Comparative Effectiveness Review Disposition of Comments Report

Research Review Title:
Treatment for Bipolar Disorder in Adults: A Systematic Review

Draft review available for public comment from November 1, 2017 to November 22, 2017.

Research Review Citation: Butler M, Urosevic S, Desai P, Sponheim SR, Popp J, Nelson VA, Thao V, Sunderlin B. Treatment for Bipolar Disorder in Adults: A Systematic Review. Comparative Effectiveness Review No. 208. (Prepared by the Minnesota Evidence-based Practice Center under Contract No. 290-2012-00016-I.) AHRQ Publication No. 18-EHC012-EF. Rockville, MD: Agency for Healthcare Research and Quality; August 2018. Posted final reports are located on the Effective Health Care Program search page.
DOI: https://doi.org/10.23970/AHRQEPCCER208.

Response to Peer and Public Comments on this Research Review

The Evidence-based Practice Center (EPC) Program encourages the public to participate in the development of its research projects. A draft form of each research review is posted to the AHRQ Web site for public comment. Comments can be submitted via the Web site, mail or email. At the conclusion of the 3-4-week public comment period, authors use these comments to revise the draft research review.

In addition to public comments, each draft research review is independently evaluated by peer reviewers before it is finalized. Because they are chosen for their expertise in the subject matter and research methods, and freedom from conflict of interest, peer reviewers help to assure that the final report is accurate and free from bias.

The table below includes the original comments by peer reviewers and the public, as well as the authors' response for each comment that was submitted for the draft research review. Comments are not edited for spelling, grammar, or other content errors. Each public comment is listed with the name and affiliation of the commentator, if this information is provided. Peer reviewers are listed by number. 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 AHRQ.

Archived: This report is greater than 3 years old. Findings may be used for research purposes, but should not be considered current.

Source: https://effectivehealthcare.ahrq.gov/topics/bipolar-disorder-treatment/final-report-2018
Published Online: August 7, 2018

Section	Reviewer, Affiliation	Comment	Response
Introduction	**TEP Reviewer #1**	Clearly written.	Thank you.
Introduction	**TEP Reviewer #3**	The text in the introduction would benefit from review and careful editing. Specific comments are included in the attached pdf file.	Thank you for the comments. The review has been revised in response to comments and subsequent editing.
Introduction	**TEP Reviewer #3**	It is obvious that a great deal of effort went into the creation of this report, including the screening of the literature and the detailed extraction of the information that is presented in the tables. Where meta-analyses were possible, these appear to be well-done. It is clear that the evidence-based practice center brought considerable expertise to this systematic review.	Thank you.
Introduction	**TEP Reviewer #4**	Sufficient - please see above. I would recommend a careful review of the language used throughout the full report in order to adopt a "patient first" stance (e.g., referring to "patients with bipolar disorder" vs. "bipolar patients"). On p. ES-1, it is recommended that the data and citations re: risk for suicide in bipolar disorder be updated to reflect data recently reported by the Suicide Taskforce of the International Society for Bipolar Disorders: Schaffer et al. Epidemiology, neurobiology and pharmacological interventions related to suicide deaths and suicide attempts in bipolar disorder: Part I of a report of the International Society for Bipolar Disorders Task Force on Suicide in Bipolar Disorder. Aust N Z J Psychiatry. 2015 Sep;49(9):785-802. doi: 10.1177/0004867415594427. Schaffer et al. A review of factors associated with greater likelihood of suicide attempts and suicide deaths in bipolar disorder: Part II of a report of the International Society for Bipolar Disorders Task Force on Suicide in Bipolar Disorder. Aust N Z J Psychiatry. 2015 Nov;49(11):1006-20. doi: 10.1177/0004867415594428.	Thank you for the suggestion. We have carefully reviewed the report for “patient first” edits and updated the sources for the risk of suicide.

Section	Reviewer, Affiliation	Comment	Response
Introduction	**Peer Reviewer #5**	Good discussion. Perhaps more could be said about the apparent broadening of the definition of bipolar disorder in recent years, e.g. "bipolar III" and varying ideas over the years as to whether medication-induced mania represented "true" bipolar disorder. See next section for further comments on diagnosis.	Thank you for the comment and suggestion. On page 2 of the report, we revised and expanded the paragraphs on bipolar II and NOS to include more information on diagnostic criteria and variants including cyclothymia. To address the issue of medication-induced mania, we also added the sentence “Based on changes in the DSM-5 criteria, in individuals with no prior bipolar disorder diagnosis, drug treatment induced manic and hypomanic episodes that last longer than the expected pharmacological effects are now considered “true” episodes and count towards a bipolar disorder diagnosis.”

Section	Reviewer, Affiliation	Comment	Response
Introduction	**Peer Reviewer #5**	Another aspect of diagnosis that could be mentioned is the difficulty – perhaps increasing – of differential diagnosis of bipolar disorder, especially without use of a full structured interview, such as the SCID. One survey of UK psychiatrists found that even with a good understanding of DSM-IV-R criteria, clinicians had difficulty distinguishing bipolar disorder from borderline personality disorder. The concept of bipolar disorder II seems to be broadening, but given its frequent co-occurrence with other disorders, and the absence of a "clinical marker" such as hospitalization to lend face validity to the diagnosis, unintended heterogeneity may be introduced to some clinical trials. Avoidance of simple "symptom checklists" in favor of complete structured interviews would seem preferable going forward. REFERENCE: Saunders KEA, Bilderbeck AC, Price J, Goodwin GM. Distinguishing bipolar disorder from borderline personality disorder: A study of current clinical practice. European Psychiatry 30 (2015; 30: 965–974.	Thank you. We have added a brief statement to the introduction that notes the importance of structured interviews in differential diagnoses. This issue is addressed in the discussion section on the limitations of the evidence base.
Introduction	**Peer Reviewer #6**	"Drugs for acute mania are generally used for shorter periods than drugs used for treating depression or for maintenance." – This is often not the case. Given the chronic, relapsing/remitting course of bipolar disorder and the need for maintenance treatment in most if not all patients, drugs that are started for an acute mood episode (including mania) are often carried forward into maintenance therapy.	Thank you for the clarification. We have used the suggested sentence. " Given the chronic, relapsing/ remitting course of bipolar disorder and the need for maintenance treatment in many patients, drugs begun for an acute mood episode (including mania) are often carried forward as maintenance treatment."

Section	Reviewer, Affiliation	Comment	Response
Introduction	**Peer Reviewer #6**	"Drugs that alleviate depression may cause mania, hypomania, or rapid cycling (four or more episodes in 12 months), and drugs that alleviate acute mania may cause rebound depressive episodes." – This statement is somewhat vague and would benefit from clarification. First-generation antipsychotics (and possibly risperidone) can increase the risk of depression. Certain antidepressants can increase the risk of hypo/mania (particularly tricyclic antidepressants and serotonin-norepinephrine reuptake inhibitors; it is not clear whether other antidepressant classes increase the risk above baseline). The statement is not true of other drugs.	Thank you for the clarification. We revised the sentence to read that some drugs may increase these risks.
Introduction	**Peer Reviewer #6**	"Key Question 3: What is the effectiveness of treatments to reduce the metabolic change (metabolic syndrome, glucose dysregulation, weight gain) side effects of first line pharmacologic treatments?" There are no data reported on this.	The reviewer is correct. The draft noted the lack of evidence for KQ3 in the discussion section. We added an additional sentence to Chapter 3 Search Results noting that no studies for KQ3 regarding treatments to reduce metabolic change side effects of drug treatments were found that met inclusion criteria.

Section	Reviewer, Affiliation	Comment	Response
Introduction	**Public Comment Anonymous**	My husband got 200 mg of risperidone and got a STROKE..the neurologist told us point blank that the stroke was CAUSED by the rsiperidone, accoding tohis nuerologist!!!! . QUALITY OF LIFE???? MINUS ZERO. THANKS, risperidone, for making our lives worse than we ever thought possible, and that it was risperidone which caused this travesty. The Benefits do NOT, in ANY WAY, outweight the risks of this drug. PERIOD> It did not help him in nay real way with his coping (or NOT coping skills) and now he is a walking shell of the man he used to be. Thanks for NOTHING for prescribing this useless and S\DESTRUCTIVE excuse for a DRUG....It Kills, maims and has mad our lives a MISERY. Being a lab rat for this drug is bad enough...having to live with its consequences is a thing for which I wish I could sue the doctor, the manufacturer and the FDA for approving it. This drug is a health hazard. My husband is somehow still living proof of this. Get this drug OFF of the market, thank you all very MUCH for NOTHING but MISERY.	Thank you for sharing your personal story. We are sorry to hear of your unfortunate experiences.
Methods	**TEP Reviewer #1**	Overall the methods section is thorough.	Thank you.
Methods	**TEP Reviewer #3**	The statistical methods appear to be appropriate. Some of the definitions require additional detail/clarification (see attached comments). There are some issues with the search strategy as described in the general comments attachment. The inclusion and exclusion criteria seemed reasonable when the review was proposed, however, given the large number of studies that were excluded, including FDA pivotal trials and virtually all trials of key treatments, they appear much too stringent in retrospect.	Thank you for the comment. We followed the search process as outlined in the a priori protocol.

Section	Reviewer, Affiliation	Comment	Response
Methods	**TEP Reviewer #3**	A substantial portion of the evidence is described as being "insufficient", which seems to misrepresent the body of available research on bipolar disorder and presents an overly negative view of the available research findings. This is true for psychopharmacological studies as well as for studies of psychotherapy and other psychosocial interventions. It is understandable that the evidence would be deemed insufficient when no studies are done or perhaps when there are one or two studies with very small samples (e.g., < 50 subjects). However, it is difficult to see how the evidence can be categorized as insufficient when many hundreds of subjects have been included in multiple randomized trials (including some pivotal trials that led to FDA indications).	Thank you for the comment. Strength of evidence assessments followed AHRQ methods guidance and the domains and process are described in the methods section. Detailed information for each specific strength of evidence assessment is provided in the appendixes. A rating of insufficient is given when an outcome is too weak, sparse, or inconsistent to draw a defensible conclusion. In these instances the raters do not have even limited confidence in the stability of reported findings.
Methods	**TEP Reviewer #3**	By categorizing much of the available evidence as "insufficient", there is a lack of actionable information about many interventions. However, for the anticonvulsant mood stabilizing medications, valproate/divalproex and lamotrigine, this absence of information leads to significant challenges for guideline development. Valproate/divalproex is one of the most frequently prescribed medications for individuals with bipolar disorder and having no statement about this drug is problematic. In addition, lamotrigine has an FDA approval for use in bipolar I disorder and prior guidelines have suggested its possible use off-label in bipolar-depression. Some specific outcome data and summary statements are essential to have in the review, even if based upon best-available but limited evidence.	Thank you. We understand the strength of evidence assessments create a challenge. The EPC program uses insufficient as the lowest strength of evidence assessment, which is different from GRADE's lowest rating of "very low". However, the findings abstracted from the included studies are provided in outcome tables in the appendix even for outcomes that were assessed as having insufficient evidence.

