies were used for the analysis. AHRQ's study selection criteria indicate that studies that do not report primary data should be excluded (page 15). We therefore recommend using the primary data source, Haese et al.	Ankerst provided an analysis of two data sets. The North American dataset was published in 2008 by Deras and Ankerst. Deras provided a slightly larger sample size with a clear presentation of data for three of the five analyses performed (e.g., comparative ROC curves for PCA3 and tPSA), and was selected. The European dataset was published in 2008 by Haese and Ankerst. The Haese dataset was very slightly larger, but all repeat status, central measures, positive rate, and other characteristics made it clear the two articles were reporting primary data from the same study population. Unfortunately, Haese did not provide what we found to be the most useful data, the comparative ROC curves, so Ankerst's data presentation was selected. Had Haese been selected, the study would have had little impact in the analyses and a large study would have been lost.
Groskopf/ Gen-Probe # 10	Results	PCA3 is FDA-approved for use in conjunction with PSA and other clinical information in the repeat biopsy setting only (Appendix II).	Noted, but outside the scope of this review since the study supporting the FDA review of the Gen-Probe PROGENSA assay has not yet been published. We do not know if a matched comparison of initial and repeat biopsy cohorts was performed.

Commentator and Affiliation	Section	Comment	Response
Groskopf/ Gen-Probe # 10	Results	The AHRQ analysts pooled data from initial and repeat biopsy studies. The basis for combining these groups was an analysis showing that, across different studies, the relative accuracy of PCA3 vs. PSA level was not associated with the fraction of subjects undergoing initial vs. repeat biopsy (Figures 5 and 6, pages 30-31). Although we appreciate that the pooled analysis increased the statistical power to compare PCA3 vs. serum PSA, this approach introduces the following issues: 1) PSA is used differently for repeat vs. initial biopsy decision making. For men with one or more prior negative biopsies, PSA predictive accuracy is decreased (since in most cases the initial biopsy was triggered by elevated PSA) and there are no widely accepted guidelines for managing patients in this 'dilemma' population. For men with prior negative biopsies and PSA<10 ng/mL (representing the majority of patients considered for repeat biopsy), the absolute PSA level is less meaningful and clinicians may therefore monitor for changes in PSA levels or other clinical parameters (Reference: National Comprehensive Cancer Network Guidelines for Early Detection of Prostate Cancer, Version II, 2012). 2) When the AHRQ protocol was posted for comment, initial and repeat biopsy indications were listed separately as KQ1 and KQ2. Combining these two questions therefore creates a procedural issue, since there was no opportunity to comment meaningfully on the validity of this approach. Had that opportunity been given, we are confident that the public would have emphasized the importance of analyzing the two questions separately. We therefore think that the combined KQ1/KQ2 analysis is inconsistent with clinical practice and recommend that the repeat biopsy studies be analyzed separately (including the more recent US pivotal clinical study and EDRN clinical validation data).	Issue 1. The aim of the review is to examine the underlying diagnostic performance of PCA3 versus a comparator, when measured in the same set of men (whether initial, repeat or a mixture). We agree that the median tPSA level of men with repeat biopsies will be between 6 and 10 ng/mL. In Table 10, we note the two studies of all repeat biopsies had median levels of 6.7 and 8.1 in repeat biopsy negative and 8.2 and 8.8 in repeat positive biopsies. However, in the remaining studies with a mixture of initial and repeat biopsies, the tPSA levels were essentially the same. This indicates that the tPSA levels may not differ to any great extent in the two populations. Note that the analysis easily detects that grey zone studies, as expected, have much smaller SDs (peaked tPSA distributions for Perdona and Ferro). Thus, we can only report that our data does not show the effect that you suggest. This may be relate to the matched study design or some other factor as yet unknown. Issue 2. That is true, and the corresponding groups of men having initial (KQ1) and repeat (KQ2) biopsies were, and continue to be, analyzed separately. We believe this resolves any procedural issue related to the protocol. You are taking the opportunity to comment on the validity of this approach at this time, prior to the addition of new studies and reanalysis of the data for the final draft. The issue you raise actually relates to analytic follow-up of an observation made during the early data analysis phase. We believe we would have been remiss had we not compared results stratified by initial/repeat biopsy, and if we had not considered the larger number of studies with detailed data on proportions of initial and repeat biopsies (rather than simply excluding them without review). With the subsequently identified additional studies (three of which provide separate results for the initial and repeat groups), the data continue to suggest that repeat status is not an important covariate.

Commentator and Affiliation	Section	Comment	Response
			Limitations of study quality and strength of evidence are discussed, along with gaps in evidence that need to be addressed. In fact, this could be an important finding with regard to clinical practice, and an impetus for those publishing such studies to provide raw data or matched analyses.
Groskopf/ Gen-Probe # 10	Results	Table 11 (page 40) indicates that, in Ankerst et al (51) median PCA3 Scores for biopsy positive and negative subjects were 34.3 and 34.2, respectively. An erratum to this study was published (J Urol 181(3):1507, March 2009) and the correct median PCA3 Score for biopsy negative subjects is 19.4. This is a significant error (Figure 9, page 41) that must be corrected in the report (if AHRQ does not use the Haese study instead as recommended above).	This has been corrected and we appreciate direction to this erratum. These data were reanalyzed with the newly identified studies.
Groskopf/ Gen-Probe # 10	Results	Data from larger, multi-center studies have more weight and should be considered more strongly in the assessment of PCA3 performance. Although the impact of the smaller studies was mentioned in the report, it is not clear whether this affected the overall grade for the strength of evidence.	It is true that a weighting scheme could have been developed. However, give the uniformly poor quality of studies and lack of the needed data and analysis (e.g., matched contingency tables), we did not feel that the results of the analyses would have been improved by the added complexity of weighting. A sentence was added under Limitations of the Database.
Groskopf/ Gen-Probe # 10	Results	Directness: Diagnostic accuracy (i.e., accuracy for predicting cancer at biopsy) was determined to be an "Indirect" outcome. Although the report states that diagnostic accuracy can be considered a direct outcome, the rationale given is that "some biopsies may be more indicative of serious existing (or future) disease than others", and that "none of the included studies provided evidence that positive biopsies identified by PCA3 were at least as serious as those identified by tPSA". These statements are true, but also irrelevant to KQs1 and 2. The ability to discriminate indolent from significant cancers is covered only by KQ3. We therefore disagree with the analysts' reasoning that diagnostic accuracy (biopsy outcome) is an indirect outcome. Furthermore, since prostate biopsy is the only reference standard method for diagnosing prostate cancer, there is no other clinically feasible means for assessing diagnostic accuracy.	The decision of the reviewers is to maintain the Directness domain as Indirect for the outcome of diagnostic accuracy with the comparator tPSA. We believe this is warranted based on the presence of both types of indirectness: 1) one body of evidence links the text to the intermediate outcome of diagnostic accuracy and another links the test-related intervention(s) to health outcomes; 2) based on the lack of matched analyses, it is not possible to determine the extent to which the PCA3 and comparators are identifying the tumors with the same or different characteristics (e.g., aggressiveness), and yet another body of evidence is needed to resolve this question. Since these are observational studies, the GRADE initial presumption is a high risk of bias (i.e., verification bias, selection bias and spectrum bias are possible here) and an initial SOE of Low. Even if this outcome was modified to Direct, we do not believe it would be sufficient to raise the SOE to Moderate.

