Comparative Effectiveness Review Disposition of Comments Report

Research Review Title: *Diagnosis and Treatment of Clinical Alzheimer's-Type Dementia: A Systematic Review*

Draft review available for public comment from April 29, 2019, to May 27, 2019.

Research Review Citation: Fink HA, Hemmy LS, Linskens EJ, Silverman PC, MacDonald R, McCarten JR, Talley KMC, Desai PJ, Forte ML, Miller MA, Brasure M, Nelson VA, Taylor BC, Ng W, Ouellette JM, Greer NL, Sheets KM, Wilt TJ, Butler M. Diagnosis and Treatment of Clinical Alzheimer's-Type Dementia: A Systematic Review. Comparative Effectiveness Review No. 223. (Prepared by the Minnesota Evidence-based Practice Center under Contract No. 290-2015-00008-I.) AHRQ Publication No. 20-EHC003. Rockville, MD: Agency for Healthcare Research and Quality; April 2020. Posted final reports are located on the Effective Health Care Program search page. DOI: https://doi.org/10.23970/AHRQEPCCER223.

Comments to Research Review

The Effective Health Care (EHC) Program encourages the public to participate in the development of its research projects. Each research review is posted to the EHC Program Web site or AHRQ Web site in draft form for public comment for a 3-4-week period. Comments can be submitted via the Web site, mail or E-mail. At the conclusion of the public comment period, authors use the commentators' submissions and comments to revise the draft research review.

Comments on draft reviews and the authors' responses to the comments are posted for public viewing on the 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.

Section	Reviewer & Affiliation	Comment	Response
Executive Summary	Public Reviewer #6 (AAFP)	Tables 0.1, 0.2, 0.3 1. Add LR+ / LR- 2. How were TP/TN/FP/FN calculated? Using the same prevalence estimate across tests, or using the prevalence from the studies of each individual test (which would make it much less useful and harder to compare tests). Table 0.4 3. This is not as helpful as it could be. Please provide quantitative measures of benefit, for example absolute risk reduction, NNT, etc. Just saying 'Favors" is too vague. 4. Also need to be clear what the minimal clinically important difference is.	1. With sensitivity and specificity, and TP, TN, FP and FN already reported, and diagnostic accuracy results tables already cluttered and difficult to fit on a page, we decided that any additional information that would be obtained from the extra work calculating and reporting LR+ and LR- was not warranted. 2. TP, TN, FP and FN were calculated based on the CATD or AD prevalence in the studies that examined diagnostic accuracy and not based on a standardized prevalence in a "typical" primary care population. We agree this makes it harder to directly compare TP, TN, FP and FN between studies. However, there are several issues with recalculating these measures based on a standardized CATD or AD prevalence. One is that recalculating these measures based on a standardized CATD or AD prevalence would require multiple assumptions and substantial work. A second is that the non-CATD and non-AD comparison groups differed in composition between studies and this likely affects sensitivity and specificity. So, it is likely that sensitivity and specificity of the diagnostic tests would not be the same with a different CATD or AD prevalence. Third, our clinical question was not about the diagnostic accuracy of these tests in a typical primary care population, but about the accuracy of brief cognitive tests to distinguish between CATD and either MCI or normal cognition in the subset of older adults with suspected CATD, and the accuracy of biomarkers to distinguish between neuropathologically confirmed AD and non-AD dementias in older adults with CATD. So, it is unclear what the most appropriate baseline CATD and AD prevalences should be for those clinical questions. 3. Throughout the revised report, we added ARDs and their 95% CIs, and NNTB and NNTH and their 95% CIs. We also added language to the Methods section explaining the criteria for when we did or did not include language suggesting a direction to results. 4. In the Methods section, we included language indicating that for mean between-group differences in change in continuous outcomes reported as SMD, we considered >0.2 to be a small effect, >0.5 to be a moderate effect, and >0.8 to be a large effect. Beyond that, we did not attempt to translate what difference in mean change between groups equals a minimal clinically important difference within individuals. We discussed this issue in the discussion of the revised report. For categorical or responder outcomes, several treatment trials reported likelihood of outcomes like any improvement on a scale (e.g., MMSE, ADAS-Cog), improvement by at least 4 points on a scale (e.g., ADAS-Cog), and improvement by a certain percentage from baseline on a behavior scale (ranging from 20% to 100% improvement). For these, we did not identify what threshold change met or exceeded a minimally clinically important difference.

Section	Reviewer & Affiliation	Comment	Response
Introduction	Peer Reviewer #2	In setting up the introduction and results for the imaging studies and CSF markers, it might be valuable to use the AT (N) framework. https://www.sciencedirect.com /science/article/pii/S1552526018300724?via%3Dihub	We revised the introduction to address the ATN framework. The biomarker literature review was limited by the methods reported in the published literature. We expect future studies will be more likely to use the recently published ATN framework.
	Peer Reviewer #3 (TEP)	I find the use of some clinical test iike the MMSE as diagnostic and as outcome measures is one of the many complexities in dementia research. While many commonly used tests like the MMSE and MOCA have been validated as diagnostic instruments, I dont believe there is validation for changes in score as trajectory (outcome) measures. While the tests are widely used in the sphere of change, I don't see a comment in the introduction or methods about that problem.	Most treatment trials reported change in cognitive and/or functional measures as outcomes. Because group differences in mean change are difficult to interpret for individuals, we emphasized categorical outcomes (e.g., incidence of change beyond a threshold), but also reported mean changes and, when neither of these were available, scores at followup.
	Peer Reviewer #4 (TEP)	The introduction was comprehensive in terms of the questions being asked, and in adding biomarkers to cognition it does present a very complicated picture and thus difficult to follow despite the comprehensive introduction. Lastly, there is mention of the NINDS-ADRDA as one of the diagnostic criteria used. But in the Analytical framework, biomarkers as measures after positiive cog tests are not called for in NINCDS-ADRDA. They are called for in the NIA-AA criteria which is barely addressed. I think this is a key missing component especially if this review will have a shelf life and remain relevant for the next few years.	Differing reference standards were applied for the cognitive tests and biomarker diagnostic questions. For the cognitive tests question, a clinical diagnosis reference was required, but NINCDS-ADRDA, DSM-IV, NIA-AA or other clinical criteria were all eligible. In the biomarkers question, neuropathologic evidence of an AD process was required. The NINCDS-ADRDA criteria were the clinical diagnostic criteria most frequently mentioned in the report because they were the most commonly reported in eligible cognitive testing studies. We revised the introduction to better address the NIA-AA criteria, including where they incorporate biomarker information.

Section	Reviewer & Affiliation	Comment	Response
Introduction (cont'd)	Peer Reviewer #5 (TEP)	Introduction 1. Page 30, lines 25-32 – The following information seems important as a rationale for the focus of this review. Consider including these sentences in earlier section(s) of the report, e.g., Structured Abstract and Evidence Summary: Because a separate ongoing AHRQ review is focused on the efficacy and harms of nondrug treatments for patients with CATD, these interventions are not addressed by the present review except when included as a control group for a drug intervention. Therefore, the scope of the present review is limited to cognitive and biomarker diagnostic testing for CATD and AD, and prescription drug and supplement treatment of patients with CATD. 2. Page 31, line 5 – Depression is not in the list of patient characteristics. Was this intentional or an oversight? It is included in the list of patient characteristics for the other Key Questions. 3. Page 31, line 5, line 14 – Do you mean to have a slash mark (/) between "cognitive or functional level" and "CATD stage" as you do earlier in the report, e.g., page 30, line 41 and in other Key Questions? Should be consistent in how you label that patient characteristic, with or without the "/." I think it was just a typo here. 4. Page 31, line 32 – What are the "nondrug interventions" mentioned in the Key Questions and then in the PICOTS tables that follow? Should you define/list somewhere in the report the ones you examined or found in the studies you reviewed? I may have missed it.	1. The abstract stated that summarizing evidence on benefits and harms of prescription drugs and supplements in patients with CATD was an objective of the review. We were required by AHRQ formatting guidelines to not include the Methods in the Evidence Summary other than to refer the reader to the main report. The scope statement referenced by the reviewer already is included in the Introduction of the main report. 2. It was our error that depression was not included in the list of possible effect modifiers in KQ 1a. It has been added to the list. 3. We intended to have the slash and added it where missing in KQ 1a and KQ2. 4. Nondrug interventions would include social/behavioral/environmental interventions such as cognitive training, diet, exercise, etc. We revised the report to list some examples of nondrug interventions. These were outside the scope of this review and are the topic of a separate ongoing AHRQ review. 5. Based on GRADE guidance and discussion with AHRQ, we added TP, FP, TN and FN to sensitivity and specificity in the report text. We still calculated PPV and NPV as specified in the protocol, but only reported those results in the appendix tables.

Section	Reviewer & Affiliation	Comment	Response
Introduction (cont'd)	Peer Reviewer #5 (TEP) (cont'd)	5. Page 33, Table 1.1, KQ1-2 – Outcomes column, top line refers to PPV and NPV, but PPV and NPV were not addressed in the Results section on cognitive tests. Should remove PPV and NPV here?	

Section	Reviewer & Affiliation	Comment	Response
Introduction (cont'd)	Peer Reviewer #6 (TEP)	1. P 17, Lines 50ff. The introduction aims for a broad user profile for this report, beginning with clinical primary care; for the review of brief cognitive screens, one expects to see a detailed evidence review in the following sections that is tailored for that user group. Instead, what we find is a very limited review of evidence about cognitive screens, virtually none of it derived from primary care samples (evidenced by the high prevalence of dementia in the samples studied). Page 20. Table 0.1. The lowest prevalence of CATD in the cited studies was 16%, at the upper end of the prevalence range expected in primary care of older people. Most included studies had CATD prevalences of well above that. The principles of test performance analysis emphasize that the prevalence of a target condition in a sample has powerful impact on the performance of a test when the condition of interest cannot be measured with high precision. There are implications of the choice made here to include studies of high-prevalence samples: Test performance will be inflated (tests look 'better' than they will perform in a more representative sample). Since test accuracy of .8 was selected as the desideratum for a 'good test' and is difficult to achieve in population-representative samples, the interpretation of this section should be much more modest. 2. Several tests included here and characterized as 'brief' are not so brief when the focus is on primary care. There is a substantial literature on this topic which has not been considered in this review.	1. The purpose of the review on the accuracy of brief cognitive testing was to evaluate the accuracy when used in individuals with suspected cognitive impairment and not to review accuracy of cognitive screening in a healthy older population. The prevalence of CATD in a population in whom it is suspected is higher than it is in the primary care population as a whole. We revised the report to better discuss the implications of high CATD prevalence study sample on results and the applicability of these results to clinical practice: "In both cognitive and biomarker test accuracy studies, the high prevalence of CATD and AD, respectively, could have increased diagnostic vigilance and led to sensitivity results higher than what would be expected in typical clinical populations,(Fowkes FGR. Lancet 1986;1:493-4) even those in whom CATD is suspected." 2. During development of the protocol for this review, we were asked to include cognitive tests as long as 30 minutes. Many included tests thus may not be feasible for the primary care provider to administer herself/himself, but potentially could be administered by a psychologist embedded within the primary care clinic. 3. We agree with this reviewer that performance of cognitive screening tests for identifying dementia in typical, heterogeneous clinical populations is unlikely to be as good as has been reported in study populations that have more homogeneous comparison groups. We addressed this issue in the Applicability section of the revised report. 4. The review looked for studies on the accuracy of Mini-Cog and other brief cognitive screening tests for distinguishing CATD from normal cognition or MCI in patients in whom there was suspicion for possible CATD. Studies of these tests were identified in our literature search, but they did not meet inclusion criteria for several reasons, including: the tests were not administered in English, sample size was <25, studies were designed to distinguish nonspecific dementia rather than CATD from MCI or normal cognition, studies looked at screening in primary care or community settings regardless of whether there was suspicion for cognitive impairment or even excluded patients with dementia, and/or studies did not evaluate all participants who completed brief cognitive testing with a full diagnostic evaluation. We revised the report Limitations section to explain why these studies were not included and discuss how their results pertain to the rest of the report findings.

Section	Reviewer & Affiliation	Comment	Response
Introduction (cont'd)	Peer Reviewer #6 (TEP) (cont'd)	3. No brief cognitive screening test has been shown to detect AD as defined by biomarkers or neuropathology. When patients with AD make up a large fraction of a study sample, cognitive impairment associated with (and believed to be pathogenically related to cardiovascular, respiratory, or renal disease, and rarer forms of dementia (e.g. FTLD) are excluded, tests will look as though they 'detect AD extremely well.' The situation in medical practice is not as clean as presented here. 4. The review excludes some useful and widely adopted screening tests that were developed specifically to detect dementia in primary care. The Mini-Cog is one such test. In contrast to most of the screening tests included in the review, there is support for its good performance in a population sample with a 'typical' (i.e. non-inflated) prevalence of CATD and evidence that it is as effective as the MMSE and a full neuropsychological battery in identifying individuals with CATD. It's not clear why this screening test (considered acceptable in a rigorous NIA review) was not considered here.	
	Peer Reviewer #7	Good	Thank you for this comment.

Section	Reviewer & Affiliation	Comment	Response
Introduction (cont'd)	Peer Reviewer #8 (TEP)	1. General: I feel it would be useful in the Introduction to specifically identify the target population/audience. This is mentioned in the executive summary but would be best placed in the Introduction. As a reader, this is where I would quickly look to find this information. 2. Consider adding to the Main Points that evidence regarding effect modifiers was lacking for most of the key questions. As looking at this question was a major goal of this review, I feel that the lack of data is a Main Point and should be highlighted.	1. We revised the introduction to better identify the target population/audience. 2. We revised the Main Points to note that evidence about effect modifiers was very limited.
	Peer Reviewer #9	Introduction well written and justifies the need for this report and provided good coverage of the topic at hand. It also builds on the previous work by AHRQ on this topic.	Thank you for this comment.

Section	Reviewer & Affiliation	Comment	Response
Introduction (cont'd)	Public Reviewer #1 (Anonymous)	1. Page 1 - - report needs to emphasize that populations are HIGHLY selected - - this is more evident later in the report when you look at ages of those populations in diagnostic marker studies. MOST IMPORANTLY - the autopsy studies typically do not allow classification of false negatives since very few persons get autopsy unless they have dementia - -you need cohort studies to achieve this and cohort studies have shown convincingly that many persons with Neupathologically proved AD that are more community based have substantial numbers of persons with AD changes (see Sonnen et al Arch Neurol 2011 and Montine et al CURR Alzheimer's RESEARCH 2012. In many ways the PICOTS and search strategy guarantees that you won't find this but it is important. The introduction mentions heterogeneity but doesn't do justice to the fact that within Alzheimer's disease there is likely much heterogeneity - so much that you can almost guarantee attempts to define a disease described in a 50 yo persons with Alzheimer's is almost certainly not the same condition we see so commonly in older persons today even though there are striking similarities to the neuropathology seen in that one case.	1. The revised applicability and limitations sections addressed several ways in which study populations were selected so that the applicability of results to typical clinical populations may be limited, including: enrollment of individual with limited life expectancy, enrollment of mostly white patients, enrollment of patients who had been well characterized over years of followup, enrollment of participants from specialty research centers. 2. Run-in phases (pre-randomization screening) were used in early AD clinical trials, particularly after the first tacrine study was published (JAMA,1992), which documented an elaborate "enrichment" design that was clearly described in its Methods section. We did not document pre-randomization screening of potential trial participants with either placebo (to gauge likely adherence and placebo response) or active treatment (to also gauge likelihood of treatment benefit or harms). However, to address this question, we re-examined the prior donepezil trial systematic review and the two oldest donepezil versus placebo trials (Rogers, 1998; Burns 1999). None described a pre-randomization run-in phase. Because these trials predated required registration in clinicaltrials.gov, there was no additional information on possible pre-trial screening available there. So, while we did not identify an example of pre-randomization screening among analyzed trials, if a trial used such a design, and excluded individuals from randomization who during screening were less adherent, experienced more adverse effects or had smaller therapeutic responses, it would be likely to bias results to make the treatment look more effective and more safe, and the results would be less generalizable to nontrial populations. 3. We did not change population column from reporting quantitative data on the population characteristics to interpreting the population representativeness. We commented on the representativeness or lack of representativeness of the study populations in the applicability and limitations sections of the report narrative, including that the age distribution in many studies appeared younger than the CATD population in typical clinical settings

Section	Reviewer & Affiliation	Comment	Response
Introduction (cont'd)	Public Reviewer #1 (Anonymous) (cont'd)	2. You don't mention that drug trials, at least early in development of drug treatments for AD, had a unique design. Persons were tested first on the drug - if they didn't have a side effect and didn't get worse or might have gotten better - - then they were eligible to be randomized and studied. This was OK with the FDA. However, I checked a couple of papers in your reference list and it is very hard to determine this is the case from published papers. I do know of one instance where we were asked to provide potential subjects and after an average of about 80% of the subjects had side effects, we chose not to continue participating but that study did go forward and was published led by the VA group in Seattle. Unless you can figure out to what extent that was the recruitment strategy (You might need to check with FDA records) I worry that you and the field greatly overestimates effectiveness and side effects of when these drugs are prescribed in clinical practice. This is especially true of harms of course	

Section	Reviewer & Affiliation	Comment	Response
Introduction (cont'd)	Public Reviewer #1 (Anonymous) (cont'd)	3. In your PICOTS - you have a column of population - - it seems to me that it might be more important to have the variable "representativeness" of the population study as something you classify. The CATD field has been bedeviled by highly selected populations leading to erroneous conclusions - - As I read later in the report - - so many populations of studies you describe have mean ages 70 or less. This really can't be representative to an older population which tends to develop this condition in large number much later than 75 and 80	
	Public Reviewer #2 (American Psychological Association)	The introduction alludes to methods to reduce dementia risk. Among other citations, a citation to the recent WHO report should be provided (https://www.who.int/mental_health/neurology/dementia/guidelines_risk_reduction/en/). Although the introduction of the current document states that cognitive training is promising for reducing dementia risk, largely consistent with WHO recommendations, the WHO review notes that evidence to support this recommendation is low.	Our introduction states: "Moderate-strength evidence showed that 1) cognitive training could improve the cognitive domain trained in patients with normal cognition." This was the conclusion of a 2017 AHRQ systematic review by our group. The report further stated that in patients with normal cognition, cognitive training did not improve cognition in domains other than the domain trained, that effects in patients with MCI were mixed and less favorable, and that there was minimal evidence regarding whether cognitive training delayed clinical progression to MCI or dementia. The WHO recommendations focused on risk of cognitive decline and dementia and for those outcomes do not materially differ from the 2017 AHRQ report. We revised the report to provide more context on the cognitive training results, including on risk of clinical disease progression.