Section	Reviewer, Affiliation	Comment	Response
Methods	**TEP Reviewer #3**	In a large number of instances, the rating of precision and the overall strength of evidence appears to be downgraded due to the use of last-outcome-carried-forward (LOCF) methods of analysis in individual studies. Although the review provides a detailed discussion of difficulties with the LOCF approach, it seems problematic to downgrade a significant fraction of the evidence for this reason, particularly since the U.S. FDA (and thus many psychiatric journals), preferred or required this type of analysis for many years (See Hamer RM, Simpson PM. Last observation carried forward versus mixed models in the analysis of psychiatric clinical trials. Am J Psychiatry. 2009 Jun;166(6):639-41.)	Thank you. The use of LOCF is a problem with the literature. Unfortunately, we cannot change the impact of this method on the science of systematic reviews. The report uses as reference material a 2011 report from the National Research Council, of the National Academies of Science, that describes and discusses the problems of missing data due to attrition. The National Research Council Workgroup concluded that missing data does not have an easy fix at the analytic stage (such as using LOCF). To do so requires using assumptions that are not testable. Thus, they recommended study designs that prevent loss to followup to minimize problems created by missing data. (Little RJ, D'Agostino R, Cohen ML, et al. The Prevention and Treatment of Missing Data in Clinical Trials. New England Journal of Medicine. 2012;367(14):1355-60. doi: doi:10.1056/NEJMsr1203730. PMID: 23034025)

Section	Reviewer, Affiliation	Comment	Response
Methods	**TEP Reviewer #3**	Based upon the full text review, a large number of references (600) were excluded as not being an included study design or not being a treatment related study. It would be helpful to know how many references were in each of these subcategories since these reasons are distinct from each other. It would also be helpful to know the reasons for exclusion on the basis of study design (e.g., study duration exclusions vs. study type exclusions).	Thank you. We did not prioritize reasons for excluding studies. Exclusion reasons were reported on a first-noted, first-recorded basis. Once excluded, we did not continue to classify exclusions reasons. It is possible that excluded articles may have in fact had more than one reason for exclusion. However, since the decision to exclude studies for attrition was made after the initial screen, that category does hold studies excluded specifically for duration but met all other inclusion criteria.

Section	Reviewer, Affiliation	Comment	Response
Methods	**TEP Reviewer #3**	An additional 129 studies were excluded on the basis of outcome variable and 70 studies were excluded for excessive attrition. Given the large fraction of studies that were excluded for these reasons, I have some concerns about whether potentially useful information was lost. It may be preferable to include more studies and outcomes in the report accompanied by caveats about their limitations.	Thank you. The protocol specified a priori which outcome categories would be gathered. It is possible that articles excluded for lacking included outcomes may also have been excluded for other reasons. While there is no satisfactory techniques yet for addressing multiplicity of outcomes in systematic reviews, systematic reviews do face the same concerns regarding using multiple outcomes as primary research (when techniques such as the Bonferroni adjustment are used). Since the outcomes outlined in the protocol (which were already numerous) were generally frequently reported, we do not believe adding outcomes at this time is warranted. The decision to exclude studies for attrition was made after the initial screen, therefore that category does hold studies excluded specifically for attrition but met all other inclusion criteria. In our judgment, studies with greater than 50 percent attrition are fatally flawed, unless the outcome specifically accounts for attrition, such as “time to event” outcomes. The report did include studies with greater than 50 percent attrition, but only abstracted the “time to event” outcomes.

Section	Reviewer, Affiliation	Comment	Response
Methods	**TEP Reviewer #3**	Attrition is of particular concern in rating confidence in the study findings. Nevertheless, in studies of acute mania, attrition may not be due to lack of efficacy per se. Rather, in an individual with severe symptoms, attrition may simply be related to the lag in time to a medication effect. It is also quite difficult to retain patients with severe mania in studies due to typical restrictions on adjunctive medications such as benzodiazepines. Greater rates of attrition may also be expectable particularly in longer term studies of maintenance therapies. Since this pattern replicates the frequent dropout of patients in clinical practice, it is still preferable to know the effects of an intervention for those individuals who are able to be maintained in treatment.	Thank you. Attrition, and the resulting missing data, is particularly challenging for experimental trials. The reviewer points out several reasons why attrition may occur. Unfortunately, these reasons for attrition support the concern that any assumptions that data is missing at random would be inappropriate, leading to further support for not including studies with attrition and missing data problems.
Methods	**TEP Reviewer #3**	In terms of the study durations that were used as cut-points for inclusion, the document refers to specific periods of "followup" for acute mania, acute depression and maintenance treatments. For some studies, the total study duration includes a period of active treatment but may also include a period of open-label followup or followup assessments after the cessation of active treatment. The document should clarify whether the 3 week, 3 month and 6 month "followup" periods refer to active treatment only, active treatment plus subsequent followup (if applicable) or only post-treatment followup. In describing the review criteria, the study designs and the findings, it would be preferable to specify the duration of each of these phases.	Thank you for the comment. We have clarified in Table 2 the inclusion cut-points were treatment duration plus post-treatment followup (if any). Evidence tables and tables in the report generally reported details if treatment duration was less than the outcome measurement period, including the minimum cut-point for outcome measurement. We have carefully checked these tables and added any missing information regarding time frames.

Section	Reviewer, Affiliation	Comment	Response
Methods	**TEP Reviewer #3**	If the pivotal trials used for FDA approval of specific medications did not meet inclusion criteria for this review, it would still be helpful to describe the trial(s) and their findings and the specific reasons for exclusion. Because the evidence in those trials was viewed as strong enough for FDA approval, readers and guideline developers need to understand the strengths and limitations of those studies.	Thank you. Requirements for FDA approval are based on a different assessment process to meet different programmatic goals than the EPC program and its systematic review methodology. Reporting on studies not included in the review is outside the scope of the review. Because we anticipated disappointment with our excluding studies for greater than 50 percent attrition, we did provide a table in the appendix with minimal information abstracted on those studies.

Section	Reviewer, Affiliation	Comment	Response
Methods	**TEP Reviewer #3**	By excluding articles prior to 1970, articles may be missed, particularly those related to lithium or chlorpromazine.	As we noted in the limitation section, the choice of search dates and search process may have missed some studies on lithium. However, since we handsearched reference lists of relevant reviews to locate studies, we are fairly confident we found the relevant studies that were likely to provide usable information. In our experience, searches of very old literature is often not a good use of resources available for this kind of scientific inquiry. With such old literature, abstracts are not available, thus requiring hunting down (often poor quality) pdfs of the papers for full text review of papers that often did not meet reporting requirements to allow for adequate assessment of risk of bias.
Methods	**TEP Reviewer #3**	The more recent GERI-BD study (Young et al., 2017) also does not seem to be included but should be mentioned.	Thank you. This study was published after the last search date of this review.
Methods	**TEP Reviewer #3**	I am not an expert on the use of the methods employed in the statistical analysis in this report, and defer to others to comment on this aspect of the review.	No response required.

Section	Reviewer, Affiliation	Comment	Response
Methods	**TEP Reviewer #4**	I do question the decision to limit inclusion of studies focused on depression to only those who report a minimum of 3 months of outcome (especially in contrast to the much shorter follow-up adopted for mania). The authors justify this decision much later in the report, stating (p. 108) "... if a treatment response to depression is not sustained, does it matter if the initial response to one treatment was faster than another?" I would argue that, in the case of bipolar depression, a faster treatment response may indeed make an important and substantive clinical difference, especially with regard to more immediate risk for suicide (which is incredibly high in bipolar disorder and correlates more strongly with depression than mania) and need for hospitalization.	Thank you. The decision to limit inclusion of studies examining bipolar depression treatments to a minimum of 3 months of outcome (i.e., active treatment plus follow-up if any) was based on several factors. As another Reviewer has noted, bipolar disorder is often chronic and recurrent disorder with frequent relapses, thus having a minimum duration that allows for at least full remission (e.g., DSM-5 defines it as two months without symptoms) was important. Given that depressive episodes are typically of longer duration than mania (i.e. at least two weeks of duration, often longer), it seemed appropriate that the minimum followup for depression treatment studies would be also longer than for acute mania treatment studies. Moreover, initial response to antidepressant medication treatments can often take weeks, also indicating a need for a longer minimum duration. Finally, in the preparation of inclusion criteria for studies, we relied on recommendations of key informants with expertise in the bipolar disorder treatment area, who unanimously agreed with the proposed 3-month minimum duration criteria for depression studies.

Section	Reviewer, Affiliation	Comment	Response
Methods	**Peer Reviewer #5**	The approach is completely reasonable and well-described. The "problem" seemed to be that all of the appropriate trials that should be available to include simply do not exist in many cases.	Thank you for your comment.
Methods	**Peer Reviewer #5**	I wonder whether in the case of electroconvulsive therapy (therapy), the follow-up requirement was too strict, given that – in contrast to pharmacotherapy – ECT is often a separate and discrete time-limited intervention for an acute episode (most often of depression), which is then often followed by a completely different regimen (most often medication). Unfortunately, bipolar depression is often an exclusion criterion in recent large clinical trials of ECT. "Electroconvulsive therapy (ECT) is administered to bipolar depression patients who are not improved by pharmacotherapy, or have severe symptoms such as suicidal ideation or catatonic features. ECT is associated with high remission rates, particularly in cases with short symptom duration, psychotic symptoms, or older age" (Haq et al., 2015). REFERENCE: Haq AU, Sitzmann AF, Goldman ML, Maixner F, Mickey BJ. Response of depression to electroconvulsive therapy: A meta-analysis of clinical predictors. Journal of Clinical Psychiatry 2015; 76:1374–1384.	Thank you for the comment. The ECT literature was challenging, since the articles did not provide treatment duration or study followup in units of time; instead, they provided outcomes based on number of treatment sessions. We agree that the question of short-term treatment for acute episodes is an important question. However, given this review's focus on the chronic nature of bipolar disorder, especially bipolar depression, the question whether the treatment effect endures for a full 3 month followup period and thus does not require retreatment in less than 3 months is also, we believe, an important question to address from the patient perspective.
Methods	**Peer Reviewer #6**	p. 11 – the final bullet point should read "weight gain of > 7 percent", not "< 7 percent".	Thank you, this has been corrected.

