Comparative Effectiveness Research Review Disposition of Comments Report

Research Review Title: Terbutaline Pump for the Prevention of Preterm Birth

Draft review available for public comment from February 01, 2011 to March 1, 2011.

Research Review Citation: Gaudet, L., Singh, K., Weeks L., Skidmore, B., Tsouros, S., Tsertsvadze, A., Daniel, R., Doucette, S., Walker, M., Ansari, M.T. Terbutaline Pump for the Prevention of Preterm Birth. Comparative Effectiveness Review No. 35. (Prepared by the University of Ottawa Evidence-based Practice Center under Contract No. HHSA290-2007-10059-I.) AHRQ Publication No. 11-EHC068-EF. Rockville, MD: Agency for Healthcare Research and Quality. September 2011. Available at: www.effectivehealthcare.ahrq.gov/reports/final.cfm.

Comments to Research Review

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

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

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

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer 1	Executive Summary	What is meant by difference sin “means” – was this weeks?	Weeks
Peer Reviewer 5	Introduction	" Unfortunately, digital examination of the cervix in early labor is not highly reproducible, adding to the limitations of preterm labor diagnosis.26-28" Cervical exams have been shown to be reproducible by many investigators. The three studies cited to support this statement compare physcial exams to ultrasound- and thus do NOT address the reproducibility of the cervical exam itself. Most of the well designed studies in this review used cervical exams to diagnosed preterm labor, and their data should not be discounted because of a misunderstanding about the accuracy of the cervical exam.	Given that this appears to be controversial, we’ve elected to remove the statement altogether.
Peer Reviewer 5	Introduction	*"The diagnosis of preterm labor is made based on contraction frequency of ≥6 per hour and cervical dilation ≥3 cm and/or effacement ≥80 percent, or if membranes rupture or bleeding occurs, is reasonably accurate.24,25"* Most clinicians view any cervical change accompanied by regular contractions to indicate preterm labor, and try to intervene before the cervix reaches 3 cm or 80% effacement. Conversely, vaginal bleeding alone is not diagnostic of preterm labor .	Amended
Peer Reviewer 4	Introduction	Over the last 20 years the preterm birth rate in the U. S. has risen, but in 2007 and 2008 there was modest decrease in the preterm birth rate (Martin5) and the rates for 2009, which will be available this spring, supports that they are continuing to fall. Also important, is the reason for preterm birth. Ananth et al6, in viewing over 17 million births from 1990-99 found that while deliveries <37 weeks continued to climb during that decade, it was due to medically indicated deliveries and attempted to salvage very small babies as well as late preterm births. The contribution of preterm labor and rupture membranes, most often with preterm labor, actually fell between 20 and 30%. This is important data because, in the case of medical complications of pregnancy, one has to perform a preterm delivery, whereas in appropriate patients tocolytics and other means of prolonging pregnancy appeared to yield a positive effect.	Valid point, taken. Suggested information added with new references
Peer Reviewer 1	Introduction	Could update national preterm delivery prevalence - now 12.3%. Note, the prevalence climbed throughout the period of peak usage and is declining during a period of waning use. The authors may want	Prevalence data revised. Previously, we reviewed the literature related