Section	Reviewer & Affiliation	Comment	Response
Methods	Peer Reviewer #2	1. Page 16. It would be helpful to be more explicit in the report how the issue of attrition bias was used to include or exclude studies. The explanation in Appendix B provides information about how attrition bias is defined but not about how it was used to exclude studies. Attrition bias is a particularly difficult issue in studies of persons with CATD and the use of 10-20% (moderate) threshold may be too stringent. 2. For the reviews of the effectiveness of drug therapy, were unpublished studies (which can be found in prescribing information) included? If not, these should probably be considered. Other publications (e.g., The Medical Letter) include unpublished studies that were used in getting FDA approval.	1. Appendix B details that overall attrition for a given outcome at a given timepoint within a study was rated as high when it was >30% regardless of whether analyses were performed to try to account for attrition bias. Overall attrition also was rated high when it was >20 to 30% and no appropriate analysis was performed to try to account for attrition bias. When data for a given outcome at a given timepoint was rated as having high attrition bias, those results were excluded from analyses. Other results from the same study may have been included in analyses if attrition bias for those other results was not high. Outcome measures with 10-20% attrition were rated as having low attrition bias if appropriate analyses were performed to address attrition or were rated as having medium attrition bias if no such analyses were performed. These results were not excluded unless there was sufficient risk of bias from additional domains, as detailed in Appendix B. Some other systematic reviewer have been more permissive in including CATD drug trials with >40% loss to followup and even labeling them as having low attrition bias. Attrition is a very challenging issue with the CATD population, especially with longer term followup. We addressed this issue in the discussion section of the revised report. 2. As detailed in the Literature Search Strategy section of our report Methods, we searched for unpublished trials using ClinicalTrials.gov. AHRQ also opened a Supplemental Evidence and Data for Systematic Reviews (SEADs) portal for 30 days to solicit pharmaceutical manufacturer protocols with additional information about published or unpublished drug studies. No pharmaceutical manufacturers submitted information.

Section	Reviewer & Affiliation	Comment	Response
Methods (cont'd)	Peer Reviewer #3 (TEP)	1. I liked the exclusion of high risk of bias studies from further consideration. 2. I find relatively little use of 'meaningful clinical difference' with regard to outcomes. There is some mention (97-98) for ASAS-Cog, but I see little to no mention of MCD for MMSE, MoCA, etc. I believe the reliance on statistical significance of changes in outcome measures is misleading to clinicians and families. The reporting of satistically sigificant changes in low SOE studies may over represent the usefulness of interventions. Section summary statements which 'call out' interventions of slightly significant outcomes for low SOE studies is unwise in my opinion, and may encourage the use of these interventions.	1. Thank you for this comment. 2. In the revised report, we made several changes to better convey the magnitude and clinical meaning of treatment outcomes. For continuous measures, whenever data allowed, we transformed between-group differences into standardized mean differences (SMD) to give them a sense of scale. We added language to the methods to indicate that SMD $\geq$0.2 were considered small effects, $\geq$0.5 were considered medium-sized effects, and $\geq$0.8 were considered large effects. Then, compared to the continuous outcomes, we placed greater emphasis on responder analyses (likelihood of response exceeding a prespecified threshold), absolute risk differences, and number needed to treat for benefit or harm. We also added more language characterizing the magnitude of between-group differences in mean changes, including the uncertainty about whether mean differences of this size are clinically meaningful.
	Peer Reviewer #4 (TEP)	The methods are comprehensive and appropriate for the effort being undertaken. The inclusion and exclusion criteria are clear and make sense. Again the use of NINCDS-ADRDA, vs NIA-AA, should be explained, especially when using biomarkers as well as an outcome. It may be that there are less studies available but could have still been addressed and some included. All else appropriate.	As noted above, we revised the introduction to address NIA-AA clinical and research criteria. In the revised Methods section (Table 2.1 Study inclusion criteria), we clarified our broad inclusion criteria for defining the CATD reference group for studies on the accuracy of brief cognitive tests: "reference diagnosis group is CATD based on full clinical evaluation and/or neuropsychological testing with explicit diagnostic criteria (e.g., DSM-IV, DSM-5, ICD, NINCDS-ADRDA, NIA-AA), with or without expert consensus." We revised the report to explicitly note in all the results sections on cognitive test accuracy that none used NIA-AA clinical diagnostic criteria. We also noted this in the Limitations section and revised the future research section to recommend that future studies use the most updated available standardized criteria to clinically define participants with CATD and MCI.

Section	Reviewer & Affiliation	Comment	Response
Methods (cont'd)	Peer Reviewer #5 (TEP)	1. Study inclusion/exclusion criteria are well-described and seem appropriate if one accepts the basic premise that this review should be limited to clinical Alzheimer's type dementia only. 2. Search strategies are explicitly stated and logical. 3. Outcome measures are well-described and seem appropriate. 4. Page 43, Table 2.1 Study Inclusion criteria, footnotes below the table a. Typo – "Sithout pooling…" should probably be "Without pooling…" b. Abbreviations – Should add ROB. 5. Page 44, line 22 – Is "AMSTAR" an abbreviation that should be spelled out on first use? 6. Statistical methods appear appropriate. 7. Page 47 – Strength of evidence levels are defined on Pages 46-47 and the term "Insufficient" strength of evidence is introduced on Page 47, although the strength of evidence terms have been used repeatedly in previous sections of the report. Would be helpful to have some definition of these terms earlier in the report. 8. Page 47, "Insufficient" strength of evidence – Why are "findings" based on "insufficient" strength of evidence presented in the Results tables through rest of the report? Doing this seems to be misleading to the average reader of the report. Perhaps there is an explanation of why such results are presented and I missed it. When I got to the Results tables, I found it hard to interpret "findings" based on "insufficient" evidence.	1. Based on discussions with AHRQ staff and a Technical Expert advisory panel, the scope of this review was limited to clinical Alzheimer's-type dementia. 2. Thank you for this comment. 3. Thank you for this comment. 4. A. This error was corrected. B. ROB was added to the list of abbreviations. 5. AMSTAR stands for A Measurement Tool to Assess systematic Reviews. We spelled it out on first use and updated the list of abbreviations at the end of the report. 6. Thank you for this comment. 7. We added a definition of strength of evidence to the revised evidence summary and referred readers to the Methods of the main report for further detail. 8. Because evidence for many questions was judged insufficient, but we believed that readers would want to see some of these results, we decided not to cut all insufficient results out of the report. To cut length, we consolidated text and tables to minimize redundancy, cutting out text as much as possible. This left the summary tables as the place to report results, even when they were graded insufficient strength evidence.
	Peer Reviewer #6 (TEP)	My comments regarding methods pertain principally to the section on cognitive tests and have all been logged in previous sections of my comments.	Thank you for this comment.

Section	Reviewer & Affiliation	Comment	Response
Methods (cont'd)	Peer Reviewer #7	The methods and analytic plan quite clear. I think the bar may be too high for the N in trials.	There was no lower limit on trial sample size.
	Peer Reviewer #8 (TEP)	1. General: Statistical methods, literature search, and review of bias/quality well described and appropriate. 2. The inclusion criteria appeared justifiable. The exclusion criteria was clearly defined. 3. Page 43, Line 9: "Inclusion criteria may need to be restricted..." I understand that this is the original inclusion criteria set out prior to the literature review. However, I found it confusing to read and was left wondering, "Was it?" Perhaps a legend at the bottom with an explanation of the final inclusion criteria.	1. Thank you for this comment. 2. Thank you for this comment. 3. Thank you for catching this remnant from the protocol. We did not end up restricting biomarker studies by the interval between biomarker collection and autopsy. We deleted this comment from the report.
	Peer Reviewer #9	The authors have done a excellent job in stating their inclusion and exclusion criteria. To this reviewer the search strategy seemed comprehensive and well described with appropriate justification for their choices, In other word explicitly stated. The report and its key questions and the scope was guided by a analytical framework. Overall all the methods presented in this report seem appropriate.	Thank you for this comment.

Section	Reviewer & Affiliation	Comment	Response
Results	Peer Reviewer #2	1. Chapter 4. For each of the cognitive tests, it would be valuable to indicate to indicate the number of items and the length of time to administer. This will help with the usability of the report. In addition, the report would benefit from consistency of descriptions of the tests, perhaps with a standardized 2 or 3 sentence description that would be placed in the same spot for each test. 2. The section might be better organized by classifying the tests into those that a primary care provider is likely to perform and those that would be considered neuropsychological testing and performed as a battery of tests by a trained psychologist. Moreover, neuropsychogical testing includes multiple tests that are performed with clinical interpretation integrating the results of these. Focusing on individual components may be artificial and misleading. 3. I was surprised that the CANTAB, 3-item recall, and Mini-Cog were not included. 4. Chapter 5. I'm not sure how the order of presentation of the tests was determined but I suggest beginning with those that are readily available in clinical practice. 5. Although CT scans with and without contrast are not very good for CATD diagnosis, they are commonly ordered and I recommend that the review covers their test characteristics. 6. If appropriate, it would be good to include recently published results from the IDEAS study.	1. As recommended, we revised the report to add information about each of the cognitive tests for which we analyzed information on accuracy for distinguishing CATD from either normal cognition or MCI. For tests often used for screening, we added a short, standardly formatted description of each of the cognitive tests to the report text (including scoring range, direction that indicates better score, and time of administration). For brief batteries and domain-specific tests, we created an appendix table that includes this standard information plus additional notes. 2. In our revised report, we did not change the organization of brief cognitive tests evaluated, but tried to better explain how those tests are usually used in clinical settings. 3. Our search criteria were broad and captured these tests. However, no studies of these tests both met eligibility criteria and were low or medium risk of bias. Reasons for exclusion varied across studies, but included that the tests were not administered in English, sample size was <25, studies were designed to distinguish nonspecific dementia rather than CATD from MCI or normal cognition, studies looked at screening in primary care or community settings regardless of whether there was suspicion for cognitive impairment or even excluded patients with dementia, studies did not evaluate all participants who completed brief cognitive testing with a full diagnostic evaluation, and did not meet any of the accepted CATD reference standards (e.g., used another slightly longer cognitive test as the reference standard). We revised the Limitations section to clarify this issue. 4. The order of presentation for brain imaging tests was based mostly on the number of analyzed studies. 5. We found no studies of CT brain imaging that met criteria for analysis. However, CT is widely available and frequently performed inpatients being evaluated for CATD. We commented in the report that the lack of CT brain imaging studies meeting criteria for analysis was a limitation of the review. 6. The IDEAS study was published after the search date for our review, so we did not formally screen, rate or extract it. However, because of the relevance of the IDEAs study to our question on the diagnostic accuracy of brain imaging for determining whether CATD is due to Alzheimer's disease, we added discussion of this study and its clinical implications to our revised report. 7. We found no studies on the accuracy of blood tests for diagnosing AD versus non-AD dementia that used a neuropathological reference standard. We stated this in the Key Messages and results of the Evidence Summary and in the Results and Limitations sections of the main report. We also recommended that future research be conducted to evaluate the accuracy of blood markers.

Section	Reviewer & Affiliation	Comment	Response
Results (con't)	Peer Reviewer #2 (cont'd)	7. Studies of blood tau predicting CATD are beginning to be published and it might be worth a short section on this. https://www.ncbi.nlm.nih.gov/pubmed/29641555 8. The extremely high specificity of amyloid PET is likely the result of the approach used (autopsy determined CATD). In fact, in more general populations, there are substantial false positives. https://jamanetwork.com/journals/jama/fullarticle/2293295 9. Chapter 7. Some of the commonly used supplements (e.g., Prevagen [apoaequorin], phosphotidalserine, Huperzine) are not included. There is also an evidence base on B6, B12, and folate for dementia that is not covered. 10. Chapter 9. At least one of the studies supporting the effectiveness of citalopram used doses that are no longer recommended. This might be commented on. Other drugs for behavioral symptoms (e.g., dextromethorphan-quinidine) have been tested for effectiveness for agitation in dementia and might be included. https://jamanetwork.com/journals/jama/fullarticle/2442936 11. Page 172, paragraph 2. There is a typo. "Trails" should be "trials".	8. We already discussed that the high diagnostic accuracy of amyloid PET in analyzed studies likely exceeds what would be achievable in clinical practice for multiple reasons. We expanded on this discussion in our revision. 9. For supplements, our review commented on the evidence for vitamin B and folate in the Chapter 7 Additional Supplements section. We found no eligible studies for the Prevagen (apoaequorin), phosphotidylserine, or Huperzine and added language explicitly stating this in the revised results and discussion. 10. We added language to the section on citalopram to clarify that the dose utilized in the trial (30 mg/day) exceeds the maximum dose now recommended (20 mg/day) and that results with 20 mg/day are not known. Based on discussions between the review team, AHRQ and the Technical Expert Panel, dextromethorpan-quinidine was not included among the prescription drugs/classes to be covered in this review. Therefore, trial data on this treatment was not formally evaluated, extracted or analyzed. In the revised discussion, we briefly described this dextromethorphan-quinidine trial and noted that its exclusion was a limitation of the review. 11. We corrected this typo.

Section	Reviewer & Affiliation	Comment	Response
Results (con'td)	Peer Reviewer #3 (TEP)	1. Excellent job. I think many clinicians find Forrest plots (in appendix) more useful than the evidence tables; I would have preferred to see the plots in the report itself. 2. I am not convinced that the inclusion of Chapters 7,8, and 9 are really useful. All of the evidence is insufficient or very low quality. I realize that once a key question is approved, the evidence report is supposed to address the question, but the chapters lengthen the review without offering much information.	1. We wrestled with how best to present the results, what information to include in the main report versus the appendix, and how to shorten the main report. We decided to limit the forest plots to the appendix. 2. We condensed these chapters in the revised report. However, we didn't eliminate them so readers could see the basis of our grading that evidence was low or insufficient.
	Peer Reviewer #4 (TEP)	1. The results though complex are laid out very well and organized. It is clear through the tables that the Sn Sp for each is best at CATD vs. NC. And much less accurate for CATD vs MCI. 2. Biomarkers look very interesting and screening and batteries as well. The tried and true ADAS does not fare so well in this. 3. Additionally, diagnostic confidence has been reported very recently in a significantly large population through the IDEAS study, published last month. This report will be woefully out of date if that data cannot be at least mentioned in the report discussion. 4. Again, this is especially relevant because the report does try to capture fluid biomarker and imaging marker evidence, and again the NIA-AA revision of the NINCDS ADRDA are not mentioned. Again this is a lack in the paper and will shorten the lifespan of it's relevance. 5. Well organized and understandable.	1. Thank you for this comment. 2. Thank you for your comment. 3. As noted above, the IDEAS study was published after the search date for our review, so we did not formally screen, rate or extract it. However, because of its relevance to our question on the diagnostic accuracy of brain imaging for determining whether CATD is due to Alzheimer's disease, we added discussion of this study and its clinical implications to our revised report. 4. We revised the report to address the NIA-AA criteria. 5. Thank you for this comment.