Section	Reviewer, Affiliation	Comment	Response
Methods	**Public Comment Anonymous**	It took the "doctor" about three minutes to prescribe this "drug" to my husband and now it's a year and a half later and I, his wife, have lived through three hospital stays, due to STROKES caused by this drug and three extended stays in nursing homes for my husband, who is a shell of whom he used to be. His walking and cognitive skills are richly impaired...thanks to this drug. A one size fits all approach to treating mental illness is directly to BLAME for this and it is driving ME crazy, having to see a good man drop to this poor level of "functioning" due to new "fashions" and experiments in treatments of HUMAN BEINGS for depression and misery. Thanks for worse than NOTHING, for treating my husband like a lab rat with this drug. You all should be ashamed of yourselves!!!!	Thank you for sharing your personal story. We are sorry to hear of your unfortunate experiences.
Results - Overall	**TEP Reviewer #1**	Overall, very comprehensive; it would helpful though to include information on which studies might have used statistical techniques to account for missing data (e.g., imputation), as well as whether quality of the studies, ROB or other factors might have differed by study setting. For example, it might be easier to follow up on patients from academic medical centers versus community health settings. Such differences would place some of the variation and heterogeneity of the studies in better context.	Thank you for the comment. These are interesting ideas, and worth consideration for a derivative publication based on this report to help researchers design future studies. However, given generally sparse literature for any given intervention/comparison set, drilling down into the studies at this level is not likely to shift the findings in any material way or alter review interpretation.

Section	Reviewer, Affiliation	Comment	Response
Results – Overall	**TEP Reviewer #3**	Although the results section appears to contain a significant amount of detail, much of the text is too non-specific to support making inferences about clinical implications. In the body of the text, it is not helpful to state that one treatment was superior to another or that adverse effects were more frequent than another without giving information on the summary data for each treatment arm and the magnitude of the treatment effect. Similarly, knowing that two head-to-head treatments did not differ is not as informative as knowing whether both showed no affect relative to baseline or whether both were associated with clinical improvement. The key messages, as currently formulated, are not sufficient to assist with clinical care or guideline development. They focus primarily on the weaknesses of the data with little concrete information about specific interventions in specific clinical circumstances. The appendices contain a wealth of information and seem quite thorough, although it would be preferable to have all of the information included rather than cross-referencing other tables or Forest plots.	Thank you for the comment. The goal of the report text was to report in a summary manner that balances readability with detail. We have added, similar to the table provided in the Evidence Summary, tables in the results sections that provide summaries of specific findings and the related strength of evidence. Given the head-to-head comparisons were assessed as having evidence that was insufficient to draw conclusions, greater detail was not provided in the report body. To satisfy those looking for detail, the appendix provides considerable detail.
Results – Overall	**TEP Reviewer #3**	The Forest plots are helpful and the shading helps to show the strength of evidence of the studies.	Thank you.
Results – Overall	**TEP Reviewer #3**	I felt that the review ignored information on the long-term studies of lithium on reducing suicide risk, for example, as well as longitudinal studies on lithium harms. Information on other treatment harms was not very illuminating. Additional comments are included in the attachments.	Thank you for the comment. We have added in the Limitations of the CER section of the Discussion chapter the difficulty assessing suicide when much of the literature suffers from attrition, making meaningful comparisons difficult.
Results – Overall	**TEP Reviewer #3**	A crucial aspect of guideline development is weighing the benefits of possible treatments for a condition against the harms of those treatments. By relying primarily on the information on adverse effects in package inserts, the review does not seem to address the breadth and depth of information available on the potential harms of the included treatments. This will make it challenging to use the review for guideline development. An expanded discussion of available data on adverse effects is essential to incorporate.	Thank you for the comment. The appendix provides details on the harms selected per protocol. Given the wide range of drugs and drug classes examined, we limited the harms abstracted, particularly for nonsevere harms. The information from the package inserts was provided in part to offset this.

Section	Reviewer, Affiliation	Comment	Response
Results – Overall	**TEP Reviewer #3**	When drawing conclusions, it is important to state them as clearly, accurately and consistently as possible so that readers can understand the implications for clinical application or guideline development. In several locations in the document, study findings are described as "mixed" but no other details or context is given. It would be preferable to state the specific ways in which the study findings differed.	Thank you for the comment. The goal of the report text was to report in a summary manner that balances readability with detail. We have added, similar to the table provided in the Evidence Summary, tables in the results sections that provide summaries of specific findings and the related strength of evidence. Those looking for greater detail will find them in the appendix.
Results – Overall	**TEP Reviewer #3**	Statements about findings are expressed in non-specific terms. Include relevant numerical values and statistics (e.g., mean difference, 95% CI, N) with all statements that summarize study findings, even if the evidence is rated as insufficient. Knowing that "response and remission were not significantly different between the groups" is not nearly as helpful as being apprised of this fact but also knowing the actual rates of response and remission in each treatment arm. Other examples are that "placebo had a lower rate of withdrawal due to adverse events than [drug]" or "participants with [drug] had significantly more extrapyramidal side effects than those on placebo." Again, these statements are useful but clinicians and patients would also want to know the proportions of individuals who experienced extrapyramidal side effects or withdrew due to adverse events as compared to placebo.	Thank you for the comment. The goal of the report text was to report in a summary manner that balances readability with detail. We have added, similar to the table provided in the Evidence Summary, tables in the results sections that provide summaries of specific findings and the related strength of evidence. Those looking for greater detail will find them in the appendix.
Results – Overall	**TEP Reviewer #3**	Head-to-head comparison study results are relevant to two (or more) sections of the document. Describe the study findings in each relevant section of the document. Currently, the study is described in one part of the document but not others. When the reader is sent to a different document section for details, it disrupts the readers' train of thought. The amount of redundancy is fairly small and offset by a more coherent user experience.	Thank you for the comment. The goal of the report text was to report in a summary manner that balances readability with detail. To reduce redundancy and improve readability, we provided signposts to the sections that provided greater detail.

Section	Reviewer, Affiliation	Comment	Response
Results – Overall	**TEP Reviewer #3**	Studies that include BP-I and BP-II patients. Include the percentage of each BP subtype in the evidence tables, even if outcomes are not shown with stratification by subtype. Knowing the proportion of individuals with specific BP subtypes in the sample will help the reader assess the applicability of the study finding to their patient.	Thank you for the comment. The evidence tables in the appendix and populations tables in the report provided the percentage of patients with bipolar type I. We have added the percentage with bipolar type II or other types if the study included more than one type of bipolar disorder.
Results – Overall	**TEP Reviewer #3**	Studies that permitted patients to be on one of several possible mood stabilizers with random assignment to an adjunctive medication (vs. placebo). Report whether outcomes were stratified by the mood stabilizer that was taken by the patient. It is possible to have synergistic benefits of an adjunctive medication that may different depending upon the mood stabilizing medication that is used.	Thank you for the comment. Studies generally provided specific information about the mood stabilizers that were allowed and did not report outcomes stratified by type of mood stabilizer.
Results – Overall	**TEP Reviewer #3**	Studies that involve blood level monitoring of specific medications (e.g., lithium, valproate, carbamazepine). Include information on the average blood levels achieved in the study as part of the tables and text of the document. This will allow readers and guideline developers to assess the adequacy of doses during the trial and assess whether adverse event rates may have been related to higher than usual blood levels.	Thank you. This level of detailed abstraction is outside the scope of this review and was not a protocol abstraction element.
Results – Overall	**TEP Reviewer #3**	Additional calculations may be needed to calculate summary statistics such as effect sizes, number needed to treat (NNT) or number needed to harm (NNH) for inclusion in tables that summarize medications or psychosocial treatments.	Thank you. We have provided NNT for response or remission outcomes with at least low-strength evidence.

Section	Reviewer, Affiliation	Comment	Response
Results – Overall	**TEP Review #3**	When the ratings of strength of evidence are discussed, many statements are non-specific (e.g., "moderate or high study limitations"). Where possible, it would be helpful to include the specific reasons that a study was viewed as having limitations.	Thank you for the comment. The report text provides the domains that contributed to lower assessments of strength of evidence. The risk of bias tables in the appendix provides detailed information regarding reasons individual studies were assessed as having moderate or high risk of bias.
Results – Overall	**TEP Reviewer #4**	The results are very detailed and I found the tables to be incredibly useful in synthesizing the results.	Thank you.
Results – Overall	**Peer Reviewer #5**	The included studies are well-described and the informative tables are excellent.	Thank you.
Results – Overall	**Peer Reviewer #5**	Due to logistical and ethical challenges, the field seems to have given up efforts at conducting RCTs of electroconvulsive therapy (ECT) in mania. However, it is perhaps worth noting – if only in a footnote, that historically ECT has been seen to be a safe and effective treatment for acute mania: "The evidence indicates that ECT is associated with remission or marked clinical improvement in 80% of manic patients and that it is an effective treatment for patients whose manic episodes have responded poorly to pharmacotherapy." REFERENCE: Mukherjee S, Sackeim HA, Schnur DB. Electroconvulsive therapy of acute manic episodes: A review of 50 years' experience. Am J Psychiatry 1994; 151:269-276.	Thank you for the comment. We would have included ECT for acute mania if the study reported outcomes at 3 weeks (treatment duration plus followup).