Commentator and Affiliation	Section	Comment	Response
Groskopf/ Gen-Probe # 10	Results	Strength of Association: The report states that "Although there is evidence that PCA3 will be slightly better at identifying high risk individuals with prostate cancer, both PCA3 and tPSA are relatively weak predictors with low sensitivity and specificity." It is not clear what the latter part of this statement means. What reference method was used to conclude that PCA3 and PSA are relatively weak predictors of biopsy outcome? Both tests were judged to be sufficiently accurate to obtain FDA approval for their respective intended uses.	Methods the FDA uses to 'approve' a test are not relevant to this comparative effectiveness review. In some reviewed studies, the tPSA ROC was only slightly above the 'useless' test (sensitivity = 1-specificity). Although PCA3 was somewhat better, these two tests would not be ranked very high with regards to other screening tests for cancer or for other diseases. This is not to say that tPSA should not be used, only there is room for improvement.
TEP # 11	Results	See methodological comment: My personal preference would be to describe/list the methods for the various syntheses you did in this section. Instead you describe the methods in the respective parts of the results. (In the end, it works, apparently either way is OK.)	We find it more approachable to have the methods located close to the results. Had the results become repetitive with multiple explanations of the methods, we would have chosen to move a more complete description into the Methods section.
TEP # 11	pp. 63-65	This is an FIY, and you need not address it more materially than perhaps discussing it. Kester et al have proposed how to do a meta-analysis of whole curves. Essentially you would have to digitize the whole curve (and if there is partial verification bias, assume missingness at random [conditional on index test results]); fit a ROC line with 2 parameters per study; and then do a bivariate meta-analysis of the 2 parameters to get the overall curve. Stijnen et al have proposed a mixed-models approach to synthesizing whole ROC curves (Stat Med in the last 3-4 years, I believe). The current approach is a descriptive one, and it is fine.	The reviewer notes that this is a point for discussion and does not propose that this be done (see final sentence). We are aware of these more sophisticated approaches, but concluded that the low quality data available does not really warrant use of these methods in a preliminary analysis
PR # 12	Results	There is a great deal of detail, which seems appropriate in an evidence review, and the investigators seem to have included all appropriate diagnostic studies.	Comment acknowledged.
TEP # 13	Results	Very nice and logical presentation of results. Descriptions of gray literature searches are excellent. I do not know of any relevant studies that could have been overlooked.	Comment acknowledged.
TEP # 14	Results	Results presented in clear format, with appropriate detail.	Comment acknowledged.
TEP # 14	Results	Finally, the PCA3 "score" is essentially the ratio of PCA3 mRNA copies / mL divided by the PSA mRNA copies / mL. Although this relationship is mentioned in the text, and a trial using different housekeeping genes is mentioned, the potential effects of the inclusion of the PSA mRNA copy number in the denominator of the PCA3 score on the relative diagnostic sensitivities and specificities of the PCA3 score and tPSA was not explored or discussed. It would be helpful if this issue and its likely significance is explicitly mentioned, and if possible, addressed.	The data are limited, but have been added to the section of the Introduction on Development of a New Biomarker: PCA3.

Commentator and Affiliation	Section	Comment	Response
TEP # 14	Results	Moreover, the text refers to "FDA approval" of PCA3 at the bottom of page 3, and "FDA clearance" of PCA3 at the top of page 27. These represent different regulatory pathways with different implications for the rigor of FDA evaluation and the nature and extent of supporting data.	They certainly do – FDA clearance has been corrected in all cases to FDA approval.
KI # 1	Discussion/ Conclusions	The implications for major findings are clearly stated.	Comment acknowledged.
KI # 1	p. 93	Some support should be provided for the statement on page 63 (actually p 93): "We found data to support the conclusion that PCA3 had slightly higher performance compared to the extent of tPSA elevation. Based on limited data, the two markers seemed to be relatively independent." It does not seem intuitively correct to me that the two markers should be independent.	We did further research and identified a total of 10 studies[5-14] that reported the observed independence of PCA3 and tPSA. A subset of this data can be found in Table 15.
KI # 1	p. 93	On page 63 (actually p 93): "The issue of potential verification bias was raised early in discussion of the analytical framework and key questions, and discussed with members of the.." No question?	This comment was truncated and it was not clear if there was a question.
TEP # 2	Discussion/ Conclusions	The implication of the findings is clearly stated. The future research section is a novel addition to a report of this magnitude. I believe that it achieved the goal desired. The main report is sufficiently detailed and actually draws positive conclusions especially for KQ1 and KQ2. The report clearly identifies the weakness of the data to inform KQ3. The findings should inform policy.	Comment acknowledged.
TEP # 3	Discussion/ Conclusions	The implications of the major findings are clearly stated. The limitations are adequately described. No important literature is omitted to my knowledge. The future studies sections are clear and can be translated into new research. The report is well structured and organized. The main points are clearly presented. Conclusions are limited by paucity of studies and poor study quality, however this can be useful to inform policy and practice decisions pending future/better evidence.	Comment acknowledged.
TEP # 4	Discussion/ Conclusions	The Discussion provides a nice summary of findings, implications of the findings, limitations, and research gaps. The future research section is clear and can be easily translated into new research.	Comment acknowledged.
PR # 5	Discussion/ Conclusions	For the most part, the discussion and conclusions are appropriate.	Comment acknowledged.