Section	Reviewer & Affiliation	Comment	Response
Results (con't)	Peer Reviewer #5 (TEP)	1. Page 48, Chapter 3 Search Results, lines 8-10 – Numbers for Key Questions don't seem correct. Shouldn't it be treatment KQ 3-8; cognitive testing KQ1; biomarkers KQ 2? 2. Page 48, line 34 – Typo - Should be "Figure" 3.1. 3. Page 50, Key Messages, lines 14-22 – a. Does first bullet mean that these cognitive tests identify CATD specifically (as opposed to indicating potential cognitive impairment generally that requires further evaluation or as opposed to some other type of dementia specifically)? Can this be clarified? b. First bullet seems to imply (falsely, I believe) that these brief cognitive tests can diagnose dementia/diagnose CATD. Distinguishing statistically between known groups is not the same thing as using a test prospectively in a clinical setting to diagnose a disease in an individual, right? Cognitive tests are one component of a diagnostic workup, but they are not diagnostic by themselves. The key point that brief cognitive tests are never diagnostic by themselves is important and should be made repeatedly. 4. Page 51, lines 53-54 – Is a comma missing in the following sentence after the word "criteria"? Participants with mild cognitive impairment (MCI) were diagnosed using Petersen criteria42 by specifying a Clinical Dementia Rating (CDR) score of 0.5, or both.	1. We corrected this ordering mistake. 2. This typo was corrected and thefigure numbering was corrected to match the order of the key questions. 3a. We looked for the accuracy of brief cognitive tests to distinguish CATD specifically from MCI or normal cognition, which is what our wording stated. The reference for these analyses was CATD defined by a clinical evaluation and/or neuropsychological testing, and not a reference of autopsy. 3b. It was not our intent to imply that brief cognitive tests can "diagnose" CATD. Our analyses examine how well performance on these cognitive tests can distinguish between groups in a research study. This is different from diagnosing individuals in clinical settings. We revised the entire report language away from "diagnosis" and "diagnostic accuracy" to "classification," "classification accuracy" and "distinguishing between" where appropriate. 4. A comma was missing and has been added. 5. We inserted the following information into the introduction to clarify the target audience for the report: "The target audiences of this report are primary care clinicians who diagnose and treat the vast majority of older patients with cognitive disorders, psychologists who may perform additional diagnostic cognitive testing in primary care settings, and dementia specialists who may be most likely to consider biomarker testing for further diagnostic clarification." With the section on brief cognitive tests, we were trying to get at the question of how good different tests were and whether there was a best test or combination of tests and best cut points. Though we retained much of this information in the revision, we tried to simplify the presentation. 6-17. We revised the tables to ensure that every abbreviation in each table was defined in the footnotes below it and that no abbreviations were included in a table's footnotes that didn't appear in the table. 18 & 22. We changed the phrasing to "Likelihood unchanged/improved." When strength of evidence was at least low, we used qualitative phrasing to describe results as "favors X" or "no difference" or "increased risk" or the like. However, when strength of evidence was insufficient, we did not include a qualitative phrase to suggest that results between groups were different or similar, but only reported the numerical results. For clarity, we added language explaining this approach to the report Methods section. Because there was very little even moderate strength evidence and even little low strength evidence for some treatment comparisons and outcomes, we elected to report results for the analyzed but insufficient strength evidence in the main report so it didn't get buried only in the appendix where few readers would be likely to see it.

Section	Reviewer & Affiliation	Comment	Response
Results (con't)	Peer Reviewer #5 (TEP) (cont'd)	5. Page 52, line 50 to Page 75, line 11 – Dense, technical results of cognitive test review. Who is the audience for this information? Seems like this extensive detail of sensitivity/specificity/ cut points from individual studies is appropriate for psychometric experts (e.g., psychologists) and researchers but not for clinical users. If primary audience for this report is primary care clinicians (or administrators, policy makers?), how should they make sense of all this technical information from individual studies? Seems like this whole section (and similar sections below for the other key question detailed study results) could go in the appendices and keep only overall interpretation/conclusions from this very limited and conflicting body of evidence here in the main body of the report. That would make this report much easier to read/comprehend for the average reader, and the study result details would still be available in appendices. 6. Page 54, line 17 – Should spell out "ROC" first time used. 7. Page 79, Table 5.2 – Abbreviations below the table need to add TP, TN, FP, FN. 8. Page 82, Table 5.3 - Abbreviations below the table need to add TP, TN, FP, FN. 9. Page 85, Table 5.4 - Abbreviations below the table need to add TP, TN, FP, FN. 10. Page 88, Table 5.5 - Abbreviations below the table need to add TP, TN, FP, FN. 11. Page 90, line 22 – Should spell out "AUC" first time used.	19. The first paragraph was incorrect and should have said that in both participants with mild to moderate CATD and those with moderate to severe CATD, donepezil was associated with an increased risk of withdrawals due to adverse events compared with placebo. This error was removed in the revised report. 20. The wording should have been "no withdrawals due to adverse events." This error has been removed in the revised report. 21. This sentence should have said the 5 mg/day was favored over placebo in just two trials. In the revised report, this error has been corrected and results have been updated and pooled. 22. Same question as #18, so answered together above. 23, 25, 27 & 28. When strength of evidence was at least low, we used qualitative phrasing to describe results as "favors X" or "no difference" or "increased risk" or the like. However, when strength of evidence was insufficient, we did not include a qualitative phrase to suggest that results between groups were different or similar, but only reported the numerical results. For clarity, we added language explaining this approach to the report Methods section. Because there was very little even moderate strength evidence and even little low strength evidence for some treatment comparisons and outcomes, we elected to report results for the analyzed but insufficient strength evidence in the main report so it didn't get buried only in the appendix where few readers would be likely to see it. 24. The statement about MENFIS results being reported in one study but not being included in the narrative of the report was removed from the revised report. 25. Same question as #23, #27 and #28, so answered together above. 26. In the revised report, we eliminated all the narrative results text that followed the summary results tables and relied on modified tables to communicate the results without being redundant between tables and text. With this revision, the text the reviewer refers to here was removed. 27. Same question as #23, #25 and #28, so answered together above. 28. Same question as #23, #25 and #27, so answered together above. 29. This was supposed to say "statistically differ" and was corrected in the revised report. 30. The "depress" in the anticonvulsants versus placebo key messages should have been "depression" as the reviewer stated. This was corrected in the revised report. 31. The correct spelling is Yokukansan and we corrected the misspelled Yokansan in the revised report. Also, we added language in the revision to inform the reader of what Yokukansan is beyond that it is a supplement (i.e., traditional Japanese herbal mixture).

Section	Reviewer & Affiliation	Comment	Response
Results (con'td)	Peer Reviewer #5 (TEP) (cont'd)	12. Page 94, Table 5.7 - Abbreviations below the table need to add TP, TN, FP, FN. Also, AUC is in the list of abbreviations, but is it in this table? Same for NR – it's in the list of abbreviations, but is it in this table? In general – Should check that the list of abbreviations below each table includes all the abbreviations in that table. 13. Page 96, Table 5.8 - Abbreviations below the table need to add TP, TN, FP, FN. Also, AUC is in the list of abbreviations, but is it in this table? Same for NR – it's in the list of abbreviations, but is it in this table? 14. Page 98, Table 5.9 - Abbreviations below the table need to add TP, TN, FP, FN. Also, AUC is in the list of abbreviations, but is it in this table? Same for NR – it's in the list of abbreviations, but is it in this table? 15. Page 99, Table 5.10 - Abbreviations below the table need to add TP, TN, FP, FN. Also, AUC is in the list of abbreviations, but is it in this table? Same for NR – it's in the list of abbreviations, but is it in this table? 16. Page 99-100, Table 5.11 - Abbreviations below the table need to add TP, TN, FP, FN. Also, AUC is in the list of abbreviations, but is it in this table? Same for NR – it's in the list of abbreviations, but is it in this table? 17. Page 107, Table 6.2 – Abbreviations below the table need to spell out all abbreviations in the table, including ones in Findings column (RR?). At this point, I'm going to stop examining each table at this level of detail and just say that all abbreviations need to be spelled out below each table throughout the report.	32. In the revised report, we corrected this error in the Chapter 10 key messages to state at 24 weeks or longer"

Section	Reviewer & Affiliation	Comment	Response
Results (con'td)	Peer Reviewer #5 (TEP) (cont'd)	18. Page 107, Table 6.2, line 10, line 29, line 41 – Phrasing ("favors") is unclear: "Likelihood of no change or any improvement in ADAS-Cog (>0-point from baseline): Favors donepezil…." Both "Likelihood of no change or any improvement" (what does that mean? Do you mean no change NOR any improvement?) and "favors" (what does that mean in the context of this sentence?) are unclear. 19. Page 110, lines 23-35, Withdrawals due to adverse events – Statements unclear. First paragraph says participants not more likely to withdraw…second and third paragraphs say participants are more likely to withdraw? 20. Page 110, lines 36-37 – Is a word missing or should there be a different word than "trials" at end of this sentence: "Two other small trials with participants with moderate-to-severe CATD (n=63) reported no withdrawals due to trials." 21. Page 111, lines 33-35 – This sentence is unclear. Says 10 mg/day favored in all trials, then 5 mg/day favored in two trials. "For clinical impression of change (CIBIC-plus or CGIC), results favored 10 mg/day donepezil over placebo in all trials,154,161, 168 but favored 5 mg/day donepezil compared with placebo in just two of trials.154, 168" 22. Page 118, Table 6.6 – Similar to my comment above on Page 107, Table 6.2 – For this table and any others throughout the report that use this phrasing – Both "Likelihood of no change or any improvement" (what does that mean?) and "favors" (what does that mean in the context of this sentence?) are unclear.	

Section	Reviewer & Affiliation	Comment	Response
Results (con'td)	Peer Reviewer #5 (TEP) (cont'd)	23. Page 118, Table 6.6 – Comment for this table and any others in report that use similar phrasing: What does it mean when the Strength of Evidence is labeled "insufficient" in the far-right column, yet detailed findings are reported in the middle column? How should the reader interpret this information? Should the reader accept the "findings" even though they are based on "insufficient" strength of evidence? Should "insufficient" strength of evidence "findings" be reported at all in this or other tables? 24. Page 127, lines 29-30 – What is the purpose of this sentence, to mention certain results of a study are available and say they are not reported here? Does the sentence add anything to this report? "Results for the Mental Function Impairment Scale (MENFIS) were available from one trial,198 but are not reported here." 25. Page 130, Table 6.10 – Same comment as for Page 118, Table 6.6 (and for subsequent tables in the report that also do this) – Why include all the "findings" from "insufficient" strength of evidence?26. Page 141, lines 37-40 – You mention 3 studies here. Can you give a brief summary statement of their results here (as you do for results in the next 2 paragraphs, for example), rather than just refer reader to the Appendix? 27. Page 142, Harms, lines 8-25 – Again just wondering why all the findings from "insufficient" strength of evidence are reported, rather than just saying there was insufficient strength of evidence on those topics.	

Section	Reviewer & Affiliation	Comment	Response
Results (con'td)	Peer Reviewer #5 (TEP) (cont'd)	28. Page 153, lines 48-56 – Just as another example – Are the results you report in this paragraph the ones you are saying in the opening sentence are graded "insufficient" strength of evidence? Not clear why such results would be reported in detail, rather than just saying the evidence on this topic is insufficient. I have the same question about all other paragraphs in which you give detailed results apparently from "insufficient" evidence. Seems like those results could just go in the Appendices. 29. Page 170, lines 46-47 – Typo, "statistically" is repeated in this sentence; some other word is missing (probably should be "did not statistically differ…": "For aggression, standard- and low-dose haloperidol did not statistically statistically in mean reduction of either of two aggression symptom scale scores." 30. Page 179, line 34 – Typo, should be "depression" rather than "depress." 31. Page 183, line 14- What is "Yokansan"? Also, on line 38 and below it is spelled differently as "yokukansan." 32. Page 183, line 18 – Typo, word missing. Should be "…followup > 24 weeks"?	
	Peer Reviewer #6 (TEP)	No major concerns other than noted above.	Thank you for this comment.

Section	Reviewer & Affiliation	Comment	Response
Results (con't)	Peer Reviewer #7	Chapter 4: 1. Very extensive. I think it would be more helpful to focus on brief screening tests than so many individual cognitive tests that are rarely used on their own (versus a battery of tests). 2. The patient characteristics that would be most helpful to investigate are high vs low education, English as second language. Chapter 5: 3. A definition of biomarker would be helpful. 4. The comparison of AD to ALS does not make much sense clinically. ALS-FTD is a syndrome but not ALS per se. Chapter 6: 5. The addition of QOL to the outcomes is important and overdue. Chapter 8 & 11: 6. It seemed a bit strange to have so much on comparative effectiveness when the SOE for each drug is fairly low.	1. During the process of determining the scope of this review, our Technical Expert Panel advised including brief cognitive tests in addition to screening tests (e.g., brief batteries, domain specific), and including both tests that a primary care provider may administer and that a psychologist embedded within primary care could quickly administer (generally <30 minutes). One of the reasons for examining the classification accuracy of any brief cognitive test, as opposed to only those developed for screening, is to assess the test's potential application for this use (CATD case finding in primary care), even if the test is not commonly used for this purpose at present. We added this rationale to Table 2.1 "Study Inclusion Criteria" in the revised report. 2. We looked at years of education and evaluated when studies reported whether classification accuracy differed as a function of education (defined as high vs. low or otherwise). We included studies in which at least 75% of the participants had the tests administered in English. We did not systematically look for whether studies reported information about whether participants were nonnative English speakers or whether classification accuracy varied as a function of this patient characteristic. We suspect this information was rarely if ever reported, but this may be a good question to examine for future research. 3. We clarified the meaning of biomarker in the revised report. 4. We agree with the reviewer comment that distinguishing between AD and ALS is not a clinical question, and we removed this comparison from the revised report. 5. Thank you for this comment. 6. One of our objectives with this review was to report on the comparative effectiveness between different drug treatments.

Section	Reviewer & Affiliation	Comment	Response
Results (con't)	Peer Reviewer #8 (TEP)	1. Page 20, Line 6: There is an error in the table. There are 2 columns titled "SN, median." There is not SP column. 2. Page 23, Line 18: I found the cell in Table 0.4 for memantine co-administered with AChEI in Mod-Severe CATD describing cognition confusing. Why would a screening test be performed in known moderate to severe CATD? What does this cell mean to convey to the reader? This cell is presented differently from the others in the table and I was uncertain as to how to interpret it. 3. Page 50, Lines 11-36: Recommend adding to key messaging that no studies reported data on harms of brief cognitive testing for diagnosis CATD. This is important knowledge for primary care physicians to have when considering conducting these tests or screening instruments. 4. Page 53, Line 35-38: Discussed stratification of CATD severity (ex: very mild, mild, moderate) but there is not a definition of how these are determined and if this is beings standardized throughout the report. This occurs at multiple instances throughout the report. Please provide the criteria used to determine CATD severity 5. Page 76, lines 12-40: The Key messaging should include a description of the findings regarding harms. This is important for primary care physicians and is a major component of this key question. 6. Page 90-91: The Key messaging should include a description of the findings regarding harms (including when there are no studies on the harms). This is important for primary care physicians and is a major component of this key question.	1. The second "SN, median" column in table 0.1 was supposed to be headed 'SP, median." This was corrected in the revised report. 2. In CATD drug treatment trials, we looked at the results of different cognitive tests. This included looking at results of tests that are commonly used for diagnostic screening, but that also are used to measure change in cognition in patients with established CATD. For the sake of brevity, instead of saying "test commonly used for screening," we said "screening tests." There really isn't another good brief descriptor we could think of that accurately characterizes this collection of cognitive tests. 3. We revised the report to add a comment to the key messages that no studies reported data on harms of brief cognitive testing for distinguishing CATD from normal cognition or MCI. 4. CATD severity was stratified either based on the descriptor used by the study author (e.g., if the author stated in the article that the patients had "mild to moderate" dementia, we characterized their CATD as "mild to moderate." If the authors didn't apply a descriptive severity term, we used the range of baseline MMSE to characterize the severity, including 20-30 as mild, 10-19 as moderate, and <10 as severe. We revised the report to clarify this. 5. We revised the report to add a comment to the key messages about harms of brain imaging techniques for distinguishing AD from non-AD dementia. 6. We revised the report to add a comment to the key messages about harms of CSF tests for distinguishing AD from non-AD dementia. 7. No studies of brain imaging or CSF testing for diagnosis of AD reported on whether accuracy varied by severity of disease severity. We specified in Key Question 2a and PICOTS table 1.1 which patient characteristics we identified a priori to examine for whether they modified biomarker accuracy (i.e., age, sex, race/ethnicity, depression, education, pre-testing cognitive or functional level/CATD stage). In the draft results sections on Variation in Diagnostic Accuracy by Test of Patient Characteristics, we did not list when no studies evaluated these characteristics for potential effect modification. Rather, we noted when they were examined and otherwise made a general statement about no studies reporting data on whether accuracy varied by patient characteristics. 8. We revised the report to add to the key messages a comment about there being minimal data about whether prescription drug or supplement treatment efficacy varies by participant characteristics. 9 & 12. In our updated literature search, we found evidence on tests often used for screening in addition to MMSE. Though these tests often are used for screening and have been called "screening tests," they also often are used to measure disease progression, as is the case for these trials. For the sake of brevity, instead of saying "test commonly used for screening," we said "screening tests." There really isn't another good brief descriptor we could think of that accurately characterizes this collection of cognitive tests.