Section	Reviewer, Affiliation	Comment	Response
Results – Search	**TEP Reviewer #3**	A number of questions arise with respect to the search strategy for the literature review. The search terms include title words for bipolar*, cyclothymia, rapid adj cycl*, mania, hypomania, manic and hypomanic; however, additional references may have been retrieved by searching these terms in the abstract as well as in the title. Other unique references might be retrieved by including search terms for "affective disorder". The current strategy also seems to include some unusual delimiters (e.g., line 123 (resection or prostate or radiofrequency or sealer or ablation or hip or fibrillation).ab.; line 83 exp postoperative complication/; line 84 exp intraoperative complications/. Lines 57 and 58 are identical. Although this duplication would not change the search findings, it suggests that a different term in line 58 may be missing. These questions about the search strategy raise concerns about whether any relevant articles have been missed although the systematic review does include cross-checks by examining reference lists of other systematic reviews.	Thank you for the comment. The search algorithm was designed to balance precision, or specificity, with sensitivity. Some terms were used to remove references related to unrelated medical or surgical topics, such as bipolar hip replacement, from the database "hits." The cross-checks of systematic reviews was indeed intended in part to assure all relevant articles were identified.
Results – Search	**Peer Reviewer #6**	Table 4 – Why is paliperidone listed under "Other drugs for acute mania" instead of "Antipsychotics"? I presume it's because of lack of FDA approval? This does not change the fact that it is a 5HT2/D2 blocker/ antipsychotic. Perhaps for clarity "Other drugs for acute mania" could be relabeled "Drugs Not Approved by FDA For Acute Mania in Bipolar Disorder", as in the Chapter on Acute Mania?	Thank you for the suggestion. We have relabeled that section of the table.
Results – Search	**Peer Reviewer #6**	Table 5: There is a study of ECT vs drug treatment in acute depression that does not appear to be included, nor do I see it in the list of excluded studies (Appendix D): Schoeyen HK, Kessler U, Andreassen OA, Auestad BH, Bergsholm P, Malt UF, Morken G, Oedegaard KJ, Vaaler A. Treatment-resistant bipolar depression: a randomized controlled trial of electroconvulsive therapy versus algorithm-based pharmacological treatment. Am J Psychiatry. 2015 Jan;172(1):41-51.	Thank you for the suggested article. Unfortunately, the study addressed depression treatment but the study duration (including followup) was 6 weeks, not the minimum 3 months.

Section	Reviewer, Affiliation	Comment	Response
Results – Search	**Public Comment Anonymous**	Results obtained by treating very sick and sad patients, depending upon doctors for expertise, yet just being treated like lab rats for the testing of new drugs and their side effects, without warning patients by said doctors...(the lables which come with these drugs, with their chemical compounds shown but real results and hazards NOT CLEARLY EXPLAINED, the "trust me" attitude of doctors prescribing this rug and the "what did you THINK was going to happen by using this drug" attitudes by attending neurologists after the strokes, is APPALLINGLY IRRESPONSIBLE and very angering to me, the caretaker of a once-vital but now child-like patient who was given this killer drug. Thanks, but no thanks, to YOU ALL for just shoving this drug at my husband with no REAL REGARD for him and his (OR MY , the CARETAKER!) regard for safety of drug efficacy, which was NEGATIVE int the extreme!!!!!!	Thank you for sharing your personal story. We are sorry to hear of your unfortunate experiences.
Results – Acute Mania	**TEP Reviewer #3**	In several places in the document, there are inconsistencies in key statements. This reduces confidence in the document content and could result in erroneous guideline recommendations, with implications for patient care. For example, on p. 29, comparisons of olanzapine and asenapine are said to provide low strength evidence for a greater response to asenapine whereas p. 25 notes that there is insufficient evidence but a greater response with olanzapine. Also, p. 48 notes that topiramate was associated with a higher rate of study withdrawals due to adverse events than lithium whereas p. 47 notes the opposite.	Thank you for the correction for the signposting of asenapine vs olanzapine in the olanzapine section. We have corrected this to state evidence was insufficient for this comparison. The statements regarding withdrawals for adverse events for lithium being higher than for topiramate was correct in both places. (One said lithium withdrawal was higher while the other stated topiramate withdrawals were lower.) We have carefully reviewed the report to assure consistency in the statements.

Section	Reviewer, Affiliation	Comment	Response
Results – Acute Mania	**Peer Reviewer #6**	My main concern across all treatments reviewed in this section (and in the other treatment sections) is that the highly stringent approach to selecting and classifying trials, coupled with the inherent difficulties in conducting clinical trials in psychiatric patients, leads to the exclusion of many trials and a likely underestimation of the effectiveness of treatments. This makes the manuscript an excellent review of the limitations in the evidence base for the treatment of bipolar disorder, but markedly limits its usefulness to clinicians.	Thank you for the comment. We respectfully disagree that the exclusion of studies with high attrition hinders our ability to draw appropriate conclusions. Adding studies with risk of bias so high as to be a fatal flaw does not improve our ability to draw conclusions. The main purpose of a rigorous systematic review is to improve our understanding of the state of the science and what is known with higher or lower degrees of confidence in the findings.
Results – Acute Mania	**Peer Reviewer #6**	The olanzapine studies provide a good example. Fifteen studies met criteria for inclusion of the review, while 16 were excluded. Some of the exclusion criteria are, to a degree, arbitrary (e.g., a study with an attrition rate of 50% would be excluded while one with an attrition rate of 49% would not). The nature of psychiatric illness is such that it is even more difficult to avoid attrition than in other medical illnesses, and attrition rates in most studies are therefore high. Furthermore, the subjective nature of psychiatric symptoms makes psychiatric trials especially prone to inconsistent and imprecise results (the other main reasons for not including study data). The end result is that even though the excluded studies were generally more similar than different to the included ones, in both design and results, conclusions are drawn based on a relatively small subset of all studies that were conducted, which may or may not be representative of the true efficacy and tolerability (or lack thereof) of olanzapine, and the other treatments.	Thank you. We agree that the line drawn for attrition was arbitrary, however, specific cut-offs reduce the opportunity for bias in the review. When pooling was possible, the meta-analyses differentiated between low/moderate risk of bias studies and high risk of bias studies. Adding further high risk of bias studies would not improve our ability to draw conclusions from the literature and may, in fact, reduce our confidence in the findings by weighting the set of included studies more toward high study limitations when assessing strength of evidence.

Section	Reviewer, Affiliation	Comment	Response
Results – Acute Mania	**Peer Reviewer #6**	The prioritization of MID as a measure of effectiveness (e.g. it is featured prominently in the Abstract and the concluding section of the manuscript) is also problematic. Olanzapine, like every other treatment reviewed in the document, did not reach a MID of 6 and so is judged to be of limited effectiveness. Another metric, e.g. the OR for response, shows that olanzapine treated patients are about twice as likely to achieve a treatment response as placebo-treated patients. This result is likely to be viewed differently by mental health practitioners.	Thank you for the comment. Minimally important differences (MID) are used to distinguish between those findings that are merely statistically significant from those that are clinically significant in the sense that a typical patient would likely be able to notice an improvement in their experience of that particular outcome. MIDs are a standard systematic review methodology for outcomes like the Young Mania Rating Scale.
Results – Acute Mania	**Peer Reviewer #6**	P 21, "Eligible studies for antipsychotics", line 3: aripipraZOLE, not aripipraZINE.	Thank you, this has been corrected.
Results – Acute Mania	**Peer Reviewer #6**	"There were no studies assessing drug effectiveness in treatment of hypomania." – This is at least partly incorrect. The McElroy 2010 study cited in Table 13 (ref 83) included manic and hypomanic patients. (The authors themselves note this in p 31: "Another small study enrolled participants with mild to moderate hypomania or mild mania regardless of type of BD.") However, it did not report the results separately for hypomanic OR manic patients. The authors should thus decide whether it should be included in studies of quetiapine in acute mania, since it did not enroll an exclusively manic sample.	Thank you, the sentence has been revised to state no studies specifically assessing drug effectiveness in treatment of hypomania. The McElroy 2010 study was not pooled with the other quetiapine vs placebo studies for acute mania.
Results – Acute Mania	**Peer Reviewer #6**	P 29, "Results for olanzapine versus asenapine were reported in the asenapine versus active control section above (e.g., low-strength evidence for greater response for asenapine but no difference in remission rates)." This is incorrect and is the opposite of what was reported in the asenapine v active control section.	Thank you for the correction for the signposting of asenapine vs olanzapine in the olanzapine section. We have corrected this to state evidence was insufficient for this comparison.

Section	Reviewer, Affiliation	Comment	Response
Results – Acute Mania	**Peer Reviewer #6**	P 40: "Evidence was insufficient for all outcomes from two RCTs (n=670) to address whether divalproex sodium was better for acute mania than placebo in adults with BD-I… Both studies reported no differences between groups in response, remission, and symptoms at 3 weeks." This is incorrect. The Bowden (2006) study reported "Improvement from baseline on the MRS was significantly greater among patients who received divalproex ER compared with placebo at the first on-treatment rating assessment, day 5, and all subsequent ratings through day 21 (p = .013). Furthermore, the proportion of patients achieving at least 50% improvement from baseline in MRS was significantly higher in patients receiving divalproex ER (48%) than in patients receiving placebo (34%) (p = .012)."	Thank you for the correction. Bowden did report positive results while the second study by Tohen found no difference. The report text has been corrected to match the details reported in the appendix. The strength of evidence remains insufficient for the same reasons (high study limitations, inconsistency, imprecision).
Results – Acute Mania	**Peer Reviewer #6**	On p 42 the authors state "Low-strength evidence from one RCT and one meta-analysis of independent data from 4 RCTs (n=847) showed lithium increased response and remission rates in BD-I participants compared to placebo for acute mania. (86, 128)." Ref 128 is in fact for a review of 4 trials of topiramate in acute mania. It is thus unclear what evidence the authors evaluated for the effectiveness of lithium vs placebo.	Thank you for the comment. Ref 128 is a report of 4 trials that were not previously published. The publication combined the studies at the level of individual patient data. Our review used the reports of the 4 trials provided at the study level.
Results – Acute Mania	**Peer Reviewer #6**	On p 12 the authors state "Study outcomes were grouped by treatment duration or followup period. For acute mania treatment, outcomes were grouped by 3-4 weeks and then final measurement (generally 6 to 12 weeks) if available." Many if not most acute mania studies had their primary outcome at 3 weeks, but some (e.g. several of the lithium studies) were longer. It is not made explicitly clear in the text whether the outcome data for all treatments were from the same time point (the authors refer to including studies "with at least 3 weeks followup". If results for different medications were from different time points, this is problematic. Trials in which patients were exposed to active treatment for shorter periods might be expected to have less positive results than longer trials.	Thank you for the comment. While the protocol stated we would abstract outcomes at the 3 weeks and final time frame for acute mania, in practice the number of studies with usable outcomes post-3 weeks was rare due to problems with attrition. Outcomes at times different than 3 weeks are specifically noted. All outcomes for acute mania treatment with at least low strength of evidence were at 3 weeks.