Commentator and Affiliation	Section	Comment	Response
PR # 5	Discussion/ Conclusions	One important point that seems to be inadequately emphasized is that the use of PCA3 testing to improve the specificity of screening (i.e., to reduce the number of negative or potentially avoidable biopsies*) is dependent upon the provider's and patient's willingness to decline to do a biopsy in the face of a 'positive' tPSA and 'negative' PCA3. * [I object to the use of the term "unnecessary biopsies" to describe negative biopsies, since it implies that one could forego the biopsy, when, in fact, this is the only way we know how to detect important cancers. One would not refer to a negative mammogram or fecal occult blood test or serum glucose level or other negative screening or diagnostic test as "unnecessary" just because the result turned out to be negative.	It is true that even if subsequent PCA3 testing improves performance, impact on clinical practice and decision making will be limited if declining a biopsy in this scenario is not acceptable to patients and providers. This point is included in the Discussion under Applicability/ Interventions. We have modified the text to talk about decreasing the number of true negative biopsies among men with a positive tPSA screening test (i.e., reducing false positives), while maintaining or increasing the number of detected cancers.
PR # 5	Gap in knowledge	Thus, one important additional research aim or gap in knowledge would be to identify ways to facilitate the implementation of this knowledge into practice, for example by developing educational materials for patients and providers.	Added to gaps in knowledge.
PR # 6	Discussion/ Conclusions	Limitations are explained.	Comment acknowledged.
PR # 7	Discussion/ Conclusions	I think the conclusions are reasonable in that all the evidence on the utility of PCA3 is not in as yet. Most of the studies involve men already going to biopsy or for a prostatectomy. One might be able to determine which of this already selected group are likely to be negative for prostate cancer in these studies or likely to have less aggressive cancer in the context of those headed for a prostatectomy. But the more important questions in evaluating PCA3 are to evaluate if this marker can select (in a more prospective sense) who needs to go on for invasive testing. Patients already going to invasive testing come with higher tPSA values and will have a higher proportion of positive DRE results. So men that might have been selected for invasive testing solely on the basis of PCA3 may never occur in these studies.	None of the included studies had men enrolled due to results of PCA3 testing. This is why there is not a verification bias for PCA3. You comment is an interesting one, but outside of the aims of this review.
PR # 7	Discussion/ Conclusions	Because 13 of 15 studies selected in the context of answering questions 1 and 2 showed that PCA3 had a higher AUC, it may indicate that more research is needed.	Or it may mean that PCA3 actually performs better. However, we agree that higher quality studies that allow matched analysis would be more convincing.
PR # 7	Discussion/ Conclusions	I think more longitudinal followup data with the right measurements is really needed to assess question 3 and as noted by the authors of this review (page 72), the studies were of poor quality in answering the questions posed. None of the studies appear to really manage patients based on PCA3 as a stand-alone marker so the data analyses should be focusing on improvements in performance once PCA3 is added onto the usual batteries of tests.	Comment acknowledged and we agree.

Commentator and Affiliation	Section	Comment	Response
PR # 7	Discussion/ Conclusions	None of the papers addressed key elements of utility beyond diagnostic accuracy, namely in these studies patients were not really managed using PCA3 and one cannot tell clinical outcomes can be improved by selecting treatments and aggressiveness of treatment using the assay.	Agreed – comment acknowledged.
PR # 7	Discussion/ Conclusions	Availability of line data would be important in making new advances and studies paid for by companies seeking to show their assays have use should be encouraged to provide such data to others for scrutiny.	Comment acknowledged and we agree that access to data would be important.
PR # 7	Discussion/ Conclusions	In comparing AUC metrics for tPSA and PCA3, there is a comment that the AUCs cannot be different because the confidence intervals overlapped. In general, this is not true unless the confidence intervals are set to be 97.5% intervals rather than the usual 95%. For paired data, this assumption could be off by a lot since the confidence interval for the difference of the AUCs is in part driven by the correlation of the two markers. There are appropriate methods for dealing with paired data but the authors of this review did not have access to the line data required to perform that test.	You are correct. That is a 'quick and dirty' method that is likely to be acceptable due to the low correlations between PCA3 and tPSA. However, we have decided to limit attempts to provide strict statistical interpretations of p-values and in most studies (and summary analyses) this is not provided or cannot be reliably computed.
PR # 8	Discussion/ Conclusions	Discussion detailed and appropriate	Comment acknowledged.
PR # 8	Discussion/ Conclusions	This review asks and answers clear questions about PCA3, and does not postulate on best clinical use of PCA3. However, the readers of this review will obviously be considering the merits of this test and how it should be used.	Comment acknowledged and we concur.
KI # 9	p. 14, line 55	The authors make a comment about the indirect evidence for positive biopsy. This is not discussed in the Results.	This has been added to the strength of evidence discussion.
KI # 9	Applicability	The authors could discuss the evidence review in the context of the recent USPSTF D recommendation for PSA screening and the callout to identify better biomarkers for prostate cancer detection.	This has been addressed in the applicability section.
KI # 9	Applicability	The authors discuss uptake among physicians but not acceptability among patients. Given that the test requires prostate massage, and the possibility it requires a second DRE to perform the massage, this test may have acceptability concerns to patients.	This has been addressed in the applicability section.
KI # 9	Discussion/ Conclusions	Discussion of research needs is excellent and comprehensive.	Comment acknowledged.