Section	Reviewer & Affiliation	Comment	Response
Results (con'td)	Peer Reviewer #8 (TEP) (cont'd)	7. Chapter 5: I was left wondering if any of the studies looking at imaging for diagnosis evaluated the disease severity at that time of imaging and if that impacted sensitivity and specificity. 8. Page 105, Line 12-30: Recommend adding to Key Messages that there are no studies that evaluate effect on donepezil as a function of patient characteristics. This is a major question for PCPs, one of the target audiences, and would be valuable information in the key messages. 9. Page 108, Line 22: This section is titled screening tests but these tests are not being used for screening. This is confusing as these individuals have already been identified as having CATD and the tests are being used for monitoring for change. Recommend title as MMSE, which is the test being evaluated here. This holds true in Table 6.2 as well. 10. Page 113, Line 9: I appreciated that this table included criteria for how severity of CATD was defined. This was useful information and I feel should be replicated throughout the manuscript where applicable. 11. Page 113, Table 6.4, Line 23: This table is challenging to read. As most of the Findings column has "moderate to severe" and "mild to moderate" and the strength of evidence column also has two subcategories, I think that splitting the table or adding additional columns/rows to accommodate these subcategories may add clarity.	10. Rather than repeat the phrase "Severity was defined by the individual trial inclusion criteria, typically DSM-III or IV and/or NINCDS-ADRDA criteria in addition to a MMSE score within a pre-specified range" below every Baseline Study Characteristics table throughout the report, we revised the Methods in the report to detail how disease severity was defined. 11. We revised the formatting of the outcomes tables to try to make them easier to read within the constraints allowed by AHRQ. 12. Same as question 9, so answered together above. 13. Serious adverse events (SAE) is specifically defined by the FDA. This is specified in the PICOTS table in the Methods section of the report. We reported results for SAE only when the trials specifically listed results for that exact term. We did not decide that some adverse events were serious and count these as serious adverse events, either alone or as a composite. 14. We revised the report to specify in the antidepressants results section that "We identified no eligible trials that evaluated the efficacy or harms of serotonin-norepinephrine reuptake inhibitors (SNRIs) (e.g., venlafaxine, duloxetine), tricyclic or tetracyclic antidepressants (TCAs/TeCAs) other than mirtazapine (e.g., amitriptyline), serotonin modulators (e.g., trazodone), or norepinephrine-dopamine reuptake inhibitors (NDRIs) (e.g., bupropion) for treating BPSD in patients with CATD." 15. Thank you for this comment. However, in our revision, we tried to reduce the redundancy between the tables and the text to shorten the results sections. 16. We revised the formatting of the outcomes tables to try to make them easier to read within the constraints allowed by AHRQ.

Section	Reviewer & Affiliation	Comment	Response
Results (con'td)	Peer Reviewer #8 (TEP) (cont'd)	12. Page 114, Line 49: This section again uses "Screening Tests" as a header incorrectly. This is a recurrent issue in the manuscript as the test in question (MMSE) Is not being used as a screening instrument in this setting but rather for tracking disease progression/severity. Would change category header to MMSE throughout the manuscript. 13. General: It was not clear to me what is being defined as a "serious adverse event." 14. Page 171, Line 10: I think it would be worth noting in this section that the trials included only involved SSRIs or mirtazapine. SNRIs, bupropion, and other TCAs were not evaluated in the trials that met inclusion criteria. 15. General: The amount of detail presented in the results section for each key question was appropriate. 16. In general, I found the tables to be challenging to read and navigate. There is a lot of text in bulleted or paragraph form and it can be difficult to get the key information quickly.	
	Peer Reviewer #9	The results are clearly presented as well the main findings are clearly articulated in the results. This reviewer found the result section to be clear and results clearly addressed each of the key questions. It is always challenging to summarize results in a systematic review that has multiple key questions but authors have done of good job providing enough detail and at the same time presenting the key findings in a succinct and clear fashion.	Thank you for this comment.

Section	Reviewer & Affiliation	Comment	Response
Results (con't)	Public Reviewer #1 (Anonymous)	1. page 23 - - I really doubt that you have achieved study selection for 42 studies of low or medium risk of bias - - It would be helpful to detail how you came up with this classification scheme. You might consider generalizability or representativeness in addition to bias. 2. page 24 - - These really aren't screening tests - I think you should describe them as case finding. 3. Page 35 and elsewhere - - Memory tests are said to have a participant mean age of 70 years - as above this is hardly an age range that represents a community population in a provider's office. I suggest you acknowledge this an also make sure that for every study your report you add a column indicating mean age and range or interquartile range for the studies you summarize in tables in this chapter. I suggest the same for chapter 5 and other chapters with tables summarizing results. Remember the strongest predictor of this disease is age	1. It isn't clear what the reviewer means by "achieved study selection." In our methodology, we described how we identified potential studies, eligibility criteria for inclusion in the review, how we triaged them for eligibility, and how we rated them for risk of bias to determine which of the eligible studies to extract and analyze. 2. These tests are sometimes used for screening in patients who do not carry a diagnosis of CATD. In this review, we are looking at their diagnostic accuracy when administered in individual in whom there is a suspicion for CATD, which is case finding. Although we described how we were evaluating these tests, not all readers may have recognized this as an assessment of case finding. So, we revised the report throughout to add case finding language. 3. In the revised report we include summary age and other demographic data in the body of the report and age and other demographic data for each individual study in the appendix. In the applicability section of the revised report, we commented on the generalizability of study participants to the age of patients seem in typical practice. 4. This review evaluated the accuracy of brief cognitive tests for distinguishing CATD from MCI or normal cognition at the time the brief cognitive tests were administered, not for what would best predict these diagnoses over time. The latter is an important clinical question, but one that was outside the scope of the present review. We added language to the applicability section of the revised report stating that the current review did not address the accuracy of brief cognitive tests or biomarker tests for predicting clinical progression to AD over time. 5.We were targeting cognitive tests with administration times of approximately 30 minutes or less and clarified this in Table 2.1 on study inclusion criteria in the Methods of the revised report. We did not revise the report to indicate which cognitive tests were proprietary. 6. We tried to maintain consistent reporting on studies throughout the report. In the body of the report, we reported age, other demographic data, and MCI and CATD definitions used only at a summary level, whereas we report this information for each individual study in the appendix. Studies on the classification accuracy of brief cognitive tests were cross-sectional, so that the interval between brief cognitive testing and completion of the full clinical diagnostic evaluation was supposed to be minimal.

Section	Reviewer & Affiliation	Comment	Response
Results (con't)	Public Reviewer #1 (Anonymous) (cont'd)	4. You also don't acknowledge that the passage of time for a progressive disease like CATD is likely a more accurate tests and is consistent with your general finding that tests perform better in distinguishing CATD than they do for MCI and other lower "disease" categories. This is just the nature of an age related condition - - with a lot of heterogeneity as depending on how long people are followed you get very different test accuracy results as certain people were probably temporarily "false positives" only to never progress. this phenomenon is under appreciated in the literature but has been reported in papers by Paul Crane and others related to MCI. 5. Describe the definition of brief and also note that many providers not in research have switched from MMSE to MOCA when MMSE was copyrighted and went for profit through attorneys threatening copyright infringement lawsuits and compensation. 6. page 40 good idea to describe the source of subjects for the tests of executive function and as before add a column describing age and ranges of persons in the studies. Also which of the Petersen definitions of MCI did they use. And for this and other tables - duration of follow up See page 43 would be valuable.	7. The false positive biomarker tests are a big issue in studies examining their cross-sectional accuracy in patients with normal cognition or MCI, or examining their ability to predict future clinical progression. This should be a smaller issue in this report because we looked at the ability of biomarkers to distinguish between AD and non-AD in patients with prevalent clinical dementia. In the revised report, we tried to make these distinctions between what we did and didn't do more clear. Data reported in table 5.1 and the other baseline characteristics tables do not present the full range of all participants in all studies, but report the range of the different study means. We revised the tables to more clearly communicate this. We changed the headings from "range between studies" to "study range" to try to make this more clear. The FP, TP, FN, and TN rates were calculated by our evidence team from sensitivity and specificity and prevalence of the included studies. These are based on the study samples and may not be representative of what would be observed in typical clinical settings. We clarified this in the revised Methods Data Synthesis section. We revised the report Applicability section to state that these FP, TP, FN and TN rates may not be generalizable to populations with a lower CATD prevalence. 8. This illustrates why we included information about the cut points utilized in the results tables. 9. In the revised report, we included more discussion about the degree to which study results are generalizable to typical clinical populations. 10. As discussed above, run-in phases (pre-randomization screening) were used in early AD clinical trials, particularly after the first tacrine study was published (JAMA,1992), which documented an elaborate "enrichment" design that was clearly described in its Methods section. We did not document pre-randomization screening of potential trial participants with either placebo (to gauge likely adherence and placebo response) or active treatment (to also gauge likelihood of treatment benefit or harms). However, to address this question, we re-examined the prior donepezil trial systematic review and the two oldest donepezil versus placebo trials (Rogers, 1998; Burns 1999). None described a pre-randomization run-in phase. Because these trials predated required registration in clinicaltrials.gov, there was no additional information on possible pre-trial screening available there. So, while we did not identify an example of pre-randomization screening among analyzed trials, if a trial used such a design, and excluded individuals from randomization who during screening were less adherent, experienced more adverse effects or had smaller therapeutic responses, it would be likely to bias results to make the treatment look more effective and more safe, and the results would be less generalizable to nontrial populations. 11. It is important not to overinterpret results from this single small trial (n=188). Also, there were at least some other trials that reported incidence of SAE withdrawals due to adverse events equally as high.

Section	Reviewer & Affiliation	Comment	Response
Results (con'td)	Public Reviewer #1 (Anonymous) (cont'd)	7. Chapter 5: This is the chapter where ignoring that there are false positive tests results that are missed since autopsy rates are much higher in persons who are demented compared to those without dementia. And we now know that people with abundant amounts of plaques and tangles will die in late life without clinical dementia based on careful cognitive assessment over time. Table 5.1 - is range the full range or a summary statistic of some sort? I like the way you've summarized age categories though. Table 5.2 and elsewhere, you might want to quality things like FP rate as "reported FP rate" My impression from following these reports for years is that the populations are so carefully and highly selected that they overstate the accuracy. 8. See page 62 and the huge difference in Sens and Spec based on where you set the cut point. 9. page 70 Striking to see on study with a mean age of 81.0, 55.4% male - - would be good to state the source of this population as it is strikingly different from almost all the other studies you cite. Also of interest in page 74 which is apparently a pooled group - how can you figure out the extent to which these findings are generalizable? Ideally A systematic review would help the reader sort this out.	

Section	Reviewer & Affiliation	Comment	Response
Results (con'td)	Public Reviewer #1 (Anonymous) (cont'd)	10. page 79 - I think this would be a good place to determine to what extent subjects were prescreened for possible side or beneficial effects BEFORE being considered for enrollment. I realize this is potentially a lot of work but at least in earlier years that was the standard - at a time when "the world" was anxious for anything that would work so standards for efficacy were made to be as favorable as possible to find benefit, based on theoretical concerns. 11. on page s 127 -8 it's striking just how frequent SAE and withdrawals were in these studies - - How come they were so much higher compared with harms reported elsewhere with the same drugs? I think this should be pointed out.	
	Public Reviewer #6 (AAFP)	Table 4.2, 4.3 - Add LR+ / LR- Ideally, provide PV+ and PV- for a typical prevalence of CATD. P. 82 - Please focus where possible on absolute rather than relative risks (latter inflate apparent benefit) and provide NNT. P. 84 - Elsewhere you use "insufficient SOE" how does this dffer from low SOE? Table 5.3 - Again, how is this calculated? Using prevalence of 0.64 for row 1 and prevalence of 0.69 for row 2? That would not make sense…	

Section	Reviewer & Affiliation	Comment	Response
Discussion/ Conclusion	Peer Reviewer #2	1. My sense is that the statement on page 165 paragraph 1, "it may not be necessary in a primary care setting to use different cut points in different populations" is too strong. Literacy and primary language are important influences on the ability to obtain valid information and should be considered in the interpretation of brief cognitive tests. 2. There are some differences between how MRI medial temporal atrophy is rated between the Executive summary (more positive) and the text and Conclusions (less positive). 3. On page ES-2, the authors state that "in patients receiving cholinesterase inhibitor, memantine improved function in moderate to severe CATD but on page 114, they that function was not improved.	1. The report did not consider whether accuracy of brief cognitive tests to distinguish between CATD and either MCI or normal cognition varies by literacy or primary language, though we sought evidence about whether classification accuracy differs by years of education. We modified this language to be more cautious in the revised report as follows: "These studies were likely too small to rule out such associations and inadequate for concluding whether or not different cut points should be used for classifying between CATD and normal cognition or MCI in different clinical populations." 2. We revised the report so that assessments of MRI medial temporal atrophy evidence were consistent throughout. 3. On page ES-2, the draft report stated that in patients with moderate to severe CATD, "in patients receiving cholinesterase inhibitors, memantine…did not improve function." This is consistent with the information on page 114 from the draft report which states that results for function did not differ between add-on memantine and add-on placebo.
	Peer Reviewer #3 (TEP)	Good job	Thank you for this comment.

Section	Reviewer & Affiliation	Comment	Response
Discussion/ Conclusion (cont'd)	Peer Reviewer #4 (TEP)	1. The discussion is well outlined and acknowledges the data that is lacking to draw certain conclusions. It is expansive, and questions 7 and 8 are a bit superfluous in my view. 2. Again, some limitations can be addressed with the recent publication of the IDEAS Study regarding amyloid imaging. 3. In the limitations, this sentence is not accurate: "The applicability of study findings on the accuracy of brain imaging and CSF biomarker testing for AD also is limited because these tests are not easily available in typical clinical settings in the U.S." The IDEAS study demonstrated that amyloid imaging is indeed accessible and utilized all around the country.	1. Thank you for this comment. 2. We reviewed the IDEAS study closely. While the study did not meet criteria for inclusion in our review because it did not compare amyloid PET results to a neuropathological reference standard, we discussed it in detail in our revised discussion. 3. The IDEAS study suggested that the mostly private practice specialists participating in the study had access to amyloid PET interpreted by local imaging specialists. This suggests that accessibility to this testing, at least to dementia specialists, may be greater than we implied in the draft report. We revised the report to clarify our wording on this point.

Section	Reviewer & Affiliation	Comment	Response
Discussion/ Conclusion (cont'd)	Peer Reviewer #5 (TEP)	1. Discussion section – limitations are described well. This is a critically important section of the report. 2. Discussion section – future research recommendations are described well. 3. Page 192, line 29 – Typo, word "being" repeated in sentence – "stigma from being correctly or incorrectly being labeled…." 4. Page 197, line 7 – Is "Yokansan" correct spelling? 5. Page 197, line 9 – Typos, word repeated and word missing from sentence - "Evidence was mostly was insufficient draw conclusions about these findings." 6. Page 198, section on medical marijuana – Were results on this topic mentioned in Results section? I may have missed it. Just seeing it in Discussion section. 7. Page 198, line 34 – Is "involving" the tense you want? Or should it be "involve" or "involved"? 8. Page 198, lines 55-56 – Typo, "few" is repeated. 9. Page 199, line 21 – Should be "trials" (not trails).	1. Thank you for this comment. 2. Thank you for this comment. 3. We corrected the report to eliminate the extra "being" in this part of the discussion. 4. We corrected the spelling of Yokukansan in the revised report. 5. We corrected the revised report to state "Evidence was mostly insufficient to draw conclusions…" 6. We stated in the Methods that medical marijuana was among the medications about which we sought evidence for its efficacy and harms on BPSD. We found no eligible trials of medical marijuana. We revised the Results to more clearly communicate drug interventions for which we found no eligible trials or for which the only eligible trials were high risk of bias and weren't extracted or analyzed. 7. We corrected the tense to "involved" in the revised report. 8. We corrected the revised report to remove the second "few." 9. We corrected the revised report to correct "trails" to "trials."