		to make reference to the recent FDA warning, and to recent associations between SQ terbutaline pump therapy and the occurrence of autism and other adverse long-term childhood neurodevelopment sequelae as well as concerning animal studies on brain development.	to autism spectrum disorder and determined that this outcome was more appropriately related to terbutaline as a drug, and not the terbutaline pump per se, which was the focus of our review. However, we did look for studies investigating long-term childhood outcomes (e.g. developmental) in comparative studies. In other words, while not of particular interest to the review, we can safely say that the autism was not investigated in comparative effectiveness primary studies investigating terbutaline pump therapy.
Peer Reviewer 3	Introduction	Some discussion would be helpful regarding how commonly this therapy is used in clinical care	We don't know, we couldn't locate any numbers.
Peer Reviewer 1	Methods (Study Section)	The pump was commonly used among women less than 24 weeks, but no studies address this.	Noted
Peer Reviewer 5	Methods	Case series with obvious selection bias should not have been included 2. The studies using Matria data should not have been included, because of selection bias, and because many of those patients were likely to have been included in other papers, thus the potential for double counting patients is high. 3. Studies in which the diagostic criteria for preterm labor were not stated or were inadequate (ie preterm contractions only, without documented cervical change) should have been excluded, because of the likelihood that many of the study subjects were not in preterm labor.	For rare harms, even case reports may be included in CERs.[1] Studies of limited internal and external validity may be included in CERs. That is the whole point of assessment of study risk of bias and applicability. We graded the strength of evidence using standard methodology and took into account the quality/risk of bias (ROB) of studies.
Peer Reviewer 1	Methods	The initial therapy may confound efficacy since those who received a primary tocolysis with a beta-adrenergic agonist such like IV or s.q. terbutaline or IV ritodrine may have had downregulation of uterine myometrial beta-adrenergic receptors and thus, less efficacy. Conversely, betamethasone and other corticosteroids up-regulates such receptors, so represents another potential source of confounding.	Differences in primary tocolytic therapy were assessed as potential confounders. In the Methods section, under risk of bias, we stated "Similarity of groups in terms of administration of primary tocolytic regimen to control acute episodes of preterm labor" We also collected data about corticosteroid

[1] Helfand M, Balshem H. AHRQ series paper 2: principles for developing guidance: AHRQ and the effective health-care program. J Clin Epidemiol 2010 May;63(5):484-90. [PMID: 19716268].

Source: www.effectivehealthcare.ahrq.gov

			use.
Peer Reviewer 6	Methods	The potential for bias against publication of negative studies should be considered and discussed.	Publication bias was not investigated because: • We took an extensive grey literature search, and requested relevant scientific information from the industry • We had few studies per outcome for an statistical assessment of publication bias • Any exaggerated positive findings could have been due to the medium to high risk of bias detected in observational studies instead of publication bias. Publication bias is not an important concern of ours with respect to this CER. This is because we did searched for grey literature, scientific information packets from the industry, and had many experts in this field participate as Key Informants/TEP/Peer reviewers – and none of these experts has indicated that there are additional unpublished studies out there. We have seen both Matria based studies with positive results and RCTs failing to demonstrate efficacy. As we had few studies per outcome, publication bias was not statistically investigated. Also, exaggerated positive findings were likely due to the medium to high risk of bias detected in observational studies rather than publication bias.
Peer Reviewer 1	Methods (Data Synthesis and Analysis)	Given discordance in pathogeneses between these two groups some effort should have been made to compare the relative efficacy of this therapy among whites vs. African-Americans.	Assessment of outcomes by racial or ethnic subgroups was an essential item incorporated in the key questions. However, none of the included studies presented information for racial or ethnic subgroups.
Peer Reviewer 1	Results (Overview of Findings)	Table 6 (pg. 16). Mean cervical length in study 21 is reported as 0.2 cm. Is this sonographic cervical length? This seems VERY short to be the mean. Please verify.	The paper does not specify if this is sonographic cervical length. Exact quote from paper: “At the start of subcutaneous