Commentator and Affiliation	Section	Comment	Response
Groskopf/ Gen-Probe # 10	Discussion/ Conclusions	Gen-Probe greatly appreciates the fact that AHRQ selected PCA3 testing for comparative effectiveness review. We also note that the analysts concluded that PCA3 had greater accuracy for predicting biopsy outcome than PSA level. However, for the reasons described above, we feel that the current report does not provide a valid, accurate assessment of PCA3 Assay performance. Clinicians, patients, providers and policymakers look to AHRQ for guidance when new tests are introduced. Given the limitations of our current methods for early detection of prostate cancer, it would be unfortunate if premature release of AHRQ's report created a barrier to further implementation of PCA3 testing in clinical practice.	Comment acknowledged.
Groskopf/ Gen-Probe # 10	Discussion/ Conclusions	Recommendations: Data from the US pivotal clinical study (Appendix II, pages 21-32) and EDRN validation (Appendix III) should be included in the analysis. We recognize that this information was not available at the time of AHRQ's analysis, but the data are available now and are essential for a full and complete assessment of PCA3 effectiveness. AHRQ's procedures indicate that regulatory documents and abstracts/conference proceedings may be used. Per AHRQ's literature search strategy, only primary data are to be utilized. Haese (73) should be included and Ankerst (51) excluded from the analysis.	The FDA Summary document was reviewed, but most data did not meet inclusion criteria because it did not present data from a matched study. One element was used, as noted in Table 9 in the Results. The initial and updated searches did include identification of abstracts; abstracts were followed up to determine if full reports had been published and provided insight into types of data that might become available in the future (relevant to gaps in knowledge). A modified funnel plot, rather than review of abstracts, was used to investigate potential publication bias. The protocol notes that when published as a meeting abstract only, abstracts would be excluded from data extraction and analysis. Abstracts often represent early analyses of data that may contain errors and/or analyses may change/evolve prior to or following manuscript submission and peer review. An exception would not be possible, based on these methodological concerns, and that objectivity would require equal consideration of many other previously identified abstracts. This issue was discussed by the review team and it was decided that the abstract and poster: 1) did not meet the inclusion standards for this review; and 2) would require inclusion of many others that had previously been excluded, and that was not acceptable or feasible. Ankerst provided an analysis of two data sets. The European dataset was published in 2008 by Haese and Ankerst. The Haese dataset was very slightly larger, but all repeat status, central measures,

Commentator and Affiliation	Section	Comment	Response
			positive rate, and other characteristics made it clear the two articles were reporting primary data from the same study population. Unfortunately, Haese did not provide what we found to be the most useful data, the comparative ROC curves, so Ankerst's data presentation was selected. Had Haese been selected, the study would have had little impact in the analyses and a large study would have been lost.
		Studies that utilize assays other than PROGENSA® PCA3 should be excluded or reported separately, so that conclusions specific to PROGENSA® PCA3 Assay (the only FDA approved assay) can be evaluated separately by reviewers.	This analysis has been done, and the results added to page 46 of the final report. Stratification of AUC data by Progensa and other assays did not provide statistically significant evidence that assay methodology is an important consideration in these matched studies. Regression of AUC difference (PCA3 – tPSA) and percent initial biopsy showed a slope of -0.0472 (P=0.52). The same regression in studies using other assays also showed no relationship (slope = 0.0854; p=0.27). This translates to a 1.7% higher AUC for other studies in the same comparison.
		Because PSA and PCA3 are used differently for initial and repeat biopsy decisions, combining these two groups is not consistent with clinical practice…does not reflect the FDA-approved intended use:	The initial plan for this review was to conduct separate analyses for men having initial and repeat biopsies, and these separate analyses have been performed. We were, however, surprised by the lack of studies that exclusively studied men having initial and repeat biopsy, and the large number of studies that included both. The point of the 'combined analyses' was not to suggest that they are combined in practice, but rather to investigate whether biomarker prediction in these groups is different in matched studies. Evidence reviews are intended to identify, document and analyze all available data on a specified topic. Whether the data are consistent with current clinical practice is only relevant as a point of discussion and context. Evidence in some cases can be the impetus for further study that may change clinical practice. The rationale for the FDA intended use was not relevant to this comparative review of available evidence, unless the supporting data were available as one or more published

Commentator and Affiliation	Section	Comment	Response
		It appears that bias was the primary reason for determining that published studies are of poor quality. We acknowledge that most men in the studies had elevated PSA (i.e., selection bias), but most men with prior negative biopsies have elevated PSA, so these studies are consistent with the pivotal clinical study and PCA3 intended use. The justification for low strength of evidence (Directness, Strength of Association) seems inadequate as described above.	matched studies. More discussion of this point has been included. The justification for SOE for diagnostic accuracy/tPSA was based on the GRADE assumption of High Risk of bias in observational studies translating to a starting SOE of Low. Results were consistent, but Indirect. Precision was supported by the ability to observe the expected selection bias of grey zone studies and the difference in PCA3 and tPSA performance, but could not be directly measured (e.g., confidence intervals). Strength of association was weak. Together, these domains do not suggest downgrading to Insufficient, but, in our view, also do not warrant upgrading to Moderate.
Trikilinos # 11	p. 87, line 15 (pt 5)	Add that we need methods for the joint meta-analysis of multiple tests, as per my comment in the Exec Summary.	Added to Methodological Gap #6.
PR # 12	Discussion/ Conclusions	I am ambivalent regarding the future research section. The research suggested is only modestly interesting to me.	Comment acknowledged.
PR # 12	Discussion/ Conclusions	This test is clearly not a major improvement over PSA. I doubt anyone would fund a clinical trial using this test, and observational studies of testing strategies including this and other measures are likely to be biased.	Comment acknowledged.
PR # 12	Discussion/ Conclusions	The data presented suggest it is unlikely to be rewarding to determine if the test result can give reasonable predictive value for high-grade tumors, which is the only thing that will have a significant outcome on clinical utility	We agree that accurate prediction of aggressive or high-grade tumors is what will be most useful. Appropriately designed longitudinal studies are needed to determine if that is the case.
Klein # 14	Discussion/ Conclusions	Implications of findings, gaps and potential areas of future research clearly stated and appropriate.	Comment acknowledged.