Section	Reviewer & Affiliation	Comment	Response
Discussion/ Conclusion (cont'd)	Peer Reviewer #6 (TEP)	The guidance provided for how studies should be conducted in the future could be summarized in a higher-level view tailored to each intended audience. While suggestions for future work on each key question/research domain are excellent in my opinion, it's important not to risk leaving the reader with the general impression that evidence is SO insufficient that we might as well pack up and go home. There are many underlying and interacting probable causes of this insufficient evidence: intractable neurodegenerative disease in an organ poorly prepared for self-repair; treatments that fail to address the (currently unknown) root causes of neuronal failure; overly broad formulation of research questions; no way (based on current understanding) to meaningfully analyze subgroups in most studies; incentives for research that privilege pharmaceuticals; outcome measures that miss the mark of what might realistically be expected - and others. While a report such as this cannot address (let alone resolve) all of these problems and more - nor can we expect that - some approach to the question 'what can and should we do now in the face of this difficult problem?' seems appropriate.	Because this report is not a guideline, other than summarizing and interpreting the current evidence and making recommendations for future research, we are constrained from making clinical recommendations. However, this report was conducted in part to provide evidence to inform an upcoming AAFP guideline, which should address this reviewer's question of "What can and should we do now in the face of this difficult problem?"
	Peer Reviewer #7	1. The Discussion is quite good and I liked the tone. It would be better if some of this material could go in Executive Summary as feels like written by different people. 2. I especially liked directions for future research. 3. One issue that is glossed over is the relatively little biomarker data on diverse elders.	1. The discussion and evidence summary were revised to be more consistent in content and voice. 2. Thank you for this comment. 3. The draft report stated in the results section that nearly all biomarker studies that reported data on race/ethnicity included predominately white participants. The revised report attempted to place greater emphasis on the relative lack of racial/ethnic diversity in these studies.

Section	Reviewer & Affiliation	Comment	Response
Discussion/ Conclusion (cont'd)	Peer Reviewer #8 (TEP)	1. Page 196, Line 30- Page 197 Line 15: I was surprised that there was minimal discussion of the risks of anti-psychotic use in CATD for stroke and mortality. 2. Page 200, Lines 17-42: Additional area of future research: Studies should be conducted in clinical settings outside of large, research settings. As identified in the applicability section, the studies were often performed with participants who have been followed for years in research settings, had well defined symptoms, and had low life expectancy. Many also used methods that are not feasible in clinical settings. Studies on imaging and CSF must be conducted with clinical feasibility in mind. 3. Page 200: Line 53: Not only should studies specify analyses to examine how patient characteristics effect treatment, they should be designed to examine this question. The majority of participants in the included studies were white. There should be a concerted effort to examine other patient characteristics and non-white individuals. 4. General: The discussion was clear, concise and clearly describes the findings. 5. The limitations are well reviewed. 6. The future research is clearly described and I agree with the recommendations with the additional recommendation I made above.	1. The revised report added additional discussion on the evidence about risks for stroke and mortality of antipsychotics in patients with CATD. 2. We revised the report to add this recommendation for future research, that in order to better determine applicability to typical clinical practice, future biomarker studies should include more typical CATD participants and should use methods that are feasible to employ in typical clinical settings. 3. We revised the report to also recommend that future studies make a concerted effort to enroll diverse participants, including nonwhites. And, besides prespecifying patient characteristics to examine for possible effect modification, we added a recommendation that individual studies be designed with adequate statistical power to examine whether treatment outcomes are modified by patient characteristics. 4. Thank you for this comment. 5. Thank you for this comment. 6. Thank you for this comment.

Section	Reviewer & Affiliation	Comment	Response
Discussion/ Conclusion (cont'd)	Peer Reviewer #9	The discussion presents the key findings based on the evidence they have synthesized in this report. The overall conclusion clearly reflect the state of the evidence. The usual limitations of this literature are provided and the suggestion for future research is similar to what has been proposed in previous evidence reports. However, nothing much has changed. It will be good for this review to suggest some new approaches to addressing the limitations of this literature so future studies can provide useful clinical guidance.	Thank you for this comment.

Section	Reviewer & Affiliation	Comment	Response
Discussion/ Conclusion (cont'd)	Public Reviewer #1 (Anonymous)	1. page 164 - - the report states "the most available accuracy data" when the reader is likely to want to know something about performance in the field. What that framing also does is leave out MOCA since it was only recently that there was enthusiasm for MOCA when people became frightened about consequences of using the popular MMSE. The question is "which are best" I don't think you answer that. 2. page 165 - - it's self evident that diagnostic accuracy varies based on patient characteristics. It is also self evident and certainly not stated here that "tincture of time" and serial observation is probably the best test in this disease - - I encourage you to mention that common sense observation and make a more direct answer. 3. page 165 - I really agree ""could not determine" is correct for which imaging, biomarkers etc is most accurate. I would argue that we really can't know even how accurate these tests are and refer to the heterogeneity of the neurodegenerative processes in the aging brain (diseased or not) and highly selected populations with published results. 4. page 166 you mention gains outside research settings are likely to be less but you don't explain why - you should	1. There was little available data on diagnostic accuracy of the MoCA, which limited our ability to draw conclusions about its accuracy. There was little data directly comparing the diagnostic accuracy of different brief cognitive tests, which limited out ability to draw conclusions about which brief cognitive test was most accurate. 2. We suspect it is true that diagnostic accuracy varies based on patient characteristics, but found minimal evidence that evaluated this question. We revised the report to be worded more cautiously that data were not available to conclude whether different tests or cut points should be used to classify individuals between CATD and normal cognition or MCI based on participant characteristics. As to the accuracy of "tincture of time," it also probably is true that the longer a patient is followed and characterized, the more accurate will be the clinical diagnosis. However, we were interested in examining the diagnostic accuracy of these tests at the time the tests are administered--when there still is clinical uncertainty. The accuracy of brief cognitive tests for predicting future clinical progression was outside the scope of our report. 3. Thank you for this comment. 4. The sentences that immediately followed "gains…are likely to be smaller outside of research settings" were the reasons. We revised the report to clarify that was what we meant with these sentences. 5. These numbers were from the proportion of patients rated as moderately or markedly improved on clinical impression of change. Also, we only reported the likelihood of this outcome in comparison with a randomized control group. Assuming randomization worked, the likelihood that this improvement was attributable to these nontreatment reasons should have been balanced between treatment groups. 6. We revised the report to mention MoCA as among the common tests warranting additional future research. 7. The false positive biomarker tests are a big issue in studies examining their cross-sectional accuracy in patients with normal cognition or MCI, or examining their ability to predict future clinical progression However, in this report, we looked at the ability of biomarkers to distinguish between AD and non-AD in patients with prevalent clinical dementia. In the revised report, we tried to make these distinctions between what we did and didn't do more clear.

Section	Reviewer & Affiliation	Comment	Response
Discussion/ Conclusion (cont'd)	Public Reviewer #1 (Anonymous) (cont'd)	5. page 167 You dont' say anything about harms for prescription drugs - if you agree you should probably have some sort of statement stating we really don't know the balance between benefits and harms. How you come up with the number s for moderate or marked improvement is a mystery - - maybe it was buried earlier in the report but there are so many reasons for moderate to marked improvement including intercurrent illness, change in location that I find this a statement that should be considered for deletion. 6. page 171 - I like the applicability and limitations sections - - If anything I think they are understated and you don't mention MOCA as amongst the common tests here or on page 172 for future research. page 173 see me earlier comments about the fallacy of autopsy confirmed AD which is not associated with clinical dementia - even in very old old persons followed for more than 10 years in many cases.	
Figures		No comments made	Not applicable.
References		No comments made	Not applicable.
Appendix		No comments made	Not applicable.

Section	Reviewer & Affiliation	Comment	Response
General	Peer Reviewer #2	Overall, the report is of superior quality and provides some important clinical insights. Perhaps equally important is the identification of areas where the evidence base is insufficient to answer the key questions. One disappointment is the omission of basic laboratory tests (e.g., kidney function tests, Vitamin B12, TSH) from the review. Although these are typically used to exclude other contributors to cognitive decline or dementia, they are part of the recommended evaluation of dementia and warrant an evidence-based review. The American Academy of Neurology guidelines are dated and this is an issue in the evaluation of every patient with dementia.	The aim in evaluating the diagnostic accuracy of brain imaging and CSF biomarkers was to examine the evidence for whether or not these tests are highly sensitive and specific. None of the basic laboratory tests the reviewer lists are specific for distinguishing AD from non-AD dementia and their evaluation was outside the scope of this review.
	Peer Reviewer #3 (TEP)	This is an extraordinary effort to categorize evidence regarding CATD diagnosis and treatment. The target audience is not only defined, but addressed throughout the report. The key questions are clear; unfortunately even moderate quality evidence was not available to address several questions important for clinicians	Thank you for this comment. To further clarify the target audiences, we added the following statement to the introduction of the report: "The target audiences of this report are primary care clinicians who diagnose and treat the vast majority of older patients with cognitive disorders, psychologists who may perform additional diagnostic cognitive testing in primary care settings, and dementia specialists who may be most likely to consider biomarker testing for further diagnostic clarification."
	Peer Reviewer #4 (TEP)	Overall the report is good, especially in reference to cog screening tests and to cognitive tests/batteries used to Dx CATD/MCI. The report attempts to do too much, in also moving to fluid and imaging markers and to expand to BPSD as well in my view. Questions 7 and 8 seem out of place in this, but in an effort to cover ALL it is long and cumbersome though the authors went to great effort to organize the questions into appropriate framework that could be followed throughout the document.	We agree that the scope of the entire report was extremely large and that this led to a long report. We worked to shorten the revised report and to minimize redundancy.

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Peer Reviewer #5 (TEP)	This analysis was undertaken in response to a request by an association of Family Practitioner's to address three issues regarding Alzheimer's disease, all of which may have relevance to primary but are only loosely related to each other. One issue the sensitivity and specificity of cognitive tests in distinguishing Clinically diagnosed Alzheimer's dementia as well as mild cognitive impairment from normal cognitive aging. The second was the sensitivity and specificity of neuroimaging and fluid biomarkers in distinguishing neuropathology diagnosed Alzheimer's from non-AD dementia. And finally address was the usefulness of various medications in treating progression to dementia, function, quality of life and disruptive behaviors associated with AD While the team completing this AHRQ review conducted a generally sound process, a few decisions would benefit from further explication. In general these fall under the heading of inconsistent treatment of various types of data:	1. For the search for biomarker studies, we judged that the prior systematic literature reviews, with the most recent having a 2012 search date, were sufficient to identify potentially eligible trials published before 2012. However, for the brief cognitive test and drug treatment questions, we did not find systematic reviews that covered their full scope. Therefore, we decided to search electronic bibliographic databases back to their inception for those questions. 2. Evaluation of combinations of brief cognitive tests were erroneously not included in the draft report and have been inserted in the revised report. We have included brief cognitive tests that are not typically administered alone in clinical practice to evaluate a research question on their accuracy for distinguishing CATD from MCI or normal cognition. 3. Full neuropsychological testing is time consuming and has limited availability in many clinical settings. It is an important question how well brief cognitive tests can categorize patients with suspected CATD as there may be circumstances in which clinical decision making can be guided by more quickly available brief testing. 4. Diagnostic accuracy of cognitive tests for distinguishing between groups with smaller clinical differences (e.g., CATD vs. MCI) did appear lower than for distinguishing between groups with larger clinical differences (e.g., CATD vs. normal cognition). The statement about combinations of CSF tests possibly increasing diagnostic accuracy when added to a clinical evaluation vs. the clinical evaluation alone was not based just on the sensitivity and specificity of the combined clinical + CSF testing, but on direct statistical comparisons of the diagnostic accuracy of the combination vs. the clinical evaluation alone. 5. Cognitive measures are not biomarkers. 6. We did not specify which neuropathological criteria must have been used to define AD. We reported whatever neuropathological criteria the authors used. The reviewer correctly points out limitations of the CERAD criteria, including how CERAD criteria may inflate the diagnostic accuracy of amyloid biomarker tests. We recommended that future research investigate how the accuracy of biomarkers for AD and non-AD dementias vary as a function of which neuropathological criteria are used. 7. We reported results as a function of which neuropathologic criteria were used, but did not identify studies that directly compared the diagnostic accuracy of biomarkers as a function of which neuropathologic criteria were used. We revised the report to enhance the discussion of this issue. 8. Wherever available, we reported when studies used the newer neuropathological criteria. We revised the report to more explicitly emphasize the point that future studies should examine diagnostic accuracy of biomarkers compared with the most current neuropathologic criteria.

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Peer Reviewer #5 (TEP) (cont'd)	1. It is not clear why databases were search from inception to 2018 for the cognitive and drug treatment data but relying on systematic reviews for the biomarker studies published before 2012. It would help to understand why systematic reviews were used in this case.	9. Thank you for this important comment. By contrast, the biomarker studies were mostly conducted in patients near their end of life. We revised the report to more clearly point out this limitation in the applicability of the research and the need for studies in patient populations earlier in their disease course, when testing may be more likely to affect disease management and possibly patient outcomes. 10. The apparent discrepancy between the mean ages at imaging or CSF collectionand age at death and the mean interval between imaging or CSF collection and autopsy is a result of these data not all being available from all imaging and CSF studies For the 15 brain imaging studies, as detailed in Table 5.1 of the revised report, mean age at imaging was reported in 5 studies (weighted mean 68 years), mean age at death was reported in 3 studies (weighted mean 78 years), and mean interval between imaging and autopsy was reported in 12 studies (weighted mean 38 months). Similarly, for the 9 CSF studies, as detailed in Table 5.6 of the revised report, mean age at CSF collection was reported in 8 studies (weighted mean 73 years), mean age at death was reported in 3 studies (weighted mean 76 years), and mean interval between CSF collection and autopsy was reported in 6 studies (weighted mean 25 months). Separately, we agree with the reviewer's point about the studies being undertaken at such a late stage of the dementia that distinguishing between dementia etiologies would have limited clinical utility. This timing also likely overestimates the classification accuracy of tests performed at an earlier disease stage. In the revised report, we discussed this issue as a limit in the applicability of study results. 11. We revised the report to correct this inadvertent omission of "clinical."

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Peer Reviewer #5 (TEP) (cont'd)	2. Second, in the case of the neuroimaging and fluid biomarkers, the team allowed studies of various combinations of biomarkers to be considered. However in the case of the cognitive data, the team reported only on single tests. In this regard for the analysis reasonably distinguishes between a)brief cognitive screening instruments (MMSE, MoCA etc.), b) stand alone measures of global cognitive function (ADAS-Cog, DRS) and c) measures commonly used as *part* of a neuropsychological battery. While makes sense to consider the diagnostic utility of the cognitive screeners and possibly theglobal cognitive measures it this doesn't make sense to report individually on the many neuropsychological tests (list and prose learning, naming, executive function) as was done herein. Those tests are just not used in isolation that way so it creates a misleading representation of their diagnostic utility to be presented thusly. This analysis should be dropped orrevised to look at studies of the diagnostic utility of combinations of tests do not appear to have been considered (Edmonds et al 2016, Rabin, et al 2009, Powell, et al, 2006).	

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Peer Reviewer #5 (TEP) (cont'd)	3. The authors or proponents of the analysis argue that one reason to undertake the analysis of cognitive screeners is that neuropsychological evaluations are hard to access. At a minimum this rational should be referenced, but this statement is really indefensible relative to undertaking a subsequent analysis of PET imaging generally and amyloid PET imaging in particular. There are approximately 5,000 neuropsychologists in the U.S. Granted they cluster in and around metropolitan areas and academic medical centers. But so too does PET imaging availability. Moreover most major insurers routinely cover neuropsychological evaluations but not amyloid PET. The statement about limited availability of neuropsychological services should be dropped.	

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Peer Reviewer #5 (TEP) (cont'd)	4. There appears to bias against cognitive tests is evident in the Key Messages, Executive Summary and elsewhere in the text. Contrast the message "cognitive measures …were less accurate in distinguishing between groups with smaller differences in cognitive impairment" with the message 'combinations of CSF biomarkers may…increase diagnostic accuracy when added to clinical evaluation." The former statement appears predicted on SNs and SPs that averaged .76 and .75 respectively for cognitive measures while the later were based on Abeta to tau ratios averaging SN of .79 and SP of .59. Why are comparable findings being represented differently? Wouldn't it be better for the authors to factually describe the SP and SN ranges and avoid editorializing on implications? 5. Caution should be used with the term 'biomarker'. Cognitive measures meet standard definitions of 'biomarker' as clearly as do neuroimaging or fluid assays. So 'biomarker' should be qualified through with document generically with "neuroimaging" or "fluid" or specifically, e.g., "PET" or "CSF". Additional important limitations should be highlighted.	

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Peer Reviewer #5 (TEP) (cont'd)	6. First, it should be noted that CERAD criteria and NIA-Reagan criteria were used as the 'gold standard' in the neuropathology studies. However these are not interchangeable and CERAD criteria do not represent the current consensus regarding methods for neuropathologic diagnosis of AD. CERAD criteria rely on analysis of diffuse amyloid plaques when the prevailing standard now is to examine cored or neuritic plaque density. Moreover CERAD does not adequately use neurofibrillary tangle information in making the determination of AD. For this reason use of CERAD likely inflated the sensitivity of amyloid associated biomarkers. 7. The analysis should examine whether SNs and SPs vary as a function of neuropathology diagnostic system. 8. Finally the limitations section should acknowledge that a newer neuropathology diagnostic scheme for AD was introduced in 2012 (Hyman et al., 2012). 9. If primary care practitioners are seeking to determine the utility of biomarkers in their routine care it is likely at the point of initial diagnosis or soon thereafter. This is when such information could help guide care decisions.	