Source: www.effectivehealthcare.ahrq.gov

			terbutaline pump therapy, the mean cervical dilation was 1.7 ± 0.7 cm, and the mean cervical length was 0.2 ± 0.2 cm."
Peer Reviewer 3	Results	If fetal fibronectin status is not reported in any of the studies, this deserves specific mention, given the lower frequency of preterm birth in studies of preterm labor in which the definition of preterm labor does not require fetal fibronectin positivity. This lack of inclusion of sonographic cervical length or fibronectin among the inclusion criteria of a study of women ith preterm labor is a major limitation and ought to be emphasized in this section.	Please read under subheading "Population": "No studies presented data on concomitant medications, body mass index (BMI), history of preeclampsia, cervical position, cervical consistency, cervical station, Bishop's Score, or fetal fibronectin" We now emphasize this point in Future research.
Peer Reviewer 4	Results	First, two flawed RCT's (Wenstrom1, Guinn2) were included in the process and should be removed. See comment box. 1) Patients were not allowed to alter the amount of bolus terbutaline based on end-organ response (uterine contraction data) because the protocol forbid contraction monitoring. Therefore, bolus terbutaline which is critical in preventing recurrent preterm labor, could not be adjusted because there was no assessment of contractions. (2) Also, the investigative protocol gave every patient the same basal amount of drug without pharmacologic consultation regarding volume distribution, body mass and renal clearance (required to determine the appropriate basal dose of terbutaline infusion). Therefore, all women (BMI 25-42) received the same amount of basal terbutaline. (3) Women were likewise not approved to have a daily contact with a nurse by phone, nor did they have access to 24 hour/d/7/wk emergency contact, although this is how SCT is used by every physician in clinical practice.What was reported in the RCTs	We went back and looked at the trials again. We disagree that the trials are so flawed that they shouldn't have been included. **Guinn et al**: (1) Bolus amount could not be changed, but women could administer more boluses if uterine contractions developed. "Before activation the pumps were programmed by the research nurses. The programming capabilities were then suspended to prevent patients from manipulating the terbutaline dose." "In addition, the patient herself could administer 0.25 mL twice daily if increased uterine activity developed." (2) Correct, but BMI not reported. "Pump therapy was initiated according to the following protocol: continuous infusion of 0.05 mL/h with scheduled bolus injections…" (3) Incorrect "Nursing support was available 24 h/d to answer questions for all participants and to help monitor therapy." **Wenstrom et al**: (1) Bolus amount could not be changed, but the number of boluses were adjusted according to contraction pattern. "Pump solution…administered at a rate of 0.05 ml (mg) per hour plus 0.25 ml (mg) boluses every 6 hours. The basal rate was

Source: www.effectivehealthcare.ahrq.gov

			kept constant but the number and timing of boluses were adjusted according to the patients' unique contraction pattern." (2) Correct, but BMI not reported. "Pump solution…administered at a rate of 0.05 ml (mg) per hour plus 0.25 ml (mg) boluses every 6 hours. The basal rate was kept constant…" (3) Seems like patients had phone contact with nurse "Patients in all three groups made phone calls to the research nurse with equal frequency."
Peer Reviewer 4	Results	The likelihood of double-counting, while a possibility, is almost impossible as the studies discussed occurred during different years with little overlap and came from different areas of the country (California vs New York vs Florida) and in many cases included the patients of only one group of physicians.	None of the Matria-based studies have reported the geographic regions from which subjects were recruited. Also, some have not reported the years over which subject recruitment occurred. We have requested this information from Matria.
Peer Reviewer 4	Results	RoB: I believe the authors should spell out this section in more detail. It appears that the major flaw (which classified a study as being one with a high risk of bias) were imbalances in the baseline characteristics or prognostic factors, I would spell those out. In assessing Figure 5 (page 18) is confusing as I read for example, these nine cohort studies in detail and I would rate all of them as all of them having a low chance of bias. While the data speaks for itself, all of them had different methods and different cohort groups which inherently make them hard to compare. I would hardly call this bias. Similar comments can be made about the two case series, the cohort study, and the non-comparator trials. One of the more important aspects I believe is funding source, and only the randomized clinical trials had evidence of funding bias not the cohort or studies which lacked a comparator.	We believe our presentation of ROB assessment and details are standard. For further details of ROB assessment, we refer the reviewer to the ROB Evidence Table in appendix F (Table F3). The suggestion that our risk of bias assessment is flawed was not supported by specific criticism and rationale. We request that the reviewer examines Table F3 and explicitly identify any erroneous judgment on our part. We shall happily revise our ROB assessment if necessary.
Peer Reviewer 4	Results	RoB: I would also have composite data for the non-comparator studies divided into two parts. First, Elliott3 and Perry4 both deal with adverse effects, have large numbers (9357 and 8709), (respectively) and speak strongest to the lack of adverse cardiac side effects with this small daily dosage of terbutaline (3mg/day) compared to oral (40-60mg/d) and parenteral (60-90mg/d). Both of these studies which have a low risk of bias and do not involve double counting. (see comment box for details)	Except for rare harms of pump malfunction, noncomparative studies are not helpful in CERs since they do not compare between the intervention and control. Moreover, the reviewer does not express a scientific rationale for including non-comparator studies except that they are large and "perhaps" at low risk of bias. But these two reasons alone are not sufficient enough to meet other eligibility criteria.