Commentator and Affiliation	Section	Comment	Response
Klein # 14	Discussion/ Conclusions	Because the risk of overtreatment of indolent prostate cancer is very high, the impact of a postulated improved diagnostic sensitivity of PCA3 testing relative to PSA on net health benefits, even if true, is unclear, and may be in fact be negative. As described in the text, the review even appears to have come to different conclusions than FDA regarding the strength of evidence for increased specificity of the PCA3 score relative to PSA levels in patients who are candidates for rebiopsy, which appears to have been the basis for FDA approval. The preceding uncertainties associated with potential benefits and harms of PCA3 testing seemingly make it difficult to justify use of the PCA3 score in patient management outside the context of clinical studies undertaken under IRB supervision, in which patients are informed of the state of the art in prostate cancer management and offered the opportunity to consent to participation in a research protocol. In this regard, it would be helpful for the authors to describe and contrast in greater detail the methodologic differences that accounted for the FDA's apparently differing interpretation of the strength of evidence supporting use of PCA3 testing in patients who are candidates for rebiopsy relative to the conclusions of the review, if possible.	It is outside the scope of this comparative evidence review to describe and/or explain the methods and results of the FDA review of the Gen-Probe Progensa test. The supporting study has not yet been published, but the FDA Summary document did not describe a comparative matched study, so it is unlikely it would have met the inclusion criteria for this review.
Klein # 14	Discussion/ Conclusions	In light of the comments in preceding paragraphs (and in the paragraph below), as well as the possible stimulatory impact of FDA approval or clearance on physician acceptance, I suggest presenting in less absolute terms the finding that the PCA3 score may be more discriminatory for detecting cancer than tPSA elevation. For example, rather than stating "We observed that PCA3 *is* more discriminatory for detecting cancer" (emphasis added), I suggest using less definitive phraseology such as "We observed that PCA3 *may be* more discriminatory for detecting cancer." The objective in this wording change would be to continue to encourage further investigation, while discouraging routine clinical use based on extrapolation from a conclusion based on evidence deemed of low strength. Similarly, I suggest reporting the observation that the relative performance of the PCA3 score relative to tPSA *does not appear* to be dependent on biopsy history in this manner, rather than as a definite conclusion using words like "is not", as the conclusion is also supported by evidence the review deemed to be of low strength.	This is acceptable and we will make the modifications throughout the document.

Commentator and Affiliation	Section	Comment	Response
PR #15	Discussion/ Conclusions	If my husband' s family doctor had not done a PSA test on him in September of 2009, he would be dead and I would be a widow!!! Who are you people to decide such life changing decisions? Are you Urologists, Oncologist, Radiologist, Research scientist? Who are you and how dare you make a decision that could impact the lives of so many people-husbands,wife's, sons and daughters. These men that apparently you think so little of are precious to each of us! They are thr back bone of this country! This test does not cost thousands of dollars but it has and will save thousands of families from suffering from the lost of a loved one. I will be contacting my Congressman/ Senator to express my feelings about this and also will ask them WHO sits on this so called "preventive task force". You are obviously not interested in preventing men from being diagnosed with prostate cancer, is your agenda a s Socialist one, where you make the decisions and we have to accept them. I thought I lived in a Democracy, now I not so sure.	These comments appear to relate to the US Preventive Services Task Force Recommendation on PSA Screening, and do not represent a position of the authors or this review. As noted in this review, PSA screening remains controversial, and we simply reported relevant information from the literature.
TEP # 3	Figures/ Tables	Figures, tables and appendices require focus to understand. Figures, tables and appendices are adequate and quite descriptive.	Comments acknowledged.
PR # 7	Figures/ Tables	I found the earlier tables with one line per study very helpful and used that to determine which articles I should try and read.	Comment acknowledged.
PR # 7	Figure 7	Some of the figures could be improved. Figure 7 looks at the average AUC for PSA and PCA3 versus the difference. I might have also liked to see a plot of X=AUC(PSA) versus Y=AUC(PCA3).	Agreed. In the revised document, the current figure will be 8A and the suggested figure will be 8B.
TEP # 4	Table 10	It is not clear why several of the values in the “Initial Bx” column differ from the values in the “Initial Biopsy” column in Table 7.	The typographical errors in Table 7 have been corrected.
PR # 6	Table 12, p. 62	It would be helpful to present true positive (TP), false negative (FN), true negative (TN) and false positive (FP) counts.	Simple directions for deriving these numbers have been added.
PR # 6	Figure 10, p. 63	It would be helpful to present confidence region for summary sensitivity and specificity. Possible bivariate model considering sensitivity and specificity jointly could be considered.	We agree. Unfortunately, sufficient data were not available to perform this analysis.
PR # 6	Table F-1, p. 118	Not clear how there can be partial verification bias but at the same time “All patients received reference standard, regardless of index result”.	It is possible because these were not complete cohorts. None reported what happened to the men with negative tPSA results and, more importantly, not all men with elevated tPSA had biopsies.
PR # 8	Figure 11, p. 45	PCA3 is better than PSA at predicting prostate cancer. Is there any clinical relevance? This cannot be answered by this review, and that is discussed in the manuscript.	Comment acknowledged.
KI # 9	Table ES-1	Needs better labeling. It is not that clear what the top box represents relative to the bottom box without looking much more closely. Should clarify that they are examining sensitivity of PCA3 and PSA (top row) under constant FPR and FPR of PCA3 and PSA under constant sensitivity (bottom row).	This table has been simplified and additional information added to the table legend. The five analyses conducted for each comparator are listed in the bottom half of the table along with the number of studies providing information. It is not feasible to describe each analysis in the table.