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Peer Reviewer #5 (TEP) (cont'd)	10. Another limitation needing greater emphasis in this report then is the confusion around time between acquisition of neuroimaging and fluid biomarkers and death in the neuropathology analysis. The average time for diagnosis to death in AD is approximately 10 years. In these analyses mean latency between imaging and death s in the amyloid PET study for example is reported to be 1.1 years. However in the same study mean age at image acquisition is 69.8 years and mean age at death is 80.4 years so I wonder if the 1.1 value listed is a typo. If in fact the intervals are less than 2 years twopossibilities ensue. The first is that these studies were undertaken at such a late stage of the illness that distinguishing between dementia etiologies would have limited clinical utility. Alternately these studies involve persons with AD that died not other causes earlier in the illness, which would diminish the representativeness of these studies. 11. Lastly, stylistically the use of acronym CATD is confusing. While this acronym is defined in the title of the report, it's first use in the text is in the Key Messages where the sentence reads 'cognitive tests for Alzheimer's type dementia (CATD). Thus theacronym appears to be abbreviated 'cognitive tests of Alzheimer's type dementia'. It would help if 'clinical' were reinserted before Alzheimer's in this first instance of the use of the acronym in text.	

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Peer Reviewer #5 (TEP) (cont'd)	References Edmonds, E. C., Delano-Wood, L., Jak, A. J., Galasko, D. R., Salmon, D. P., & Bondi, M. W. (2016). "Missed" mild cognitive impairment: High false-negative error rate based on conventional diagnostic criteria. *Journal of Alzheimer's Disease*, *52*(2), 685-691. Hyman, B. T., Phelps, C. H., Beach, T. G., Bigio, E. H., Cairns, N. J., Carrillo, M. C., ... & Mirra, S. S. (2012). National Institute on Aging–Alzheimer's Association guidelines for the neuropathologic assessment of Alzheimer's disease. *Alzheimer's & dementia*, *8*(1), 1-13. Powell MR, Smith GE, Knopman DS, Parisi JE, Boeve BF, Petersen RC, Ivnik RJ.(2006) Cognitive measures predict Alzheimer's disease pathology. *Archives of Neurology. 63:865-8* Rabin, L. A., Paré, N., Saykin, A. J., Brown, M. J., Wishart, H. A., Flashman, L. A., & Santulli, R. B. (2009). Differential memory test sensitivity for diagnosing amnestic mild cognitive impairment and predicting conversion to Alzheimer's disease. *Aging, Neuropsychology, and Cognition*, *16*(3), 357-376.	

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Peer Reviewer #5 (TEP)	General Comments **1. Note: Using page numbers at Top of page for this review.** 2. Overall: This report topic is highly clinically meaningful. It is of tremendous value to have details of this literature in one place. 3. Overall: "Key Messages" wording should be the same when they are repeated throughout the report in different places. They may be already; I didn't check carefully. Just making the point that consistent wording assists readers' comprehension and retention. 4. Overall: A message I get from this report is that much is still not known about these important questions, and that conclusions must be cautious due to limited amount and quality of research. I think that message should be made clear in key places throughout the report – key messages, abstract, evidence summary, body of report. 5. Overall: When the evidence is clear (adequate amount and quality of research) and a negative effect is found (e.g., no difference, no benefit, actual harm), that should be made clear in Key Messages. That is, "negative" findings are important for the Key Messages and not just "positive" findings. This is just a general comment. 6. Overall: Somewhat excessive use of commas throughout report interferes with easy reading and understanding. Perhaps review proper sentence structure and use commas only when grammatically appropriate.	1. Thank you for your clarification. 2. Thank you for this comment. 3. We agree and attempted to improve the consistency of wording in the revised report. 4. We tried to make it clearer throughout the revised report that much is not well understood about the questions evaluated in this review because of limits in the quantity and quality of available research. 5. We tried to place equal weight on negative and positive findings throughout the report. 6. We reviewed the sentence structure throughout the report and tried to minimize unnecessary commas in the revision. 7. The review was limited to CATD partly to limit the focus to the most typical subgroup of patients with clinical dementia, and partly to limit the quantity of work in a review that already was very large in scope. We revised Table 2.1 on study inclusion criteria to clarify this first rationale in the revision. 8. To clarify the target audiences, we added the following statement to the introduction of the report: "The target audiences of this report are primary care clinicians who diagnose and treat the vast majority of older patients with cognitive disorders, psychologists who may perform additional diagnostic cognitive testing in primary care settings, and dementia specialists who may be most likely to consider biomarker testing for further diagnostic clarification." Thank you for this comment. 9. Thank you for this comment. 10. We revised the key messages to try to avoid potential bias, including negative findings and limitations. 11. Thank you for catching this error. The wording was changed to "conflicting." 12. The structure of the report described is that required by AHRQ. 13. The heading under which this bullet falls is titled: "Accuracy of biomarkers for distinguishing autopsy-confirmed AD from non-AD dementia." These tests are highly sensitive and specific for distinguishing autopsy-confirmed AD from non-AD dementia. 14. We revised the language in the report to refer to "the nutritional drink, Souvenaid" the first time Souvenaid is mentioned in the evidence summary and in the main report. 15. (a) We corrected table 0.1 to correct the sixth column heading to state SP rather than SN. (b) We revised table 0.1 to include definitions for all abbreviations in the footnotes. 16. (a) We revised table 0.2 to include definitions for all abbreviations in the footnotes. 17. (a) We revised table 0.3 to include definitions for all abbreviations in the footnotes.

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Peer Reviewer #5 (TEP) (cont'd)	7. Target population – Not really clear why the review was limited to clinical Alzheimer's type dementia (CATD) only. Can the rationale be more fully explained, e.g., in Chapter 1 Background section? Was it a matter of limited review resources or other reason(s) for limiting to CATD? 8. Intended audience of report – Was it explicitly stated? Who is the intended audience? Early section of report on "Key Messages" has a general statement ("The information in this report is intended to help health care decisionmakers—patients and clinicians, health system leaders, policymakers, and others—make well-informed decisions and thereby improve the quality of health care services."). Evidence Summary page 17 line 9 mentions "Primary care settings…." Chapter 1 Background page 28 line37 mentions "primary care settings." Perhaps intended audience can be more clearly stated or emphasized. 9. Key Questions – appropriate and explicitly stated. 10. Page 2, Key Messages – This section seems somewhat biased toward "positive" findings, even when based on very limited evidence. Perhaps consider also including some of the more negative/cautionary findings, i.e., caveats, from the results. See some examples listed in comments below. 11. Page 5, line 12, Technical Expert Panel – Should it be, "Divergent and conflicting opinions…" rather than "conflicted"?	18. Per AHRQ formatting requirements, the methods of the report are not included in the evidence summary. The main report explains how low and insufficient strength of evidence are defined. 19. In the revised report, CMAI was no longer included in Table 0.5. 20. Thank you for this comment. 21. AHRQ limits the number of bullets and characters in the key messages. However, we revised the key messages to include interpretive points like these where possible. 22. We revised the key messages to include interpretive points like these where possible, but due to space limitation, we addressed this particular point in the discussion. 23. We revised the text to include the missing "was."

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Peer Reviewer #5 (TEP) (cont'd)	12. Page 17-27, Evidence Summary – Shouldn't this lengthy section be included in the Table of Contents that begins on page 8? Somewhat confusing structure of report (Key Messages, Structured Abstract, Table of Contents, Evidence Summary [not in table of contents], Chapter 1 Introduction, etc.). Would be less confusing to include Evidence Summary as an element in the Table of Contents. 13. Page 17, lines 54-55; Page 18, lines 3 –15 "…highly sensitive and specific…" for what? 14. Page 19, line 5 – "Souvenaid" is mentioned for the first time with no explanation of what it is. It is defined on page 144 as a nutritional drink. Suggest you add that descriptor here on page 19 for clarity, i.e., "…and the nutritional drink Souvenaid did not improve…." 15. Page 20, Table 0.1 a. Sixth column should be headed SP (rather than SN repeated from fifth column)? b. Below table, list of abbreviations needs to include TP, TN, FP, FN. 16. Page 21, Table 0.2 a. Below table, list of abbreviations needs to include TP, TN, FP, FN. 17. Page 22, Table 0.3 a. Below table, list of abbreviations needs to include TP, TN, FP, FN. 18. Page 23, Table 0.4 – What is the difference between "low SOE" (low strength of evidence) and "insuf SOE" (insufficient strength of evidence, I presume)? I assume these terms will be explained later in the report. They are not clear in Table 0.4 at this point.	

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Peer Reviewer #5 (TEP) (cont'd)	19. Page 23, Table 0.5 – Below table, list of abbreviations needs to include CMAI. 20. Page 24, Limitations – Well-written. The description of what ISN'T known (no studies reported on…few studies on…etc.) is very sobering. 21. Page 25, lines 7-15, Implications and Conclusions – These sentences seem particularly important and perhaps merit inclusion in the "Key Messages" section of the report on Page 2: a. Brief cognitive tests may help identify which patients with clinically suspected cognitive impairment are more likely to have CATD, but cannot diagnose it. b. Brief test results may help clinicians decide who warrants further diagnostic evaluation, including a detailed history of cognitive symptoms, focused neurological exam, and possible neuropsychological testing and specialty referral. c. These brief cognitive test results also may be sufficient for objectively documenting cognitive impairment in more impaired patients with a recognized history of cognitive and functional decline typical for CATD. 22. Page 25, lines 25-30, Implications and Conclusions – This sentence seems particularly important and perhaps merits inclusion in the "Key Messages" section of the report on Page 2: a. However, even if future research confirms that biomarkers and their combinations improve diagnostic accuracy when added to clinical evaluation, clinical applicability is likely to be limited both because of limited access to such testing in typical clinical settings and the limited efficacy of available treatments for AD and non-AD dementias.	

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Peer Reviewer #5 (TEP) (cont'd)	23. Page 26, line 21 – Word seems to be missing, "…memantine, evidence [was?] insufficient to draw conclusions."	
	Peer Reviewer #6 (TEP)	This report clearly shows the enormous amounts of time, energy, thought, preparation, and care that have gone into reviewing the wide range of very different kinds of evidence. The questions are clearly defined and for generally appropriate. The scope of the work is very broad, ranging from performance of screening tests to biomarkers to the effects of pharmaceutical, supplement and non-pharmacological approaches to care of patients with Alzheimer-type dementia, and settings ranging from primary care to nursing homes to specialized/advanced biomedical research center and sponsored clinical trials. While overall this report is of extremely high quality, I have a few concerns functionality for the wide range of users the authors hope to reach, and how to address the problem of 'insufficient evidence' - especially for clinicians faced with the need to care for patients and families. I would like to see:	1. We revised the report to more clearly define CATD, AD and other related diagnoses. 2. (a) We revised the introduction of the report to clarify the target audience for the different components of the review and how the findings might be pertinent to their decision making. (b) Cognitive screening asymptomatic older adults in primary care was outside the scope of this report. This report sought to address the diagnostic accuracy of brief cognitive tests for distinguishing CATD from MCI or normal cognition in patients in whom there was suspicion for CATD. As stated in #1, we revised the report to more clearly define CATD. 3. We revised the report to clarify that it aimed to examine the accuracy of brief cognitive tests for case finding, or distinguishing CATD from normal cognition or MCI in individuals with suspected CATD. We further clarified that the report was not addressing screening, so that we excluded studies that examined the accuracy of brief cognitive tests for distinguishing unspecified dementia or AD from MCI or normal cognition in asymptomatic older adults in primary care or community settings. We did not change the protocol to include and formally extract and analyze these studies, but we discussed them in the discussion to provide context for the report (which evaluated the diagnostic accuracy of brief cognitive tests for distinguishing CATD from MCI or normal cognition in older adults in whom there was suspicion for CATD.

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Peer Reviewer #6 (TEP) (cont'd)	1. "CATD" is more clearly defined and used with the same meaning throughout the report. As we know, CADT in an epidemiological study (or a clinical trial conducted prior to use of advanced imaging support) is a different entity than AD in an amyloid PET or CSF biomarker study. And even the latter has continued to evolve as apparently new pathological entities with similar clinical features continue to be recognized (just one more in the past 2 months). Though the authors certainly know this, some confusion inevitably creeps in and confounds to some extent the usefulness of the work they have produced. a. Recommendation: Start with a clear definition of each use of a term denoting a neurodegenerative dementia of the Alzheimer type. This was the intent at the outset but got a bit muddy in subsequent parts of the body of the report. All designations of the range of conditions included here should be carefully identified throughout; a table of definitions for each one would be very helpful and could be readily extracted from data from the evidence reviewed for each key question. Also helpful would be a timeline showing the evolution of terms over the period covered by this evidence review.	

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Peer Reviewer #6 (TEP) (cont'd)	2. The intended users of this report are as broad and disparate in their responsibilities as are the key questions. The needs of primary care providers and health care systems are very different from those of specialized diagnosticians and memory care clinicians, basic scientists, biomedical researchers, caregiving researchers, pharmaceutical developers, and imaging tool developers. a. Recommendation: identify (without limiting or excluding specific types of users) different user groups to which each key question is most relevant. Consider structuring a section of the report that helps readers locate what is most likely to be relevant to their interests and needs. b. Recommendation: Acknowledge the evolving nature of the field - e.g. detection of dementia in primary care has become more compelling because population aging, population-health thinking, national concerns about health care utilization and costs, what kinds of care should be valued, and social determinants of health have became important in our policy thinking. These considerations are particularly relevant to primary care and administrative policy decisions about whether (and how) to develop dementia care programs within health care systems and what their content should be.	

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Peer Reviewer #6 (TEP) (cont'd)	For example: cognitive screening tests are most relevant to primary care, where dementia detection is likely to first occur in a mainly non-demented population, yet the special needs and features of primary care are not considered here. Those needs are quite different from the needs of investigators running clinical trials of, say, neuroimaging, biomarkers, or new treatment technologies. Another example: the definitions of AD used in research and now appearing in public media have evolved greatly with the development of technologies over the last 30 years - the span of time covered in this report, at least for some of the questions. Though we think we know that 'Alzheimer's disease is the most common (underlying pathology) in late-life dementia', clinical presentation and neuropathological substrate don't have one to one correlations even in the most expert hands - let alone in general medical settings, where the predominant diagnosis (in Medicare claims data) is Dementia NOS - indicating that non-specialist clinicians are not inclined or don't feel able to make a diagnosis of Alzheimer's disease or another causal pathophysiological process. While I am personally very comfortable making a full differential diagnosis of cognitive impairment in late life, I am not sure what "CATD" actually means in this report - nor do I think generalists will either.	

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Peer Reviewer #6 (TEP) (cont'd)	3. The review of cognitive tests - whether brief screens, multi domain short batteries, or detailed single-domain tests - is, in my opinion, potentially misleading to all but the most sophisticated or specialized research reader. Most studies reviewed used samples with around 50% "CATD" rather than the much lower figures found in formal epidemiological prevalence estimates. Such samples result in inflated test performance and may disadvantage the few studies that used epidemiological sampling methods. It also seems to have resulted in omission of some commonly used tests. For example, the Mini-Cog was tested in a population-based sample with a low (generally accepted) prevalence of dementia, and did quite well in identifying dementia defined by criteria that clearly identified CATD at least as well as many of the studies of other tests that were included. This is discussed further in the next section.	

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Peer Reviewer #7	1. Yes the report is clinically meaningful for much of it. I don't think it makes so much sense clinically or otherwise, however, to focus so much on use of tests to separate MCI from CATD. 2. The key questions were clear. 3. The Executive Summary is one of the most important parts and there were some elements that were less clear. For example,c an it include memantine alone in mod-severe CATD? 4. For Tables 0.1 and 0.2 it would be helpful for the reader to know how much time each test takes. 5. Adding the effect size for table 0.4 would be valuable. 6. It was not clear why SAEs are listed and not common AEs. 7. For Table 0.5, it seems very important to at least mention block box warning for antipsychotics. Leaving the SAE column as NS really sends the wrong message. 8. Future research recommendations are well written and flow from the material presented.	1. Part of our task was to evaluate the diagnostic accuracy of brief cognitive tests for distinguishing between CATD and MCI. 2. Thank you for this comment. 3. We identified no eligible trials that compared memantine with placebo in patients with moderate to severe CATD who were not receiving a cholinesterase inhibitor. We added a footnote to table 0.4 to indicate this. 4. We revised the report so that in the main report when each test is first noted, it is described, including information about time of administration. 5. Table 0.4 was revised to include limited quantitative results to enable easier interpretation compared to including only qualitative statements. 6. The evidence summary aimed to summarize key findings of the report. The report aimed to focus on efficacy and harms outcomes of greatest clinical relevance. We did not evaluate risk of the composite harms outcome of any AE, nor of all specific individual AEs. The list of AEs we evaluated were specified a priori and are listed in the methods section of this review. 7. We revised the report to include commentary about the black box warning about stroke and mortality for use of antipsychotics in patients with dementia. 8. Thank you for this comment.