Source: www.effectivehealthcare.ahrq.gov

Peer Reviewer 6	Results	Authors have included studies of high bias, which I would have considered excluding. Inclusion of these gives credence to results which the authors believe have high potential to be inaccurate or misleading. The risk of bias among the included studies overall is moderate to high. Results from an analysis of such studies should be regarded cautiously and with skepticism. The level of detail in the analyses is appropriate.	For rare harms, even case reports may be included in CERs.[2] Studies of limited internal and external validity may be included in CERs. That is the whole point of assessment of study risk of bias and applicability. We graded the strength of evidence taking into account the quality/ROB of studies. Conventionally, systematic reviews adopt a priori eligibility criteria for study designs. All studies based on those designs are included and assessed for risk of bias. Sensitivity/subgroup analyses may focus on higher quality studies. We had no studies that were rated as low risk of bias or high quality for that study design.
Peer Reviewer 4	KQ1 (Neonatal Health Outcomes)	I am surprised neonatal death showed data favoring SCT since loss in the first 28 days of life is rare. Since it was positive, I cannot understand why it was "low strength of evidence". Is there any way that the non-comparator studies (D14-15) can be of help in this area? It would seem from the data on Table 8 that almost 1000 patients (singletons and twins) give comfort to the fact that neonatal death is lower in the SCT group than other comparators. Finally, it is unlikely given the small numbers of deliveries at gestational ages below 28-30 weeks that one would see any difference in NEC, ROP, sepsis, stillbirth, or IVH. Perhaps adding the other studies even in a separate category may be helpful.	Grading the strength of evidence from SINGULAR observational studies of medium risk of bias only very exceptionally can raise our confidence above low (or insufficient) – And that is only when the study (ies) is of large sample size, is rated as low risk of bias and demonstrates a very large effect size for a clinical outcome. We stand by our grading of the strength of evidence. Please read the paper by Owens at al we referenced for the methodology we followed to grade the strength of evidence.
Peer Reviewer 4	KQ2 (Surrogate Outcomes)	Page 32, line 42-49. The Cochrane review as well as other papers such as ACOG and the Hayes Brief included the RCT (Wenstrom, Guinn) which, due to their small populations (only 94 who were randomized to the SCT group), demonstrate why there could not yield positive results. These should be removed here and throughout	To the best of our knowledge, high likelihood of type II error is not reason to exclude studies from systematic reviews. Evidence syntheses aim to increase power by including all eligible studies.

[2] Helfand M, Balshem H. AHRQ series paper 2: principles for developing guidance: AHRQ and the effective health-care program. J Clin Epidemiol 2010 May;63(5):484-90. [PMID: 19716268].

Source: www.effectivehealthcare.ahrq.gov

		the manuscript.	
Peer Reviewer 4		I do not see why the studies on surrogate outcome are rated as “low” on strength of evidence.	Owens paper we referred to earlier (and in the report) is self explanatory. Surrogacy of outcomes is one of several reasons limiting the strength of evidence.
Peer Reviewer 1	KQ3 (Maternal Harms)	The occurrence of gestational diabetes should be reported when comparators were calcium channel blockers or placebo.	We evaluated the outcome of maternal hyperglycemia for all studies irrespective of comparators. Only two studies both employing oral terbutaline as comparator reported this outcome.
Peer Reviewer 1		This section, in my opinion, must include the Perry and Elliott study at a minimum, but I would also include maternal complications from the studies without a comparator group if they address side effects. The data will be more robust and whether a complication is present or not, the SCT was used in the same way in the same patients with the exception of two randomized clinical trials. Another comment in this area includes the study15 and it was rated “a high risk of bias” because the groups were unbalanced for risk factors of preterm birth, primary tocolytic therapy and level of care. It should be pointed out that patients in the SCT group were at greater risk for preterm birth, but all women had the same primary tocolytic therapy and level of care. This again underscores my concern that the expressed statistical suspicion of bias, however small, will unduly color the results in a way that perception, not reality, will hinder the clinician in effort to use the data. The pulmonary edema comments (Results, page 68-69) are concerning as it is said those studies were likely underpowered when there were over 1200 cases and the incidence of pulmonary edema was certainly less than noted in normal patients (one to two per thousand). Likewise, “therapy discontinuation” is confusing. It is very rare when the pump is appropriately used (the way every clinician uses it), there is very little discontinuation, if any.	Our a priori criteria allowed non-comparative studies only for rare harms of pump malfunction –a non-comparative outcome, because other tocolytics are not administered via a pump. Since we could not compare gestational diabetes between terbutaline by pump and other tocolytics from non-comparative studies, they were not eligible for this outcome. Please see Table F3 for transparent details of how and why we rated Morrison et al. 2003 cohort study as of high risk of bias. We rated it as such because an important flaw was that the groups were not comparable (with respect to risk factors for preterm birth, primary tocolytic therapy, and level of care). The two studies by Lam et al, under Pulmonary oedema, were judged as underpowered not because of the sample size...but event rate. For an incidence rate of 1-2/1000...the studies were inadequately powered.
Peer Reviewer 4	KQ4 (Neonatal Harms)	Significant information about the safety of SCT infusion was omitted (Elliott3, Perry4)	These were non-comparative studies precluding comparative assessment of harms