Commentator and Affiliation	Section	Comment	Response
KI # 9	Table ES-2	Is the definition (e.g. precise vs imprecise) based on a comparison with PCA3? So are the authors saying that PCA3 is precise compared with tPSA? This should be clarified in the table legend.	Not exactly. Precision is the degree of certainty for an effect estimate related to a specific outcome. For example, the preciseness of an estimate of clinical sensitivity might be assessed using the tightness of a 95% confidence interval. So, precision in this case relates to the certainty of the estimates (e.g., differences between PCA3 and tPSA AUCs) computed based on performance of PCA3 compared to tPSA in predicting positive biopsy. The table legend has been clarified to reflect this.
KI # 9	Figure 2	Adverse events in this scenario are unlikely to be hemorrhage or infection, they're much more likely to be incontinence, voiding dysfunction, impotence, bowel dysfunction.	This has been modified.
KI # 9	Table 8 (AKA ES-1)	Again would benefit from a clarified legend.	Additional text has been added to clarify this table.
Groskopf/ Gen-Probe # 10	Table 1, p. 21	Reference to Schilling, 2010 should be reference 82, not 88. Reference 88 is another Schilling, 2010 reference, but it is a 5 subject case study not matching any of the data in the table.	Yes, corrected throughout the report.
Groskopf/ Gen-Probe # 10	Tables/ Figures	"N Biopsies Reported/N Total N Biopsies" for Auprich, 2011 is incorrectly reported as "621/805." These values come from a 2010 publication that is not cited in the report. The values from the Auprich reference used in the report are 160/305 or 305/305 depending on the analysis ("305" is correctly stated in Tables 18 and 19).	Corrected; no impact on analysis. Correct, data referred to here was RP data with N=305. Data from Auprich 2010 was not used in the analysis as noted in the text.
Groskopf/ Gen-Probe # 10	Tables/ Figures	% Positive Biopsies" for Cao, 2011 is incorrectly reported as "60." The value should be "64" (91/143).	Corrected.
Groskopf/ Gen-Probe # 10	Tables/ Figures	"N Biopsies Reported/N Total N Biopsies" for Roobol, 2010 is reported as "429/429," however, the study population was 721 subjects. Tables 7, 9, 10, and 17 correctly report 721.	Corrected; no impact on analysis.
Groskopf/ Gen-Probe # 10	Table 2, p. 22	The reference to Roobol, 2010 is shown as reference number 81 in Table 1, but as reference 96 in Table 2 and other tables. Reference 96 is a duplication of reference 81 in the "References"section.	Ref 96 is a duplication; corrected in references, text and Table 2.
Groskopf/ Gen-Probe # 10	Tables/ Figures	PSA Cutoff" for Ankerst, 2008 is reported as "≥2.5." The publication cites that it was an inclusion criterion of Haese, but that information does not appear to be in the Haese publication.	We used the Ankerst data from the European cohort (p1304 & 1307). The Table indicated that 97% of subjects had PSA ≥ 2.5; this seemed a reasonable characterization of the data.
Groskopf/ Gen-Probe # 10	Tables/ Figures	"PSA Cutoff" for Nyberg, 2010 is reported as "2.5." No PSA cutoff is specified in the publication and in the Discussion section the authors state, "…the present study had no inclusion criteria for tPSA…"	Table 1 shows that only 2 of 62 had PSA < 2.5 (3%); 60% 2.5-10; 37% > 10. This seemed a reasonable characterization of PSA values in the population.
Groskopf/ Gen-Probe # 10	Tables/ Figures	"Previous Negative Biopsy (%)" for the Schilling, 2010 reference is reported as "14." The publication states that 14 of the 32 patients who received a biopsy had a previous negative biopsy, or 44%.	Corrected

Commentator and Affiliation	Section	Comment	Response
Groskopf/ Gen-Probe # 10	Tables/ Figures	"PSA Cutoff" for Ochiai, 2011 should be listed as 2.5 - 50 (the study excluded men with PSA >50 ng/mL from analysis).	Corrected
Groskopf/ Gen-Probe # 10	Table 5, p. 27	Rigau, 2010: Sample size ("N") is incorrectly reported as 21. The correct value is 215.	Typo corrected; no effect on analysis.
Groskopf/ Gen-Probe # 10	Tables/ Figures	"Initial Biopsy" for the Rigau reference is incorrectly reported as "100%." The correct value is 74%. Supplemental Table 1 of Rigau, 2010 shows that not all men were having an initial biopsy: 55 of 215 subjects (26%) were having a repeat biopsy. This directly affects some of the analyses in the AHRQ report (Figures 5 and 6).	100% entry error was corrected in Table 5 and Table 7 to 74% from Supplemental Table. Potential impact on analysis as noted. These data were reanalyzed with the new studies identified.
Groskopf/ Gen-Probe # 10	Table 7, p. 29	"Initial Biopsy" for the Schilling, 2010 reference is incorrectly reported as "86%." The correct value is 56% because 18 of the 32 subjects had an initial biopsy.	Typo corrected; no effect on analysis
Groskopf/ Gen-Probe # 10	Tables/ Figures	"N" for the Adam reference is incorrectly reported as "106." The correct number of reported biopsies is "105" as shown in Table 1.	Typo corrected and checked throughout.
Groskopf/ Gen-Probe # 10	Tables/ Figures	"N" for the Wang reference is incorrectly reported as "516." The correct number of reported biopsies is "187" as shown in Table 1.	Corrected in Table 7; all other references correct.
Groskopf/ Gen-Probe # 10	Figure 5, p. 30	The Schilling reference is correctly reported as having 56% initial biopsy in Table 10, however the data point used in Figure 5 is the incorrect value of 86% (as reported in Table 7).	Noted; these data were reanalyzed with the new identified studies.
Groskopf/ Gen-Probe # 10	Tables/ Figures	The Rigau data point is included in Figure 5 as 100% initial biopsy should be 74%.	Corrected. These data were reanalyzed with new identified studies.
Groskopf/ Gen-Probe # 10	Tables/ Figures	The Figure 5 analysis needs to be repeated with correct data**.**	Agreed; analyses were repeated with the new studies identified.
Groskopf/ Gen-Probe # 10	Figure 6, p. 31	As in Figure 5, both the Schilling and Rigau data points use the incorrect proportion of men with initial biopsy.	Corrected as above.
Groskopf/ Gen-Probe # 10	Tables/ Figures	Table 13 reports Schilling as having a sensitivity difference of 39%, however there is no point in the graph with a 39% difference. Table 9 (page 35)	The Table 11 point indicating a sensitivity difference of 39% for Schilling at 1-Specificity of 50% was missing. The figure has been corrected and additional studies added.
Groskopf/ Gen-Probe # 10	Tables/ Figures	The Figure 6 analysis needs to be repeated with correct data.	Agreed; analysis was repeated with new identified studies.
Groskopf/ Gen-Probe # 10	Table 10, p. 37	Three "Initial Bx" numbers are reported inaccurately in Table 10, Nyberg, Roobol and Rigau.	All were corrected in Table 10; there was no impact on analysis as numbers were correct in data tables.