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Peer Reviewer #8 (TEP)	1. Page 30-32: The Key Questions are appropriate and explicitly stated. They are clear and justified. 2. General: The target audience of this review was not clearly identified in the introduction. There is a brief description on page 3, Line 15-17 in the Executive Summary that is quite broad. I would like to see this explicitly stated in the Introduction section. 3. General: This report is clinically meaningful. It asks and attempts to answer clinically significant questions facing primary care doctors every day.	1. Thank you for this comment. 2. We revised the introduction of the report to clarify the target audience as follows: "The target audiences of this report are primary care clinicians who diagnose and treat the vast majority of older patients with cognitive disorders, psychologists who may perform additional diagnostic cognitive testing in primary care settings, and dementia specialists who may be most likely to consider biomarker testing for further diagnostic clarification." 3. Thank you for this comment.
	Peer Reviewer #9	This is a well written report that covers areas that span from diagnostic accuracy of cognitive tests, biomarkers and the efficacy if therapy. The results are summarized in clear fashion and appropriate conclusions have been drawn from the evidence reviewed. The elements addressed are of clinical value and all the questions are clearly articulated. However, I suggest that authors clearly state in their structured abstract and executive summary who is the target audience for this report. Not clearly articulated.	We revised the evidence summary and main report introduction to indicate the target audiences for the report. However, due to strict word count limitations, we could not add this information to the structure abstract.
	Public Reviewer #4 (AMDA)	I read the summary statements and note the level of detail that followed. The summary statements are consistent with my understanding generally and give me some specifics I was less certain of previously. I agree with Gary, the AHRQ group seems to continue to do excellent work.	Thank you for this comment.

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Public Reviewer #4 (AMDA)	This is 200+ pages. Impossible to perform an in depth review. It is incredibly detailed. It seems as if they did a very comprehensive job. It is very important to get this information out so practitioners/public do not base conclusions on one or two studies. Unfortunately, even after combining multiple studies using the rigorous criteria (that I assume was appropriate criteria for each section to select the studies they included) that they employed, it leaves questions unanswered today. On the other hand that they found no strong evidence for use of medications for tx of BPSD may mean that there is no medication, even if we wish there was.	Thank you for this comment.
	Public Reviewer #4 (AMDA)	This is a very comprehensive systematic review. It highlights the importance and rationale (and the integration of Cognitive tests, Neuroimaging studies and CSF and blood biomarkers in the accurate diagnosis of Alzheimer's disease. It also confirms the current accepted standard of treatment for mild to moderate dementia with Cholinesterase inhibitors and for moderate to severe dementia with Memantine. It still leaves us in the dark on how to best treat and manage the Behavioral and Psychological Symptoms of Dementia using available prescription drugs.	Thank you for this comment.

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Public Reviewer #3 (American Geriatrics Society)	1. **Study Durations.** Consider clarifying the 2-week minimum study duration for agitation, aggression, psychosis, and disinhibition but 24 weeks for other outcomes. A rationale behind the delineation may be helpful. 2. **Screening.** AGS agrees that brief screens can be useful to early detection and, although it may not allow for improved outcomes or change in disease progression, earlier diagnosis is still better if it allows for appropriate referral to geriatrics and other specialty support services. Consider noting this next step to early detection in order to provide appropriate patient care moving forward. 3. **Imaging.** Although PET & FDG-PET increases diagnostic accuracy, these scans are not widely available and likely to not affect clinic treatment of dementia symptoms since no disease altering treatments currently exist. AGS wishes AHRQ to consider noting these limitations in the report. 4. **Biomarkers.** Similar to the imaging section above, markers are not readily available and may not be covered by insurance companies for patients. This reality could limit their usefulness in clinical practice. Additional details should be included on which markers are indicated and how they will be used and ordered in non-invasive ways. These considerations should also be made when extrapolating the recommendations to wider clinical use for diagnosis.	**1.** We revised the report in Table 2.1 Inclusion Criteria to clarify the rationale behind the minimum study durations for the different treatment outcomes. **2.** We revised the report to clarify that diagnostic accuracy is relevant if it leads to changes in patient or caregiver outcomes. **3.** We agree with the reviewer. The report already stated these limitations. **4.** The scope of this report extends to defining the accuracy of the different biomarkers for distinguishing autopsy-confirmed AD from non-AD dementia. We clarified that biomarker tests have limited availability in many clinical settings and that this circumstance and the absence of data linking these tests to patient or caregiver outcomes limits their applicability in typical clinical settings. **5.** AGS is correct that these drugs are not disease modifying. Though the question of stopping them when appropriate towards the end of life is an important one, it was not specifically included in the scope of this review. We revised the applicability section of the report to note that the review did not address this clinical question.

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Public Reviewer #3 (American Geriatrics Society) (cont'd)	**5. Treatment.** AGS wishes to note that cholinesterase & NMDA medications in some dementia are not disease modifying and care should be given to stopping these medications when appropriate towards the end of life. Additionally, these types of drugs are sometimes stopped during care transitions, leading to a question of their effectiveness.	
	Public Reviewer #4 (AMDA)	I don't have any comments to make on this review.	Thank you for this comment.

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Public Reviewer #2 (American Psychological Association)	This review descriptively summarizes the existing findings on using cognitive markers in helping with diagnosing clinical Alzheimer's type dementia, and using imaging and biological markers to help with identifying whether the dementia is of AD etiology. The concept and research questions are generally clear. However, there were some concerns regarding the conclusions and review methodology, and points that were insufficiently addressed. **1. The review appears to be somewhat biased from the beginning against cognitive/neuropsychological testing, given that:** A. It is stated that the analysis of cognitive screeners was justified because access to neuropsychological evaluation is poor. Access to neuropsychological evaluation may vary based on locale and other factors but PET scanning (including amyloid PET) is itself quite difficult to access. Access to neuropsychological evaluations is likely more feasible, affordable, and can yield comprehensive information, moreso than other proposed biomarkers.	1a. As the reviewer acknowledges, access to comprehensive neuropsychological may be limited. Our report also noted that access to amyloid PET and other brain imaging measurements as implemented in included studies may be limited. 1b. The sensitivity and specificity of brief cognitive tests were for distinguishing CATD from MCI or normal cognition when using a reference standard of full clinical evaluation. The sensitivity and specificity of biomarkers (brain imaging and CSF tests) were for distinguishing AD from non-AD dementias when using a reference standard of neuropathologically confirmed AD. Given these differences, it is not appropriate to directly compare the diagnostic accuracy of brief cognitive tests with that of the biomarkers and to infer bias. 2. We understand the reviewer concern with analyses evaluating the ability of tests normally administered as part of a cognitive battery for distinguishing CATD and either MCI or normal cognition. In the revised report, we tried to clarify that these tests are not normally administered as stand-alone tests. We also tried to increase the emphasis on a point we already made in the draft report that these brief cognitive tests are not sufficient by themselves for diagnosing CATD in individuals in clinical settings.

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Public Reviewer #2 (American Psychological Association) (cont'd)	B. The classification accuracy statistics cited for cognitive testing and biomarkers are more comparable than the conclusions suggest. Note that the sensitivity statistics behind the highlighted statements are nearly identical and the specificities are better for the cognitive tests: *Many brief cognitive tests had high sensitivity and specificity (>0.8) for distinguishing CATD from normal cognition, but were less accurate distinguishing between groups with smaller differences in cognitive impairment.* versus *Individual CSF biomarkers and ratios were moderately sensitive and specific for autopsy-confirmed AD, but combinations of CSF biomarkers may have the highest combination of sensitivity and specificity, and may increase diagnostic accuracy when added to clinical evaluation*.	3a. Any style of "trails" testing would have been eligible for inclusion, including CTT and others. We did not limit eligibility to the traditional Trail Making Test (e.g. the DKEFs trails subtests were also eligible). Still, many tests with a known literature base did not make it into our review. The most common reasons for exclusion (not in any particular order) were non-English test administration, lack of a CATD diagnosis group (e.g., the study only included generic cognitive impairment, dementia, mixed samples, etc), lack of a gold standard diagnostic evaluation for comparison, and methods not allowing for the calculation of sensitivity/specificity. TMT interference score also was eligible. 3b and 3c. In order to meet the nominator objectives focusing on brief testing to identify CATD, the scope of our review only included brief cognitive tests commonly used for screening, brief batteries, and domain-specific tests for memory, and selected executive and language tests. There are many tests that tap more than one conceptual ability. If a test was thought to be commonly used, had otherwise eligible studies available, and can be conceptualized in multiple ways, we opted to be inclusive. While we agree that the Digit Symbol (and Symbol Digit Modalities) task evaluates processing speed, asserting that they are exclusively assessing processing speed and without executive contribution would have excluded them from the review. We included verbal fluency tests under the broader category of language tests. 3d. Any brief battery summary score was eligible for inclusion, but not any single domain level test. Still, many individual tests that would have been eligible under domain categories did not produce any eligible studies. We were also surprised to see how little eligible data (for our specific question) were available on many commonly used tests. We explicitly commented on this in regards to common screening instruments, but could perhaps make the same comment in regards to the domain category tests more often used as part of a full battery. The degree to which the review provides even coverage over different types of testing is limited both by scope (determined in consultation with the AHRQ, the nominator, and consulting topic experts) and available literature meeting eligibility criteria.

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Public Reviewer #2 (American Psychological Association) (cont'd)	2. The inclusion of standalone tests in this review is questionable, as rarely are these tests used as individual assays of pathology, thereby making them (and their classification accuracy statistics) incomparable to the disease-specific biomarkers reviewed here. While analyzing cognitive screeners (MoCA and MMSE) and global tests (ADAS-Cog, DRS) is reasonable it doesn't make sense that they also analyzed tests that are not meant as stand alones (Boston, Trails B, etc.). It is misleading to publish sensitivities and specificities of these tests used in isolation. Using a single test for diagnostic accuracy can also be problematic. The NIAA all-cause dementia criteria require impairments in at least two cognitive or behavioral domains - something which cannot be concluded from a single test.	4. We agree with this reviewer comment that the limited number of studies, patients and data examining questions of interest limited the conclusions we could draw about the relative accuracy of tests or combinations of tests, or whether patient-specific characteristics may modify the efficacy of the tests. We tried to make this conclusion clearer in the revised report. 5. We did not limit inclusion to studies that used 2004 criteria for MCI. Studies using NIA-AA criteria also were eligible. We reported whatever MCI criteria studies used. We revised the report to point out that it is important to consider which MCI criteria were used when interpreting results on the diagnostic accuracy of brief cognitive tests for distinguishing between CATD and MCI. 6. In the revised report, we tried more clearly communicate the criteria studies used for CATD and MCI and to discuss the potential implications of using different criteria for interpreting results. 7. The purpose of the review was to evaluate the accuracy of brief cognitive tests for classifying individuals with suspected CATD between CATD, MCI and normal cognition when compared to a reference standard of a full clinical evaluation. The assumption is that the full clinical evaluation included an assessment of function to distinguish between CATD and MCI. 8a. We revised the report to include discussion of possible relevance of biomarker testing early in the CATD disease course, the lack of eligible studies with early assessment of biomarkers and a neuropathological reference group, and the uncertain applicability of the published study results to patients early in their disease course. 8b. We did not identify any eligible studies with low or moderate risk of bias that assessed the diagnostic accuracy of hippocampal atrophy with neuropathologically defined AD. We noted this as a limitation in the revised report and recommended that future studies examine this association. We also recommended that other brain imaging, CSF and blood tests be evaluated. 8c. In the revised report, we added the following comment to the start of the limitations section: "The first limitation of this review on brief cognitive tests and biomarker tests is that it did not identify eligible studies that connected test accuracy to patient or caregiver outcomes, including cognitive, functional, psychological, quality of life and others. Moreover, eligible studies did not report on the association of these tests with process outcomes like changes in pharmacological or nonpharmacological management, including lifestyle changes or changes in life planning that may or may not affect patient or caregiver outcomes."

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Public Reviewer #2 (American Psychological Association) (cont'd)	3. There were also some questions and concerns regarding the appropriateness of tests/scores reviewed and the cognitive domains to which they were assigned. a. There are only 2 studies on TMT. Should the data be supplemented by CTT studies? Furthermore, should TMT interference score be considered together with the TMT-B score? b. Digit symbol should be a task of processing speed rather than executive function. Indeed, there should be a separate category for processing speed. c. Phonemic fluency (letter) is a test of fluency of verbal concepts. It is a prefrontal measure, in addition to being a language test. d. Working memory is an important component for executive function but it was not given much attention in this report, apart from it being merged together with learning memory for a composite score in the Cogstat brief battery.	

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Public Reviewer #2 (American Psychological Association) (cont'd)	**The review methodology may be improved in the following ways:** 4. The number of studies for individual test measures is rather limited (1 or 2), particularly for the imaging measures. Thus, the authors cannot answer the question they set out to answer, such as the relatively efficacy of tests or combination of tests, or whether patient-specific characteristics may moderate the efficacy of the tests. It is also hard to compare the results for different tests since they derived from different numbers of studies.	

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Public Reviewer #2 (American Psychological Association) (cont'd)	5. It is unclear why the mild cognitive impairment (MCI) studies were limited to those using the 2004 criteria. Studies using the NIAA MCI due to AD criteria can possibly be included as well. One of the limitations in determining the diagnostic value of screening measures is the assumption of studies that their sample of MCI patients actually have AD. MCI is heterogeneous and reflects many potential etiologies including other neurodegenerative disorders, vascular diseases, mood disorders, substance abuse, and medical comorbidities. Therefore, when we try to determine if a screening measure is able to detect AD, we need to be sure that the MCI patients have this pathology. We think studies on screening measures should ensure that the MCI patients have prodromal AD based on CSF biomarker evidence in order to investigate their diagnostic utility. In addition, screening measures should be developed that take advantage of the existing literature showing the sensitivity of certain cognitive tests to prodromal AD such as facial recognition and eye tracking. Continued use of screening measures that rely on procedures such as paragraph recall that have been found not to correlate with CSF biomarkers or medial temporal lobe atrophy are not going to be effective in detecting prodromal AD.	

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Public Reviewer #2 (American Psychological Association) (cont'd)	6. It may help the readers to list the criteria for Clinical Alzheimer's-type Dementia (CATD) as well as provide a specific description of the MCI populations. MCI populations can refer to those with amnestic MCI, non-amnestic MCI or a mix of both subtypes. In regards to research on CATD, aMCI research should carry a greater weight as opposed to naMCI or studies with mixed aMCI/naMCI samples. **Important points that were insufficiently addressed in this review:** Assessments of activities of daily living were not reviewed, which is a major omission given how critical these instruments are to determining an MCI diagnosis. Even though patients with MCI have "relatively preserved" instrumental activities of daily living, there is evidence that even in this pre-dementia stage, individuals are beginning to show subtle functional declines in instrumental activities. Cognitive screening instruments should incorporate functional assessment measures in order to be more sensitive in distinguishing normal cognition from MCI.	

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Public Reviewer #2 (American Psychological Association) (cont'd)	7. A summary/discussion of broader issues would be helpful, such as the following: a. *Hypothesized usefulness of biomarkers earlier in the disease course.* The authors only reviewed the efficacy of brief cognitive tests in diagnosing clinical AD type dementia, but not biological or imaging measures. While practical reasons may demand a brief evaluative assessment of cognitive impairment in older adults, it is possible that certain aspects of early-state neurodegeneration may only be discovered on a neural or biological basis. b. *Availability of other corollary biomarkers, specifically examining hippocampal atrophy, in addition to medial temporal lobe atrophy*. c. *Overall benefits of early detection/diagnosis to the patient irrespective of the assays (cognitive or biological) used*. While it is true that there is no cure for AD, there are many advantages to early detection and diagnosis, and thus the potential harms may be outweighed. These include the opportunity of individuals to participate in treatment trials, make lifestyle changes, and have information needed to make informed decisions and to prepare for the future.	

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Public Reviewer #2 (American Psychological Association)	1) The report does not cover behavioral interventions that may be effective with respect to treating behavioral and psychological symptoms of dementia (BPSD). STAR-VA (https://doi.org/10.1093/geroni/igy023.142) and Peaceful Mind (https://doi.org/10.1177/0145445513477420) are two examples. The report states "We are not aware of a recent review on the effect of nondrug interventions for treating cognition, function, and QOL in patients with established CATD," such a review, along with the review provided in the current report, would provide the most useful information to practitioners. Since adverse events were specifically mentioned as a limitation of prescription drug treatments, it is not clear why behavioral interventions were not given more consideration. 2) [see under Introduction] 3) Only 15 unique studies, with only 3 with low risk of bias, are analyzed with respect to biomarker tests. This is compared to 42 low-to-medium risk of bias studies for cognitive tests. Differences in the amount and quality of information for these approaches might be better communicated to assist practitioners. There are obviously greater costs and risks associated with biomarker tests, and data are not as easily interpreted. Combined with the relatively small literature of support, it seems that greater caution might be warranted in recommending them as a diagnostic tool.	1. The efficacy and harms of behavioral interventions was outside the scope of this review. This topic is the focus of a separate ongoing AHRQ review. 2. Refer to response under introduction. 3. We revised the report to try to ensure that our interpretation of study results was appropriately cautious based on the data available. For example, we added language in the key messages, evidence summary and discussion that biomarker data were limited and also highlighted the small number of studies. 4. We revised the report to clarify that brief diagnostic tests alone are not sufficient for diagnosing CATD in individuals and that the studies and our analyses evaluated the sensitivity and specificity of these tests for distinguishing CATD from MCI and normal cognition in groups of study participants. We also changed the wording in the report referring to the cognitive test accuracy away from "diagnostic accuracy" to "classification accuracy" or to identify CATD. 5. In the revised report, we specified in the Applicability section of the Discussion that by intent this review did not evaluate the role of formal neuropsychological testing for clinical diagnosis of CATD. 6. In the revised report, we were more cautious (appropriately) in drawing conclusions about the possible variability in accuracy of brief diagnostic tests by patient characteristics. We also tried to emphasize the importance of future research evaluating the diagnostic accuracy of these tests in different patient groups, including by race/ethnicity and others.