Section	Reviewer, Affiliation	Comment	Response
Results – Acute Mania	**Peer Reviewer #6**	P 43 – If in the Lithium Plus OPT Versus OPT Alone study "87 percent of the participants were experiencing a depressive state", why is this included in the mania section?	Thank you for the comment. The lithium plus OPT study could be slotted into either the mania section or the depression section, and the review team was evenly divided as to which section to report the study. We have moved the study to the depression chapter.
Results – Acute Mania	**Peer Reviewer #6**	P 48 – "Patients with hypomania tended to show the same benefits for quetiapine as patients with moderate to severe mania." Presumably this conclusion is drawn deom the McElroy study. However, this study did not report results separately for manic and hypomanic patients.	Thank you for the comment. We have revised the sentence to note that patients with hypomania or mild mania tended to show the same benefits.
Results – Acute Mania	**Public Comment Jannsen Pharmaceutical**	Page 34 Table 15. Hirshfeld 2004 "Rapid antimanic effect of risperidone monotherapy: A 3-week multicenter, double-blind, placebocontrolled trial" is not listed and is a pivotal trial supporting the efficacy of risperidone monotherapy.	Thank you for the comment. This study was excluded due to greater than 50% attrition. The study is minimally abstracted in Table D1 of the appendix.
Results – Acute Mania	**Public Comment Anonymous**	DID NOT HELP, but did cause a series of STROKES. TERRIBLE and callous misuse of this drug to gather data on its side effects...my husband had to suffer because he TRUSTED his doctors who gave him a drug in doses which, as per his neurologist, were going to either kill him or cause him major brain damage. Now it is I, his caretaker, who must live with the consequences, the cleaning, the feeding, the watching him forget where he LIVES, the laundry, the constant drudgery of a Life gone terribly, maddenly awry, all because you all wanted to "see how this drug works." Well, folks, it doesn't. Happy NOW???? And now we have to lose ourt life savings too.. Icing on your cake of Mengela-like experimentation on human beings, their lives and their suffering FAMILIES. SHAME ON YOU ALL!!!!	Thank you for sharing your personal story. We are sorry to hear of your unfortunate experiences.

Section	Reviewer, Affiliation	Comment	Response
Results - Depression	**Peer Reviewer #5**	I realize that the review deals with present-day clinical realities. I wonder, though, if it is worth mentioning that historically, the monoamine oxidase inhibitor (MAOI) antidepressants were regarded as the pharmacotherapy of choice for bipolar depression – so that they are not needlessly left behind while there might still be a role for them? REFERENCE: Heijnen WT, De Fruyt J, Wierdsma AI, Sienaert P, Birkenhäger TK. Efficacy of tranylcypromine in bipolar depression: A systematic review. J Clin Psychopharmacol 2015; 35:700-705.	Thank you for the comment. The studies in the Heijnen and colleagues review were all for 6 weeks and two included participants other than bipolar disorder. In addition, one of the included studies used participants from another included study, thus double-counting them. Given our current review, we would not be in a position to comment on this treatment.

Section	Reviewer, Affiliation	Comment	Response
Results - Depression	**Peer Reviewer #6**	The methodology of this section is very concerning. The authors' decision to exclude depression trials of less than 3 months duration is, as best I can tell, arbitrary. They state that "Given the chronic nature of bipolar disorders, it is doubtful that studies reporting efficacy or comparative effectiveness for drugs with followup periods shorter than 3 months are clinically useful". Later, they similarly ask "If a treatment response to depression is not sustained, does it matter if the initial response to one treatment was faster than another?" However, these arguments also hold for mania, and their decision thus sits uncomfortably with their decision to include much shorter (3 week) mania trials. As a result of the decision, absolutely no consideration is given to otherwise methodologically sound, replicated studies showing effectiveness for the FDA-approved drugs quetiapine, olanzapine, and lurasidone, and a meta-analysis showing modest effectiveness for lamotrigine. For quetiapine in particular, the fact that it was found to be effective in treating acute depression over 8 weeks; and in increasing time to relapse to depression during maintenance treatment over 1 year; shows that it has antidepressant effects across short-term and long-term trials. This makes the decision to exclude it from consideration all the more puzzling. The medications that were deemed worthy of inclusion in this section – antidepressants, mainly as monotherapy, and memantine – are not FDA approved for bipolar depression and do not have clinical support for effectiveness. Thus, this entire section creates the false impression that there are no effective treatment for bipolar depression, and makes it of limited or no use to clinicians. This is especially unfortunate since the symptomatic burden of bipolar disorder is predominantly depressive for the majority of patients.	Thank you for the comment. The study timing inclusion criteria were the result of the topic refinement deliberative process involving discussions with key informants representing major stakeholders. (Please see the acknowledgments area for key informants, and Chapter 2 for a brief discussion of the topic refinement process.) As per the protocol, treatments were not an exclusion criteria. The findings from the review are not evidence of lack, that is, that nothing works. Rather, the findings are of lack of evidence, that is, the available evidence is insufficient to draw conclusions. This finding, then, suggests that one of the 3 legs of evidence-based medicine, evidence, is missing, leaving the 2 legs of clinical and patient experience to inform decisions. Thus, as we note in the discussion chapter, more research is needed.

Section	Reviewer, Affiliation	Comment	Response
Results - Depression	**Public Comment Anonymous**	THIS should NOT be one of them! It is an :anti-psychotic," whatever that means, and did nothing to treat his depression, except give him many small strokes, which have made him into a large baby. Thanks, but no thanks, as now I, his wife, am majorly depressed having to be his caretaker, due to the fact that Medicare does not cover the long-term effects caused by this excuse for a "drug." And if this continues, we will have to pay out all of our life savings, which we sweat blood to EARN, just to apply for Medicaid, if it still even exists, after we deplete what we have saved all of our lives. What is WRONG with you people?????????	Thank you for sharing your personal story. We are sorry to hear of your unfortunate experiences.
Results - Maintenance	**TEP Reviewer #3**	For maintenance studies, it is not clear whether these all began during a maintenance phase of treatment or whether some of these trials began as acute studies and included a continuation/maintenance phase of followup. If any studies were of the latter design, it would be helpful to include links to the reporting of the acute results in the corresponding section of the document.	Thank you. We have noted in the population tables in the text and the appendix tables the inclusion critiera and whether participants were randomized when stabilized, having responded to a treatment, usually open label, or in an acute episode. We have noted when maintenance studies were extensions of short-term studies for acute episodes.

Section	Reviewer, Affiliation	Comment	Response
Results - Maintenance	**Peer Reviewer #5**	It seems odd that the results for lithium are not stronger! For example: "Lithium has the strongest evidence for long-term relapse prevention" (Geddes and Miklowitz 2013; Joas et al. 2017). REFERENCES: Geddes JR, Miklowitz DJ. Treatment of bipolar disorder. Lancet 2013; 381:1672-1682. Joas E, Karanti A, Song J, Goodwin GM, Lichtenstein P, Landén M. Pharmacological treatment and risk of psychiatric hospital admission in bipolar disorder. Br J Psychiatry 2017; 210:197-202. And what about the evidence that lithium reduces suicide rates (Cipriani et al. 2013) and suicide attempts (Song et al. 2017)? REFERENCES: Cipriani A, Hawton K, Stockton S, Geddes JR. Lithium in the prevention of suicide in mood disorders: Updated systematic review and meta-analysis. BMJ 2013; 346:f3646. Song J, Sjölander A, Joas E, Bergen SE, Runeson B, Larsson H, Landén M, Lichtenstein P. Suicidal behavior during lithium and valproate treatment: A within-individual 8-Year prospective study of 50,000 patients with bipolar disorder. Am J Psychiatry 2017; 174:795-802.	Thank you for the comment. In relative terms, the evidence in maintenance phase for lithium is stronger than for other drugs, with an improvement equal to the minimally important difference. Thus, we also found the strongest evidence for lithium in comparison to other treatments, it is just that we deemed the absolute strength of evidence as low due to study limitations and imprecision. The references provided here in support of lithium are more traditional narrative reviews and do not assess the strength of the evidence according to AHRQ or GRADE standards. The Song article was published after the last search update (and while the report draft was being finalized). Unfortunately, the updated systematic review and meta-analysis included both unipolar depression and bipolar disorder and do not report the results separately for the different conditions.

Section	Reviewer, Affiliation	Comment	Response
Results - Maintenance	**Peer Reviewer #6**	In summarizing their conclusions on p 61, the authors state: "19 of 29 of maintenance studies (65 percent) were rated as having severe study limitations (high risk of bias). Second, the high rates of attrition often led to only one usable outcome measure—time to recurrence of a bipolar episode—since this metric accounted for high attrition rates by including information from participants who dropped out due to BD episode relapse. Moreover, 14 studies had small sample sizes of less than 200 participants and 21 studies (79 percent) had followups between six to twelve months, precluding conclusions for long-term maintenance for most of examined treatments. Third, differences in current bipolar phase criteria across studies, ranging from any current phase (i.e., depression, hypomania, or euthymia), remission from mania, remission from any BD episode, or response or partial response to a specific acute episode treatment, made it difficult to determine for whom findings might apply." If these were limitations severe enough that no usable data could be extracted from the above studies, why were they not exclusion criteria? I note that the authors explicitly INCLUDED studies with time to relapse as an outcome, and studies of 6 months duration. Finally, these criteria set an extraordinarily high bar – few if any maintenance trials in psychiatric samples could ever meet the above standards.	Thank you for the comment. The reasons listed in the section pulled out and highlighted here were discussing the challenges in the included literature that contributed to the frequent strength of evidence assessment of insufficient. The assessment, as noted in Chapter 2 Methods, depends on the study limitations (based on the individual studies risk of bias), the consistency of the findings across the studies, and the precision of the data and estimates of effects. We fully acknowledge the difficulty conducting such research. This, however, does not change the underlying scientific rigor needed to draw strong versus weak conclusions, or even the ability to draw conclusions at all.