Commentator and Affiliation	Section	Comment	Response
Groskopf/ Gen-Probe # 10	Table 11, p. 40	Median PCA3 Scores for the Ankerst reference are reported as "34.3" and "34.2." An erratum for the Ankerst paper was published (The Journal of Urology, Volume 181, Issue 3, March 2009, Page 1507) and the median PCA3 Score for the Bx Neg group was corrected to 19.4. This issue directly affects the results and conclusions (Figure 9, text pages 40-41).	We appreciate this comment as the erratum was not identified in the original search. The correction was made and these data were reanalyzed with the newly identified studies.
Groskopf/ Gen-Probe # 10	Table 12, p. 42	In the text on page 42 the AHRQ report states, "In two studies, the specificity was incorrectly reported as 1-specificity, or vice versa." The report appears to refer to Nyberg (76) and Adam (65). There are some issues with the assessment of these publications: Re Nyberg (Ref 76): Table 12 of the AHRQ report lists PCA3 "1-Spec" as "45.4" and "Sens" as "66.7." The Nyberg publication states that at a PCA3 cut-off of 35 "the sensitivity was 66.7% and the specificity was 44.4%." It appears that the AHRQ analysts believe that specificity is incorrectly reported in the publication, and that the 1-specificity value was used instead. Gen-Probe agrees that the PCA3 specificity cannot be 44.4%. Nyberg reported a median PCA3 Score of 22 for patients with a negative biopsy result. A PCA3 Score cutoff of 22 therefore corresponds to 50% specificity (in the negative biopsy group, 50% of values are below the median). Raising the cutoff from 22 to 35 would cause specificity to increase; therefore, 44.4% cannot be the specificity (it must be greater than 50%). Gen-Probe cannot confirm that 44.4% is equal to 1-specificity. If the ROC curve is correct, a sensitivity of 66.7% corresponds to a 1-specificity of approximately 35% (specificity of 65%). These inconsistencies cannot be completely resolved without contacting the authors of the publication or obtaining access to the source data.	Reading the 1-specificity off the ROC curve confirmed these assumptions (per statistician). Nyberg reported a specificity of 44% (rounding here), 1-spec of 56 and sens of 67. As you note, this cannot be correct, based on the ROC curve. We agree that the reported sensitivity and specificity at a PCA3 cutoff of 35 was not sufficiently reliable to be used and removed Nyberg from one analysis (comparisons of sensitivity and specificity at a PCA3 cutoff of 35; Table 12). However, data were sufficient to perform the other analyses comparing PCA3 and tPSA: AUC (Table 10, Figure 8); Central estimates (median/mean for negative and positive biopsy; Table 11); deriving distributions (Figure 9); and sensitivities at a range of specificities over the ROC curve (Table 13). Since the missing data was not critical, we elected not to contact the authors.
Groskopf/ Gen-Probe # 10	Table 13, p. 44	tPSA "Sens" is reported as "74.0." It is unclear how this value was reached. Based on the ROC published in Adam, neither the incorrect 1-specificity of 35.1 nor the correct 1-specificity of 50.0 would result in a tPSA sensitivity of 74.0.	In Table 12 (p. 44), you are correct that the 1-specificity that should have been used for Adam was 50% and the sensitivities for PCA3 and tPSA should have been 78% (as in Table 13) and 87%, respectively. This has been corrected.
Groskopf/ Gen-Probe # 10	Tables/ Figures	"Number" for the Adam reference is incorrectly reported as "27." The correct number of reported biopsies is "105" as shown in Table 1; 27 corresponds to the number of white men in the study. PCA3 "1-Spec" is reported as "35.1" and "Sens" as "77.1." These data correspond to the specificity and sensitivity of the black men in the study, not the total number of men.	This reference has been corrected. This has been corrected. These numbers from Table 3 were entered in error. The correct numbers for the overall population from Table 2 were 77.7% sensitivity and 50% specificity at a PCA3 cutoff of 35. This correction was made.
TEP # 13	Tables/ Figures	Figures and tables are helpful, as are appendices.	Comment acknowledged.

Commentator and Affiliation	Section	Comment	Response
TEP # 13	Tables/ Figures	Following the discussion and references is a table of abbreviations which contains "USPSTF" (p. 97), however, this abbreviation is not used in the report.	USPSTF has been removed from the table of abbreviations.
TEP # 14	Tables/ Figures	Tables and figures useful in supplementing [result] text.	Comment acknowledged.
TEP # 14	Tables/ Figures	Results presented in clear format, with appropriate detail. Tables and figures useful in supplementing text.	Comment acknowledged
PR # 7	References	Citing background references for the benefit of the reader would be useful in addition to the AHRQ document on evaluating medical tests.	Reviewers are requested to follow the methodological guidelines provided in the AHRQ *Methods Guide for Effectiveness and Comparative Effectiveness Reviews* and the AHRQ *Methods Guide for Medical Test Reviews*. This latter reference is available online for content review and the link is provided. Other methods used and referenced included Whiting et al (QUADAS) and Owens et al (GRADE).
PR # 15	References	My husbands PSA score, two biopsies were on the second one stage 4, high grade aggressive cancer was found. Those are the only references that are needed. I'm not a physician but apparently this committe is no either. Can you tell that I am angry?	Comment acknowledged.
TEP # 4	Appendix J, Fig J1	The numbers do not add up for the men with tPSA >= 7. Fix the numbers.	This typographical error has been corrected; 94 has been corrected to 84..
TEP # 4	p. J2, lines 3-31	The text will need to be corrected after the numbers in Figure J1 are fixed.	The text has been modified to fit the figure.
TEP # 4	p. J3, line 3	details should be provided for the “in-house modeling”	Agreed. This has been added to Appendix J.
TEP # 4	p. J3, line 37	a period is needed at the end of the sentence	Period was added.
TEP # 4	p. J4, line 41	replace “show” with “shown”	Done.
TEP # 4	p. J7, line 10	replace “actually” with “actual”	The text was modified.
TEP # 4	p. J9, line 35	replace “Base” with “Based”	Done.
TEP # 4	p. J9, line 48	add closing parenthesis to the end of the sentence	Done.
TEP # 4	p. J9, line 50	NPV should be 96.9%	This was modified and correct value is now 96.6%.
TEP # 4	p. J9, lines 54-58	omit this repeated information	This text duplication was corrected.
TEP # 4	p. J10, Figure J6	the calculation on the right-hand side is NPV not PPV. The fraction should be “(93/(3+93))”.	This has been corrected
TEP # 4	p. J10, line 41	replace “difference” with “different”	Corrected.
TEP # 4	p. J10, line 45	replace “loser” with “lower”	Corrected