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Public Reviewer #2 (American Psychological Association) (cont'd)	4) Although brief and accurate diagnostic tests are an important goal, it should be emphasized that these tests alone are insufficient for assessing for dementia. According to APA Guidelines for the Evaluation of Dementia and Age-Related Cognitive Change (currently being updated, https://www.apa.org/practice/guidelines/dementia): "Although objective neuropsychological testing provides valuable data for diagnostic purposes, the clinical interview remains an essential element of an in-depth assessment for dementia (ABA & APA, 2008; Mackinnon & Mulligan, 1998; National Center for Cost Containment, 1997). Obtaining contextual and historical information from interviewing knowledgeable informants improves diagnostic accuracy and may be less likely to be biased by sex and gender, education, or ethnicity in comparison to performance-based measures (Galvin et al, 2005; Monnot, Brosey, & Ross, 2005). Interview data from a corroborative source such as a caregiver or knowledgeable family member can provide information on everyday cognitive functioning (Waite et al, 1998). An advantage of informant history is the ability to assess change in performance from earlier in life which also potentially reduces test bias (Jorm, 1996). Finally, obtaining data from informant interviews can add greater precision in the design of appropriate behavioral, environmental and pharmacological treatments of dementia (Waite et al, 1998; Hartman-Stein, Reuter, & Schuster, 2002)."	

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Public Reviewer #2 (American Psychological Association) (cont'd)	5) We understand the focus on brief measures (e.g., MMSE, Clock Drawing) as these are quick screen that are easy to administer, but there is a concern that this report may minimize the importance of, and detailed clinical information, that can be derived from larger assessment batteries (e.g., Halstead-Reitan Neuropsychological Battery, Luria-Nebraska Neuropsychological Battery).	

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Public Reviewer #2 (American Psychological Association) (cont'd)	6) The report states "the small set of studies that evaluated the diagnostic accuracy of brief cognitive tests by patient characteristics reported that accuracy did not vary by age, sex, race/ethnicity, or education. These studies were likely too small to rule out such associations. However, these findings suggest it may not be necessary in a primary care setting to use different cut points in different populations as defined by these characteristics." This statement may undercut the importance of research in this area. The review should better acknowledge the limitations of cognitive tests in ethnic and racial minority group and in non-native English speakers and mention some of the existing findings. The report might provide stronger language related to need to develop culturally and linguistically sensitive tests and to validate existing tests in diverse groups. This is increasingly important as the diversity of the older adult population in the U.S. grows (for discussion and examples, see: Goldstein, F. C., Ashley, A. V., Miller, E., Alexeeva, O., Zanders, L., & King, V. (2014). Validity of the montreal cognitive assessment as a screen for mild cognitive impairment and dementia in African Americans. J Geriatr Psychiatry Neurol, 27(3), 199-203. (3), 241-250.) doi:10.1097/JGP.0b013e31825d0935; Moon, H., Badana, A. N., Hwang, S. Y., Sears, J. S., & Haley, W. E. (2019). Dementia Prevalence in Older Adults: Variation by Race/Ethnicity and Immigrant Status. The American Journal of Geriatric Psychiatry, 27(3), 241-250.)	

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Public Reviewer #1 (Anonymous)	See above - - I think there's a lot of good exhaustive work that's gone into this. But the report has shortcomings. It's also extremely long and even when you want to check out a reference - if you go directly to the journal and you don't subscribe it's costly. So I worry that many people will not read the report carefully and only draw out findings that support their point of view - - to me the report should be more analytic and less just descriptive and formulaic using EPC standards for reporting. Try to link the material you publish with more general understanding of the disease (diseases) being considered and the nature of aging and progressive diseases of aging.	In revising the report, we shortened it, eliminated substantial redundancy between the tables and text by using the tables for descriptive purposes, and limiting the text to analytic purposes as much as possible. We also tried to enhance the contextual information in the report.

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Public Reviewer #5 (ACT-AD)	May 27, 2019 Dear Mr. Khanna, Dr. Bierman, Dr. Chang, Ms. Wittenberg, The coalition to Accelerate Cure/Treatments for Alzheimer's Disease (ACT-AD) is comprised of more than 50 national organizations representing patients, caregivers, researchers, health professionals, and other health advocates. Our mission is to support efforts to expedite the development, review, and approval of transformational therapies for Alzheimer's disease. ACT-AD appreciates the opportunity to comment on the Agency for Healthcare Research and Quality (AHRQ) draft comparative effectiveness review, *Diagnosis, and Treatment of Clinical Alzheimer's-type Dementia (CATD): A Systematic Review*. A formal review of available evidence on diagnostics and treatments of Alzheimer's disease is helpful in identifying current gaps in research and literature. The below feedback bullets constitute some feedback to consider as you finalize the document.	1. The IDEAS study provided information about the availability of amyloid PET in general clinical settings and with reading and analysis methods more feasible for typical clinical settings. It reported results for changes in diagnosis and changes in clinical management, but for the latter also showed the frequency with which management with dementia medications was not evidence-based, either before or after amyloid PET imaging. The IDEAS study was not eligible for our review because amyloid PET results were not compared to a neuropathological AD reference, but its findings were highly relevant to the discussion and were incorporated into the revised report. 2 and 3. Research on the benefits and harms of cognitive screening in primary care settings and to "Improve uptake" of such screening was outside the scope of this report. 4. We revised the report to provide more context around the issue of treatment of BPSD in patients with dementia, including the issues raised by this reviewer. 5. Impact on change in residence to a different level of independence (e.g., placement in a nursing home from independent living) was a prespecified treatment outcome for our report. We found minimal data reported on this outcome. None of the three studies suggested by the reviewer were judged eligible for our review. The Geldmacher study compared participants with long-term open label adherence to donepezil vs. those without similar open-label adherence, with all participants having previously participated in one of three 12-24 week donepezil vs. placebo RCTs. They reported longer time to first dementia-related nursing home placement and longer time to permanent nursing home placement in the adherent open-label donepezil group. It is not possible to confidently attribute the reported between-group differences in these outcomes to the pharmacological effects of donepezil. This study was not included because we excluded nontrials from our analyses of treatment efficacy. The Knopman study pertained to Tacrine, which no longer is FDA approved, and so was not eligible for inclusion in our review. The Fillit study was not eligible for our review because it was not an RCT, but instead was an expert consensus paper on treatment recommendations.

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Public Reviewer #5 (ACT-AD) (cont'd)	1. We recommend the final document reference the Imaging-Dementia – Evidence for Amyloid Scanning (IDEAS) Study. IDEAS is a longitudinal cohort study examining data from more than 11,000 Medicare patients with Alzheimer's disease, mild cognitive impairment, and uncertain etiology to assess the impact of amyloid PET on patient outcomes. The results of the first phase of the study were published in April 2019, which found that PET scans that identify amyloid plaques in the brain changed the medical management in more than 60 percent of patients. For more information on the study visit, https://www.ideas-study.org/.	

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Public Reviewer #5 (ACT-AD) (cont'd)	2. There is a need to promote cognitive assessments in primary care settings for older adults. A recent survey by the Alzheimer's Association found that despite the widespread belief among primary care physicians and older adults that brief cognitive assessments are beneficial, only half of older adults are being assessed1. Research on how to improve the uptake of brief cognitive assessments by primary care physicians and the benefits of screening should be incorporated into the final review. 3. In 2013, the Alzheimer's Foundation of America and the Alzheimer's Drug Discovery Foundation convened a working group for the review of the evidence for screening for dementia. Following the review of the available evidence, the working group released ten recommendations for the early detection of Alzheimer's disease to improve clinical care and management2. The final report could reference some of these recommendations for the improvement of screening. 4. The Alliance for Aging Research and the University of California San Diego School of Medicine published a paper in the *American Journal of Geriatric Psychiatric* which underscored the high unmet medical need for behavioral and neuropsychiatric symptoms for Alzheimer's disease3.	

Source: https://effectivehealthcare.ahrq.gov/products/alzheimers-type-dementia/research
Published Online: April 28, 2020

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Public Reviewer #5 (ACT-AD) (cont'd)	We suggest that the behavioral and psychological symptoms of dementia (BPSD) section of the final paper include more information about the lack of data-driven information on which to base decisions regarding improvements to dementia care—such as when to use or reduce antipsychotic use, the best potential treatments for agitation in dementia, and the benefits and drawbacks associated with bifurcating treatment for psychosis and agitation. For example, in April 2005, the FDA issued a "black-box" warning for atypical antipsychotics in the treatment of NPS in older patients with dementia because of a 1.6- to 1.7-fold higher death rate in those taking such drugs compared with those taking placebo. However, a large longitudinal observational study published in the September 2013 issue of the American Journal of Psychiatry challenged these findings by showing that the primary correlate of negative outcomes was the psychiatric symptomatology and not the drugs used to treat these symptoms. Additionally, the American Psychiatric Association has longstanding practice guidelines that address the treatment of psychosis (including pharmacologic) in patients with dementia,4 and more recently, the results of the PRISM II study found that dextromethorphan/quinidine is shown to be an effective and well-tolerated treatment for PBA secondary to dementia.	

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Public Reviewer #5 (ACT-AD) (cont'd)	5. There have been some studies conducted that have suggested that cholinesterase inhibitors slow the rate of Alzheimer's disease progression and delay nursing home placement6,7,8. They should be reviewed and considered before the release of the final review. Thank you for continuing to pursue research that identified the gaps in our knowledge of Alzheimer's disease. If you have any questions about our recommendations, please do not hesitate to contact our organization. Inquires can be directed to Ryne Carney at (202) 293-2856 or rcarney@agingresearch.org. Sincerely, Missy Jenkins Accelerate Cure/Treatments for Alzheimer's Disease (ACT-AD) Vice President of Public Policy Alliance for Aging Research	

Section	Reviewer & Affiliation	Comment	Response
General (cont'd)	Public Reviewer #5 (ACT-AD) (cont'd)	1 Alzheimer's Association. 2019 Alzheimer's Disease Facts and Figures. Alzheimers Dement 2019;15(3):321-87. 2 Borson S, Frank L, Bayley PJ, et al. Improving dementia care: the role of screening and detection of cognitive impairment. Alzheimers Dement. 2013;9(2):151–159. doi:10.1016/j.jalz.2012.08.008 3 Promoting Wellness in Older Adults with Mental Illnesses and Substance Use Disorders: Call to Action to All Stakeholders Jeste, Dilip V. et al. The American Journal of Geriatric Psychiatry , Volume 26 , Issue 6 , 617 - 630 4 Lopez OL, Becker JT, Chang YF, et al. The long-term effects of conventional and atypical antipsychotics in patients with probable Alzheimer's disease. Am J Psychiatry 2013;170:1051-8. 5 Hammond et al. BMC Neurology (2016) 16:89 DOI 10.1186/s12883-016-0609-0 6 Geldmacher DS, Provenzano G, McRae T, et al. Donepezil is associated with delayed nursing home placement in patients with Alzheimer's disease. J Am Geriatr Soc. 2003 Jul;51(7):937-44. 7 Knopman D, Schneider L, Davis K, et al. Long-term tacrine (Cognex) treatment: effects on nursing home placement and mortality, Tacrine Study Group. Neurology. 1996 Jul;47(1):166-77. 8 Fillit HM, Doody RS, Binaso K, et al. Recommendations for best practices in the treatment of Alzheimer's disease in managed care. Am J Geriatr Pharmacother. 2006;4 Suppl A:S9–S24.	

Section	Reviewer & Affiliation	Comment	Response
Clarity and Usability	Peer Reviewer #2	Because of the granularity of the data and the amount of data (including commenting that data were insufficient or low quality), I felt that the report was difficult to navigate. Here are some suggestions to help with this. 1. From a clinicians perspective, I think that some summary tables that are simpler (even simpler than those in the ES), though less precise would be valuable. A slide deck of these to accompany the report would be very helpful. 2. In the Evidence Summary, it would be valuable to give page numbers where the data supporting each of the main points can be found.	1. In the revised report, we eliminated redundant text and relied on tables as much as possible. 2. The organization of the main report parallels that of the evidence summary and includes a detailed table of contents to assist with navigation.
	Peer Reviewer #3 (TEP)	Yes, with the exception of the later chapters noted above	Thank you for this comment.

Section	Reviewer & Affiliation	Comment	Response
Clarity and Usability (cont'd)	Peer Reviewer #4 (TEP)	1. The report is well outlined and organized and understandable with repeating framework that makes sure the reader can identify repeating topics. 2. The conclusions in most sections are relevant and can be employed in practice with two exceptions in my view. (a) The first is the failure to recognize the NINCDS-ADRDA have been replaced by the NIA-AA, and including studies that used that criteria should have been done. This will limit the usability of this document. (b) The other stand out is recent research on amyloid imaging, and the IDEAS Study report in JAMA 2019. This report is relevant to all the imaging sections in the report and again will lessen the utility and in fact makes some sections inaccurate.	1. Thank you for this comment. 2a. We included studies that used the NIA-AA criteria. Our inclusion criterion for the biomarker studies was that they had to have compared biomarker diagnostic accuracy to neuropathological AD criteria. Studies used different neuropathological criteria over time and we did not include only certain neuropathological criteria (e.g., NINCDS-ADRDA) and exclude others (e.g., NIA-AA). All were included. Where data were available, we reported how results appeared to differ as a function of which neuropathological criteria were used. 2b. The IDEAS study provided information about the availability of amyloid PET in general clinical settings and with reading and analysis methods more feasible for typical clinical settings. It reported results for changes in diagnosis and changes in clinical management, but for the latter also showed the frequency with which management with dementia medications was not evidence-based, either before or after amyloid PET imaging. The IDEAS study was not eligible for our review because amyloid PET results were not compared to a neuropathological AD reference, but its findings were highly relevant to the discussion and were incorporated into the revised report.

Section	Reviewer & Affiliation	Comment	Response
Clarity and Usability (cont'd)	Peer Reviewer #5 (TEP)	1. Report structure/organization somewhat difficult to follow. See comments above for examples. 2. It is very difficult for a non-technical expert to wade through the highly detailed and contradictory findings of the individual studies that were reviewed. Therefore, the Key Messages repeated throughout the report and the final Discussion chapter's Overview, including clearly stated limitations of the evidence, are critical for the average reader of this report (e.g., clinicians, policy makers). 3. Difficult to know how a clinician, especially in primary care (is that the intended audience?), should use this report.	1. We tried to simplify the number of layers of subheadings and the content of the revised report to make it easier to follow. 2. The reviewer points out why we included the key messages and summarized the main findings, including interpretation of the results and limitations. 3. In the revised report, we tried to clarify who our target audiences were and frame the findings in terms of how they may help inform decisions by those audiences.
	Peer Reviewer #6 (TEP)	I would like to see broader topic headings and tighter links between the key questions and (1) the different target audiences (among the several intended here) (2) the settings in and for which each key question is most relevant	In the evidence summary and introduction of the revised report, we specified the target audiences for the different key questions and the clinical settings most relevant for each key question.
	Peer Reviewer #7	The overall rationale for updating he 2008 AAFP/ACP makes sense but unfortunately, there have been no new FDA approved drugs so then much of this revolves around "off label" treatments or further developments for already approved drugs. The low SOE for so many large trials is a bit surprising and makes one wonder what can really be taken away from them.	1. Thank you for this comment. We agree that it is unfortunate that there no new FDA approved drugs for treatment of CATD. 2. We agree that the low strength of evidence limits the conclusions that can be drawn.

Section	Reviewer & Affiliation	Comment	Response
Clarity and Usability (cont'd)	Peer Reviewer #8 (TEP)	The report is well structured and the Key Messages allow for the main points to be quickly and easily identified. I often found the tables to be difficult to read and included too much dense text. The conclusion are very relevant to policy makers, primary care physicians, and organizations looking to make guidelines.	1. We attempted to reformat the tables to make them easier to read in the revised report. 2. Thank you for this comment.
	Peer Reviewer #9	yes	Thank you for this comment.