Section	Reviewer, Affiliation	Comment	Response
Results - Maintenance	**Peer Reviewer #6**	For reasons that are unclear to me, the authors' conclusions for some drugs appear to differ from those reported in the applicable trials. • For example, in a pooled analysis of the two 18-month lamotrigine monotherapy studies (Goodwin et al, J Clin Psychiatry 2004;65:432-41), the investigators report that "Lamotrigine and lithium were superior to placebo for time to intervention for any mood episode (median survival: placebo, 86 days [95% CI = 58 to 121]; lithium, 184 days [95% CI = 119 to not calculable]; lamotrigine, 197 days [95% CI = 144 to 388]). Lamotrigine was superior to placebo for time to intervention for depression (median survival: placebo, 270 days [95% CI = 138 to not calculable]; lithium, median not calculable; lamotrigine, median not calculable). Lithium and lamotrigine were superior to placebo for time to intervention for mania (median survival not calculable for any group)." One of those two studies enrolled recently-manic patients and the other enrolled recently depressed patients (and in both, lamotrigine was superior to placebo for time to recurrence to depression), suggesting the results hold regardless of baseline mood state. The authors consider them to have low and moderate risk of bias, per Table 26. However, the authors report that the evidence is insufficient to support lamotrigine as effective in maintenance treatment.	Thank you. Strength of evidence is an assessment of confidence in the findings, which is different than statements of no benefit or benefit from a treatment. As noted in the text, study limitations and imprecision contributed to the assessment of evidence insufficient to draw conclusions. Appendix I gives further detail for these studies, noting the time to event outcomes used log rank methods which assume that the censoring of participants is noninformative, that is, that the reason for the participant dropping out is not useful information. We agree with your comments below (in Discussion) that reasons for a clinical trial study drop out by individuals with bipolar disorders are often related to symptoms of their illness. Thus, assuming that reasons for dropout are random and noninformative as log rank statistical methods do, is a challenging assumption to hold for this patient population. Whether the impact of this on the strength of evidence is done through higher study limitations or through increased imprecision (similar imprecision is noted in the wide confidence intervals from the Goodwin et al study), the final assessment was insufficient.

Section	Reviewer, Affiliation	Comment	Response
Results - Maintenance	**Peer Reviewer #6**	• Similarly, in the quetiapine monotherapy study, "Time to recurrence of any mood event was significantly longer for quetiapine versus placebo (hazard ratio [HR] = 0.29; 95% CI, 0.23-0.38; P < .0001) and for lithium versus placebo (HR = 0.46; 95% CI, 0.36-0.59; P < .0001). Quetiapine and lithium significantly increased time to recurrence of both manic events (quetiapine: HR = 0.29; 95% CI, 0.21-0.40; P < .0001; lithium: HR = 0.37; 95% CI, 0.27-0.53; P < .0001) and depressive events (quetiapine: HR = 0.30; 95% CI, 0.20-0.44; P < .0001; lithium: HR = 0.59; 95% CI, 0.42-0.84; P < .004) compared with placebo". Weisler et al, J Clin Psychiatry 2011;72:1452-64.) Although this is a single trial, it had a large sample size and included recently manic/depressed/euthymic patients, suggesting the results hold in a large clinically representative sample. It is considered to have a moderate risk of bias. However, the authors report that the evidence is insufficient to support quetiapine as effective in maintenance treatment.	Thank you. The review states that evidence is insufficient to draw a conclusion for quetiapine monotherapy for maintenance. Details for the strength of evidence assessment are in Appendix I. The assessment was based on moderate study limitations, based on moderate risk of bias for the one study, unknown consistency (since there was only one study), and imprecision in the findings. Confidence intervals for the hazard ratios are wide. (Confidence intervals for negative numbers must fit within a 0 to 1 interval, so are not as intuitive as confidence intervals for positive hazard ratios, which could range from 1 to infinity.) We have added a brief summary of reasons for insufficient evidence conclusions throughout the report, in order to highlight factors contributing to these ratings.

Section	Reviewer, Affiliation	Comment	Response
Results - Maintenance	**Peer Reviewer #6**	• Likewise, the two trials of quetiapine or placebo + mood stabilizer, "Treatment with quetiapine in combination with lithium/divalproex significantly increased the time to recurrence of any mood event compared with placebo plus lithium/divalproex" (Vieta et al, J Affect Disord 2008;109:251-63), and "The hazard ratio for time to recurrence of a mood event was 0.32. Hazard ratios were similar for mania and depression events (0.30 and 0.33, respectively)" (Suppes et al, Am J Psychiatry 2009;166:476-88).	Thank you. Details for the strength of evidence assessment are in Appendix I. The assessment was based on high study limitations and imprecision in the findings. Confidence intervals for the hazard ratios are wide. (Confidence intervals for negative numbers must fit within a 0 to 1 interval, so are not as intuitive as confidence intervals for positive hazard ratios, which could range from 1 to infinity.) We have added a brief summary of reasons for insufficient evidence conclusions throughout the report, in order to highlight factors contributing to these ratings.
Results - Maintenance	**Public Comment Jannsen Pharmaceutical**	Page 53. Eligible studies for maintenance treatment. Quiroz 2010 excluded due to 50% attrition and not using time to relapse outcomes. The primary efficacy variable was the time to recurrence of a mood episode during double-blind treatment. 303 patients entered the 24-month DB phase and 77 patients discontinued; 75% of the patients either completed the study or relapsed (primary efficacy outcome)	Thank you. We have reassessed the study and included it as a maintenance study.
Results - Maintenance	**Public Comment Jannsen Pharmaceutical**	Page 59 MacFadden 2009 Demograghics. The age of the population entering the DB trial (N=124) is 18-63 (not 18-70).	Thank you, we have made this correction.
Results - Maintenance	**Public Comment Anonymous**	Forget this one. His neurologist took him OFF of this excuse for a treatment.	Thank you for sharing your personal story. We are sorry to hear of your unfortunate experiences.

Section	Reviewer, Affiliation	Comment	Response
Results – Behavioral and Other Nondrug Treatments	**TEP Reviewer #3**	For the psychotherapies and psychosocial interventions, it would be helpful to know if there are differences between individual and group therapies and between in-person and non-face-to-face delivery methods for the same type of intervention when studies of more than one method are available. Individual and group interventions have many different features and face-to-face interventions may have different characteristics and possible modes of action than internet, phone based or telemedicine interventions.	Thank you for the comment. We did not group studies because of variation in components, modality (group vs. individual, in-person vs. not in person), duration, and comparators. In addition, some studies offered a combination of modalities in the intervention and comparator arms. Due to this variation, comparisons of different modalities including individual versus group studies are questionable. We agree with the statement that group and individual interventions may have different features. We have outlined the components and modality of each intervention and comparator in the population and inclusion criteria tables.
Results – Behavioral and Other Nondrug Treatments	**TEP Reviewer #3**	Family and partner interventions are grouped together in the analysis (though currently separated from other family-focused therapy). However, it seems conceivable that partner based interventions may be associated with different outcomes than family based treatments, particularly related to one's family of origin. If any distinctions between family and partner interventions are possible, it would be helpful to note them.	Thank you for the comment. We did not separate outcomes by family versus partner interventions because the included studies often allowed participants to use either a family member or a partner. Study inclusion criterion specified that participants should have at least one close family member, significant other, or care provider willing to participate in the psychosocial intervention. This could include spouses, parents, siblings, close friends, and extended family.

Section	Reviewer, Affiliation	Comment	Response
Results – Behavioral and Other Nondrug Treatments	**TEP Reviewer #4**	With respect to the behavioral interventions, it was not clear to me why the STEP-BD trial was included in the section labeled, "other behavioral interventions." This is one of the only behavioral intervention studies focused on acute treatment of bipolar depression, and certainly the largest to date. Although this was a multiple-arm study including CBT, IPSRT, FFT, and a time-matched active comparator, the data appear to be more relevant to the prior sections focused on each of the individual treatments themselves (e.g., the CBT outcomes could reasonably be incorporated into the CBT-focused analyses, the IPSRT outcomes could reasonably be incorporated into the IPSRT-focused analyses, etc.).	Thank you. The analysis of STEP-BD groups the three intervention arms as "intensive psychosocial treatments." Relapse results are presented as "intensive psychosocial treatments" versus collaborative care as well as the individual arms versus collaborative care. However, functional outcomes are only presented as the aggregate intervention category versus comparators. While differences from baseline are provided, we were unable to calculate effect sizes by arm (i.e., for each intensive psychosocial treatment separately) for this outcome because the publication does not provide sample sizes by individual intervention arm for this sub-analysis. Due to this discrepancy, we opted to report the results of STEP-BD together in one section for a complete presentation of study findings in the "other interventions section."
Results – Behavioral and Other Nondrug Treatments	**Public Comment Anonymous**	Therapy is of limited value if half of his brain cells are not firing.	Thank you for sharing your personal story. We are sorry to hear of your unfortunate experiences.

Section	Reviewer, Affiliation	Comment	Response
Discussion	**TEP Reviewer #1**	Overall, very helpful; if room the authors should elaborate more on practical suggestions for conducting studies in this challenging and vulnerable population, e.g., best practices in patient follow-up, involving peer supports, use of subject incentives, etc.	Thank you for the comment. We have not extended the text on further specific suggestions due in part to keep the report short, and in part because we believe this is partly an empirical question in its own right.
Discussion	**TEP Reviewer #3**	The limitations of the literature are described, although in some cases the clarity of the discussion could be improved. The implications of the major findings are unsatisfying since the primary focus is on the lack of useful information from a huge number of studies over many years of research. The future research section seemed vague, non-specific and did not mention most of the research gaps that were identified by this review.	Thank you for the comment. We have revised the limitations and future research sections t. As stated in the report, since only low-strength evidence was reached for benefit or no difference between groups for any treatment, drug or psychosocial, essentially all key questions would benefit from further research. To delve into our opinions regarding specifics for all the possible research questions would have led to an unbalanced presentation of the report.
Discussion	**TEP Reviewer #4**	Findings and implications are clearly and succinctly summarized in the Discussion. In addition to sufficiently addressing the major limitations of the literature (e.g., attrition, complexities of including cases across BD subtypes, comoribidities, special populations, etc.), it would have been informative for the authors to provide some summary re: new treatments under development that may not have been included in this review but could benefit from these considerations as they are being developed.	Thank you for the comment. This was outside the scope of our review.