Commentator and Affiliation	Section	Comment	Response
PR # 6	Appendix J, p. 114	Appendix J includes discussion of partial verification bias. However it is still not clear why. “Of most importance is the finding that the ROC curves are not affected by partial verification bias for tPSA measurements.” Is this a fact or conjecture?	This finding has been reported in another article (Thompson et al. JAMA 2005;294(1):66-70; PMID 15998892). Our own in-house modeling confirmed this finding in our setting. Whether this is a universal finding in all settings was not addressed.
KI # 1	General	The report is clinically meaningful. The report is well structured and organized. The main points are presented and conclusions can be used to inform policies.	Comment acknowledged.
TEP # 2	General	The report is very well organized and very clear. I am surprised that the writers were able to distill the information so crisply and easily understood language. The insight that PCA3 works equally well in biopsy naive and post biopsy patients is a real positive. This report appeals to many audiences.	Comment acknowledged.
TEP # 3	General	The key questions are appropriate and explicitly stated. The target population and audience are explicitly defined. The report is clinically meaningful and useful.	Comment acknowledged.
TEP # 4	General	The report is well structured and organized. The main points are clearly presented and the conclusions are clear. The report is clinically meaningful.	Comment acknowledged.
PR # 5	General	The report seems to be clinically meaningful. The key questions are appropriate and explicitly stated. It [the report] is clear.	Comment acknowledged.
PR # 6	General	Clarity is reasonable but can be improved. Unfortunately the report provides impression of generally modest or insufficient evidence. Hence, results will not likely inform policy with exception of the need for future research	Comment acknowledged.
PR # 7	General	I believe the three questions posed by the authors are relevant and clinically meaningful. The audience is clearly defined, however readers that are not familiar with statistical methods used in the evaluation of diagnostic tests and issues of verification and spectrum bias may not get the concerns over weaknesses in the studies used in these reports.	Comment acknowledged.
PR # 7	General	I think the main points are clearly stated. So many diagnostic accuracy studies in oncology suffer from the biases in these articles. Verification bias is a form of non-random missing data in which patients that e.g. have negative scores for a marker never go on to have clinical truth assessed. Spectrum bias is when patients have already been referred for testing but the utility of the test is to decide if patients should be referred. It would be useful to have standards for the reporting of such studies along the lines of the REMARK guidances and STARD guidelines and actually have editors enforce this when accepting articles.	Agreed – comment acknowledged.
PR # 7	General	Long term benefits of the using the PCA3 assay for men suspected of prostate cancer or those already diagnosed have not been determined.	Agreed – comment acknowledged.

Commentator and Affiliation	Section	Comment	Response
PR # 8	General	A thoughtful review of PCA3 and comprehensive analysis of PCA3 vs. PSA data. The questions the review was trying to answer are clearly stated, and the data are clearly explained regarding what conclusions are possible. The finding that biopsy history is not relevant and that questions 1 and 2 could be combined was interesting, and the explanation elegant. The methodology was exhaustingly described. The limitations of the study and directions for future investigation are described in great detail. I find few faults with this review.	Comment acknowledged.
PR # 7	General	I think the report is very clear. It asks a question and answers it. The underlying questions - how to use PCA3, should we screen for prostate cancer - cannot be answered by this report, and the report does not make any recommendations regarding these matters. There will continue to be debate on these issues, through no fault of this report. Please see my general comments above.	Comment acknowledged.
KI # 9	General	The report is certainly clinically meaningful, especially in light of the recent USPSTF D recommendation for PSA-based prostate cancer screening. In that report, there was an explicit call for better biomarkers for prostate cancer screening and PCA3 represents the latest biomarker to receive widespread attention as an augmenting marker in prostate cancer.	Comment acknowledged.
KI # 9	General	The report is very well structure and organized. The data are presented in an understandable order and logically arranged. The conclusions may be used to guide future research where the robust gaps identified in this report lie.	Comment acknowledged.
Groskopf/ Gen-Probe # 10	General	Our comments are limited to the combined KQ1 and KQ2 analysis (PCA3 for initial and repeat biopsy), used by the AHRQ analysts to assess PCA3 predictive accuracy relative to serum PSA. All references to tables, figures, page numbers and literature citations correspond to the main body of the AHRQ draft report.	Comment acknowledged and considered in reviewing comments.
TEP # 11	General	I read the report "PCA3 testing for ... " with interest. Overall this is a well written report: - The main text that is (appropriately) on the shorter side - The methods used are appropriate - The key questions have been addressed appropriately - The authors did perform quantitative analyses, and their analyses were informative. I feel obliged to make a few comments, but I will not insist on them. I leave it to the discretion of the authors first, and the TOO secondarily (the comments are not essential).	Comment acknowledged.
PR # 12	General	This is a very important report that carefully considers key questions and addresses them appropriately. This [report] is very clear.	Comment acknowledged.
TEP # 13	General	This is an extremely well-written report. The key questions are appropriate and the report is clinically meaningful.	Comment acknowledged.

Commentator and Affiliation	Section	Comment	Response
TEP # 13	General	The report is very well structured and organized, particularly given the complexity of the subject. I believe that the results will be usable and helpful towards informing policy and practice recommendations	Comment acknowledged.
TEP # 14	General	Comprehensive analysis of matched studies relevant to comparative analysis of PCA3 testing. Report clear and well-organized. The authors did a commendable job in synthesizing and presenting a heterogeneous, body of data that was less than adequate to answer key clinical questions.	Comment acknowledged.
TEP # 14	General	This thorough, comprehensive, and well-reasoned review highlights the absence of high quality data supporting the usefulness of PCA3 testing both as a diagnostic marker in patients at risk for prostate cancer who are candidates for initial or repeat biopsy, or as a prognostic marker in patients in whom prostate cancer is identified. Given the controversy over the clinical utility of curative therapy vs. watchful waiting in patients with localized prostate cancer and the recent USPSTF recommendation against PSA screening, the introduction of PCA3 testing into clinical practice without data supporting its clinical utility is troubling.	Comment acknowledged.