Discussion	**Peer Reviewer #5**	I have several comments about the Sachs et al (2007) trial of bipolar depression treatment in the context of the STEP-BD study. The judgment of "high risk of bias" is surprising, given that not only was this Government- and not industry-funded, but it included two different antidepressants, presented in complete (apparently unbiased) equipoise. The results are somewhat unique, as to some extent the key message is independent of the two antidepressant drugs. And that is that for many patients, adequate mood stabilizer medication obviated the need for ANY antidepressant. Indeed, this was one of the lessons of the STEP-BD study as a whole: that existing treatments are generally good, provided they are used as prescribed or recommended. That the relapse rate into either mania or depression was lower than expected in STEP-BD may have been due to the overall good adherence to treatment. Although STEP-BD was designed to be fairly "naturalistic", there are some aspects of participation in a formal clinical trials that will differ from routine clinical practice. Among other factors, the structure of the clinical trial, reminder messages for appointments, and the desire to help others, including future generations, with new treatment approaches, can all contribute to a higher rate of adherence than might be seen in "real-world" clinical settings. Future studies should utilize the growing number of approaches to optimizing adherence, including motivational interviewing and other psychosocial interventions as well as "smart" bottles and smartphone apps and, beginning with aripiperazole, a "smart pill" that signals when swallowed. Back to the Sachs et al (2007) article specifically, a subsequent independent reanalysis of the data by Wu et al. (2015) found that "the estimated optimal dynamic treatment regime (DTR) for bipolar depression constructed from the STEP-BD study suggests the hypothesis that standard antidepressants should not be used to supplement mood stabilizers for patients with a prior hypomanic episode." REFERENCE: Wu F, Laber EB, Lipkovich IA, Severus E. Who will benefit from antidepressants in the acute treatment of bipolar depression? A reanalysis of the STEP-BD study by Sachs et al. 2007, using Q-learning. International Journal of Bipolar Disorders 2015; 3:7-17.	Thank you for the comments. Unfortunately, while using a very interesting study design, the STEP-BD study suffered the same challenges with high attrition as many of the other studies, making drawing conclusions difficult. The finding that the evidence was insufficient to address the question would remain if the study were given a moderate risk of bias because of the unknown consistency and imprecision of the data. The modeling by Wu et al. is also interesting, but modeling provides at best an indirect approach to the question (since it is an estimate or simulation, not a direct observation of the phenomenon in question). We have included some of your suggestions for future studies in the Future Research section of the report.
Discussion	**Peer Reviewer #5**	I wonder if the increasingly-available "relative citation ratio" could be applied to some of the key references, as a rough index of their	Thank you for the comment. The relative citation ratio is an

Section	Reviewer, Affiliation	Comment	Response
		influence on the field. One would assume that an article such as that by Sachs et al (2007) in the New England Journal of Medicine would be more widely read and carry more weight than a clinical trial report in a more specialized or obscure journal. The present review does a good job of noting funding sources of included studies. While of course, overgeneralization should be avoided, those industry-funded trials which were conducted with the clear intent of securing FDA approval for a given indication – as opposed to advancing knowledge per se – often have a "quick and dirty" feel and typically aim to meet FDA requirements, e.g. minimal acceptable length of follow-up, rather than advancing scientific or clinical understanding.	interesting idea. However, "influence on the field" is a different concept from strength of evidence. Influence can arise from evidence (whether good or bad evidence) as well as charisma, personal authority (a nobel prize winner, for instance), "buzz" or "going viral" through social connections, and other non-scientific factors. Strength of evidence is based on the Bradford Hill criteria for inferring causality from a relationship. Strength of evidence frameworks were created in part to counteract such non-scientific influences. While there is a lot of interest in bibliometrics like Altmetrics, perhaps the number of tweets isn't what we want to focus on either.
Discussion	**Peer Reviewer #5**	The last paragraph on page 101, about the "nonspecific" benefits of behavioral interventions is excellent and might be underscored and amplified in the discussion. With the possible exception of treatment of acute depressive episodes, most behavioral interventions for people with bipolar disorder are designed to be used in concert with other - generally pharmacologic - treatments, and not stand on their own as complete treatments of the syndrome. So perhaps it is unrealistic to look too closely at "effects on symptoms" of psychosocial and behavioral interventions in isolation. As noted on page 101, beyond simply augmenting medication effects, behavioral interventions can enhance adherence to treatment, reduce family friction, and promote hopefulness in patients and their families and friends.	Thank you for the comment. We agree, and have used some of your text here to add a paragraph to the Future Research section.

Section	Reviewer, Affiliation	Comment	Response
Discussion	**Peer Reviewer #6**	In my opinion, the suggestions the authors offer for enhancing retention in clinical trials – "multiple secondary contacts who do not live with participant and all inclusive contact information from cell phones, email, to social media; flexible scheduling outside of business hours, availability at the last minute notice" – are unlikely to be of major benefit. Subjects who enroll in trials generally provide contact information, and if they don't return for study visits and don't respond to attempts to contact them, these are purposeful decisions. Repeatedly attempting to contact them via other means, particularly via family and friends, is unlikely to work and runs the risk of being intrusive and even harassing. Likewise, social media as a contact tool is problematic. It is considered unethical for clinicians to contact patients via social media for confidentiality reasons, and I suspect the same is true for researchers contacting subjects enrolled in clinical trials of treatments for psychiatric illnesses. Finally, flexible scheduling and last-minute availability are generally part of most trials, to the degree that this is possible. The key difficulty for retention in clinical trials in bipolar disorder is most likely that the cardinal symptoms of the illness include fluctuating levels of motivation, activity, outlook (i.e. optimism vs pessimism) and insight, making adherence to the demands of a clinical trial, particularly a lengthy one, a challenge that is not easily addressed.	Thank you for your perspective on the future research suggestions. Given the challenges of this research area, we intended only to seed the conversation and look forward to researchers in the field engaging in discussion and experimenting with different designs to advance the field's ability to conduct research.

Section	Reviewer, Affiliation	Comment	Response
Discussion	**Public Comment Anonymous**	Get this drug off of the market and STOP using patients who need REAL HELP as mere lab rats for your experiments with new "drugs" which can easily ill and do maim their patients. STROKES are not fun to deal with, and for their relatives, as untrained and not willing caretakers, to have to deal with, just because a drug "seems promising" in mental hospital trials where the patients have absolutely NO SAY in whatever drugs are being shoved down their throats by doctors who are using them in such horrible ways. One would think that this is 19th Century England, and the hospitals Bedlam. Not much real "progress" is being made in this ARENA of high stakes drug experimentation!!!!!!!! Drug Marketing is responsible for companies to use unsuspecting patients, completely dependent upon seemingly knolwdgable "doctors" for help. With help like THIS, we might as well use leeches and witchcraft to effect meaningful cures for patients whose "doctors" have NO IDEA about how to REALLY TREAT. Trial and ERROR is not good enough to treat people with these illnesses. FOR SHAME!!!!	Thank you for sharing your personal story. We are sorry to hear of your unfortunate experiences.
Appendixes	**Public Comment Jannsen Pharmaceutical**	E20 Yatham 2003 # Randomized should be 150	Thank you for the comment. We confirmed the printed publication is 151 randomized.
Appendixes	**Public Comment Jannsen Pharmaceutical**	E-24 Risperidone Khanna 2005. Randomization procedures may not be detailed in the publication, however; this DB trial was one of the pivotal trials supporting the bipolar mania indication. “Handling of data from missing persons not described” Refer to figure 1. How efficacy data will be handled for patients who drop out without a final assessment is accounted for in studies.	Thank you for the comment. Completeness of reporting study methods in publications is a challenge, since reviewers cannot make assumptions about study conduct. Given the 20% attrition and the problems with Last Observation Carried Forward (LOCF), if we had assumed randomization was adequate and LOCF was used to address missing data, we would leave the risk of bias rating as moderate.

Section	Reviewer, Affiliation	Comment	Response
Appendixes	**Public Comment Jannsen Pharmaceutical**	I-11 MacFadden 2009-Intervention Dose The injection is given every 2 weeks. In the DB phase the modal dose was 25 mg in 44 patients, 37.5 mg in 18 patients, and 50 mg in 3 patients.	Thank you. We have corrected the modal dose.
Appendixes	**Public Comment Jannsen Pharmaceutical**	I-11 Bobo 2011 The mean dose was 27 mg ± 2 mg every 2 weeks	Thank you. We have corrected the mean dose.
Appendixes	**Public Comment Jannsen Pharmaceutical**	I-17. MacFadden 2009 Rationale "Unknown if rater blinded." Publication describes the blinding in this pivotal trial establishing the effectiveness of risperidone longacting injection as an adjunct to treatment with lithium or valproate for the maintenance treatment of bipolar disorder. For the primary analysis, relapse was determined by an independent relapse monitoring board (RMB). The RMB consisted of three psychiatrists who reviewed, in a blinded fashion, cases of relapse reported by investigators to standardize the assessment of relapse and to support the scientific validity of the study.	Thank you for the information. We reviewed the publication and searched for mention of the relapse monitoring board but were unable to locate such mention. However, with this additional information, the risk of bias assessment would remain unchanged.
General	**TEP Reviewer #1**	This is one of the most comprehensive reports I have read. There is a broad range of discussion regarding the clinical efficacy/effectiveness as well as the real-world barriers to conducting such studies in a difficult-to-reach population.	Thank you for the comment.
General	**TEP Reviewer #1**	Very clearly written report.	Thank you.
General	**TEP Reviewer #3**	The key questions are generally appropriate. In the subquestions of KQ1, the wording "behavioral health treatments" seems problematic although the thrust of the subquestions is accurate. The outcomes were prioritized using patient-centered considerations but some outcomes are more patient-centered than others. The use of the term "patient-centered" in the sub-questions implies that some outcomes may have been excluded simply because they weren't viewed as patient-centered. The broader wording in the overarching key questions seems preferable.	Thank you. The question wording was determined at the time of the topic refinement for this project.

Section	Reviewer, Affiliation	Comment	Response
General	**TEP Reviewer #3**	It is obvious that a great deal of effort went into the creation of this report, including the screening of the literature and the detailed extraction of the information that is presented in the tables. Where meta-analyses were possible, these appear to be well-done. It is clear that the evidence-based practice center brought considerable expertise to this systematic review.	Thank you.

General	**TEP Reviewer #3**	There are several ways in which the organization of the document could be enhanced. First, the titles of chapters 4, 5 and 6 should be renamed as Somatic Therapies for Acute Mania, Acute Depression and Maintenance Therapy, respectively. Each of these chapters could then be organized as follows: • Lithium • Anticonvulsant medications • Antipsychotic medications • Antidepressant medications • Other medications • Other somatic therapies Although several of the anticonvulsants have been used as mood stabilizing medications, lithium is unique in many respects and is better classified separately. In addition, topiramate should be placed with the other anticonvulsant medications and paliperidone with the other antipsychotic medications rather than grouping the medications based solely on their FDA approval status. The discussion of TMS is quite small, but it seems incorrect to group it with non-somatic therapies. (It is not clear that the TMS study meets the inclusion criteria at all since the outcomes for depression were assessed at 4 weeks.) Chapter 7 could then be renamed "Psychosocial Treatments". This would be a better overarching title since interpersonal and social rhythm therapy is not considered a behavioral treatment. The organization of results based upon specific outcomes (e.g., relapse, symptom scores, other outcomes) is consistent with the GRADE approach of rating the body of evidence for each outcome separately but makes it harder for the reader to appreciate the overarching effects of each intervention. Inclusion of additional summary tables may help the reader in synthesizing the variety of information in the review (See examples of summary tables from other reviews at the end of these comments). Quality of life seems to be conceptualized as a subset of functioning, at least in some portions of the draft (e.g., p. 84). However, at least in psychotic disorders, subjective changes in quality of life do not always parallel observed changes in functioning. Thus, it may be preferable to discuss quality of life outcomes separately from functioning outcomes.	Thank you for the suggestions. The report organization is based on feedback from Technical Expert Panel members and follows general AHRQ EPC program guidance. A number of revisions have been made in the report, based on balancing differing viewpoints of review comments and preferences, readability, and detail. We have changed the name of Chapter 7 to used "Psychosocial" rather than "Behavioral", and carried that change throughout the document.

Section	Reviewer, Affiliation	Comment	Response

Section	Reviewer, Affiliation	Comment	Response
General	**TEP Reviewer #3**	In the tables in the main body of the document that describe each of the included studies, it would be helpful to include: • Study design (e.g., RCT, controlled clinical trial) • Funding source • Specific locations of the trials and not just the number of continents that were included, as this may be relevant to the applicability to North American patient populations • Trial duration and duration of active treatment • Number of individuals in each treatment arm • Whether the intervention is an add-on to usual treatment in both arms • Medication doses (fixed dose or average dose, in flexibly dosed studies) • Findings of the study (as presented in Table 48) Although some of this information is included elsewhere in the document, it is cumbersome for readers and guideline developers to need to refer to multiple portions of the document to get an overall impression of each study. Adding hyperlinks within the document would be an improvement but would still be clumsy. The sort order of the tables of studies also seems inconsistent. The layout of the tables in the sections about psychotherapy and psychosocial interventions is not as helpful as it could be. Without knowing what the active comparator is, it is hard to interpret any similarities and differences between the two columns. The jumbling together of multiple outcomes in each cell also makes it hard to read/process. Examples of several summary table layouts are shown at the end of these comments and other viable approaches may also exist.	The body of the report is written to provide a summary that balances readability with detail. Detailed information is available in the Appendixes for all included studies. We have added summary results tables for all intervention comparisons with at least low-strength evidence (in fact, all are low-strength evidence) to each of the major results sections.

Section	Reviewer, Affiliation	Comment	Response
General	**TEP Reviewer #3**	The organization of the report could be improved (see attached comments). Many of the main points could be expressed with greater clarity (see attached comments). The conclusions of the report are not well-suited for incorporation into decisions about the care for individuals with bipolar disorder. If I had bipolar disorder, I would find the findings of this review to be extremely demoralizing and much more demoralizing than I think is warranted by the available evidence. I did not find that this review contributed new information to my understanding of the available research.	Thank you for your comment. As noted by another reviewer, bipolar disorder is a complex phenomenon to study and "the rate limiting step" to many more definitive conclusions was the nature of available evidence. We agree the findings are not satisfying, and would benefit from future research that improves the field.
General	**TEP Reviewer #4**	In general, this is a comprehensive and informative report that provides a critical overview of the bipolar disorders treatment literature. Given the vast heterogeneity of this literature, driven largely by the complexity of bipolar disorder itself (e.g, by subtype, mood episode polarity, acute treatment vs. secondary prevention aims, adherence and attrition challenges, and high rates of mental health, substance use, and physical health comorbidities), it is acknowledged that this review can only accomplish so much. The authors should be commended for fully acknowledging all of the above, while also attempting to operationalize review criteria and outcomes in an effort to synthesize the state of the field. Nevertheless, by attempting to "make order out of chaos," it is noted that this report does not cover other meaningful outcomes that are critical in BD treatment, such as suicide behaviors and treatment adherence.	Thank you for the comment.

Section	Reviewer, Affiliation	Comment	Response
General	**TEP Reviewer #4**	As previously noted in this review, research on interventions for BD is incredibly complex, due to the need to account for disorder subtypes, different symptom polarity outcomes, choice to focus on acute phase treatment, maintenance phase treatment, or secondary prevention (when people are first enrolled during a euthymic state), and a host of other complicating factors. For these reasons, the data are inherently difficult to digest and synthesize. With that said, the current report draft organizes the comparisons of interest in a straightforward and clear manner, and the accompanying tables for each analysis are incredibly helpful in following the results and conclusions. Operationalizing risk of bias (ROB) was also helpful in following conclusions drawn. Although the data do not inspire much optimism re: the status of treatment options for people with BD, they do point to the need for continued research to develop more effective interventions for BD. I believe this conclusion is particularly important as my personal (anecdotal) sense is that the general public believes that existing treatments are much more effective for BD than the data suggest (and this report further indicates). Bringing attention to this gap in our clinical knowledge base is an important step in advancing work in this area.	Thank you for the comments.

Section	Reviewer, Affiliation	Comment	Response
General	**Peer Reviewer #5**	The report is thoughtful, well-researched and organized, with appropriate key questions. It is clinically meaningful as far as the data will permit. The "rate-limiting step" is the nature and quality of the data. Regrettably, it appears that a disproportionate of the literature over the past 25 years has consisted of industry-funded clinical trials with a narrow focus (typically a new FDA indication), with a dearth of adequate comparative effectiveness trials or studies with a long enough follow-up period to meet the reasonable parameters set in this report. So we are left with a "glass half-full" situation, whereby standard clinical practice - even guideline-based - seems to be supported by only "low-strength" evidence.	Thank you for the comments.
General	**Peer Reviewer #5**	The report is well-organized and clear. The challenge for the field is that "only low-strength evidence was reached for benefit or no difference between groups for any treatment, drug or behavioral," so it will be hard to use this report to promote any specific change in current practice. On the other hand, the report could serve as a spur for more serious research into controlled clinical trials that go beyond the "bare minimum" evident in much of the reviewed literature.	Thank you for the comment.

Section	Reviewer, Affiliation	Comment	Response
General	**Peer Reviewer #6**	Thank you for asking me to review "Treatment for Bipolar Disorder in Adults: A Systematic Review." The authors are to be commended for the voluminous time and effort they put into preparing the manuscript. It exhaustively documents and critiques the evidence base for the effectiveness of treatments in bipolar disorder. It is highly valuable as a summary of the strengths and weaknesses of clinical trials of treatments in mania, depression, and maintenance treatment. However, the highly stringent criteria for selecting trials for inclusion, along with specific decisions the authors regarding the required duration of trials, and the considerable difficulties inherent in conducting clinical trials in psychiatric patients, make it less useful to clinicians, health systems, payors, and policy makers. This is particularly true in the depression section, where the decision to only include trials of at least 3 months duration (despite including mania trials of only 3 weeks duration) means that many effective treatments were not evaluated. Also, in the maintenance section, for reasons that are not clear to me, the authors' conclusions directly contradict the reported results of some of the trials. In both cases, this creates the mistaken impression that there are few if any effective treatments for depression and maintenance treatment in bipolar disorder. This could have a highly negative impact on clinical practice, health policy, and insurance coverage of many medications. I hope the authors will consider making some revisions to these two sections.	Thank you for the comment. This review process, using AHRQ methods guidance and informed by discussions with key informants and technical experts, is intended to be rigorous and hold research to standards that can support causal statements. The findings from the review are not evidence of lack, that is, that nothing works. Rather, the findings are of lack of evidence, that is, the available evidence is insufficient to draw conclusions. This finding, then, suggests that one of the 3 legs of evidence-based medicine, evidence, is missing, leaving the 2 legs of clinical and patient experience to inform decisions. Thus, as we note in the discussion chapter, more research is needed. It is our hope this report will help the field improve research conduct so that these questions might be answered in the future.

Section	Reviewer, Affiliation	Comment	Response
General	**Peer Reviewer #6**	The report is well-structured and clearly written. However, the highly stringent criteria for selecting trials for inclusion, along with specific decisions the authors regarding the required duration of trials, and the considerable difficulties inherent in conducting clinical trials in psychiatric patients, make it less useful to clinicians, health systems, payors, and policy makers. This is particularly true in the depression section, where the decision to only include trials of at least 3 months duration (despite including mania trials of only 3 weeks duration) means that many effective treatments were not evaluated. Also, in the maintenance section, for reasons that are not clear to me, the authors' conclusions directly contradict the reported results of some of the trials. In both cases, this creates the mistaken impression that there are few if any effective treatments for depression and maintenance treatment in bipolar disorder. This could have a highly negative impact on clinical practice, health policy, and insurance coverage of many medications. I hope the authors will consider making some revisions to these two sections.	Thank you for the comment. This review process, using AHRQ methods guidance and informed by discussions with key informants and technical experts, is intended to be rigorous and hold research to standards that can support causal statements. The findings from the review are not evidence of lack, that is, that nothing works. Rather, the findings are of lack of evidence, that is, the available evidence is insufficient to draw conclusions. This finding, then, suggests that one of the 3 legs of evidence-based medicine, evidence, is missing, leaving the 2 legs of clinical and patient experience to inform decisions. Thus, as we note in the discussion chapter, more research is needed. It is our hope this report will help the field improve research conduct so that these questions might be answered in the future.