Peer Reviewer 4		It must be remembered that neonatal hypoglycemia and hypocalcemia occur in untreated patients more than they were reported in studies with SCT. Likewise, none of the studies with thousands of patients receiving SCT reported a high incidence of maternal hyperglycemia to levels which neonatal hypoglycemia would even be expected. If it has only been noted in a few case reports, it should not infer that neonatal hypoglycemia, hypocalcemia or ileus happens commonly; it is very rare.	We do not think we have inferred anywhere that these are common harms outcomes. We reported that data are sparse and not powered to show a difference.
Peer Reviewer 4	KQ5 (Levels of Activity and Care)	Maternal activity is never reported in any study as almost all of these patients would have had levels of low, medium or high activity at different times during the pregnancy. These levels also vary necessarily with regards to their cervical status and number of contractions. The level of maternity care however (excluding the two randomized clinical trials) were always reported, the patients were called 7 days per week, there was two hours of uterine contraction monitoring and there was 24/7 availability of telephone consultation with a nurse. In cohort groups the patients necessarily do not receive this level of maternal care. The studies lacking a comparator group, however, were all consistent in their use of SCT, which is another good reason to include them.	Noted.
Peer Reviewer 4	KQ6 (incidence of pump Failure)	Lastly, many of the articles with good data on efficacy (and safety) were not analyzed due to the lack of a comparator (D14, 15). It may be true that they cannot be analyzed with the works which form the basis of this report, but if not perhaps, a separate section on these studies would be very comforting to the reader as it would add further support that SCT is safe and effective for maintenance tocolysis. As it stands the paper would not be helpful to clinicians. While the target audience is well defined, the key questions overreach the data in several areas.	As noted above, our review protocol did not consider non-comparative studies except for rare pump malfunction outcomes.
Peer Reviewer 4		Similarly to the question on neonatal harm the incidence of mechanical complications appears not to have enough data to discuss. Therefore, while we would like to know we simply do not have enough information to comment on it. Lam's earlier studies from the late '80s note a few pump problems, his later studies did not so perhaps there is a learning curve. Overall, if you leave this section in, I would simply call it pump malfunction, rather than failure. If the batteries stop working, certainly the pump could have missed dosages. There is not a possibility of an overdose, but even if the total amount (3mg), were injected, it would not cause a problem as we often administered 5-10mg IV for acute tocolysis). When the catheter is dislodged, it is immediately known by the patient because the fluid flows on her skin and clothing. This is corrected immediately	We think pump malfunction is actually a pump failure.

Source: www.effectivehealthcare.ahrq.gov

		by reinsertion of the line back into the subcutaneous catheter.	
Peer Reviewer 1	Discussion/ Conclusion	The major public health issue with this therapy is the occurrence of associated rare serious maternal cardiovascular events. Perhaps partnering with the FDA to provide an estimate of both the numerator of such events and the denominator of exposures would strength the presentation. The second issue of nascent, but growing, concern is the possible linkage of such therapy with autism. Some discussion of this - still tenuous - link should be included in the discussion.	We have requested harms data from the FDA. No information has been received as of April 8, 2011. Previously, we reviewed the literature related to autism spectrum disorder and determined that this outcome was more appropriately related to terbutaline as a drug, and not the terbutaline pump per se, which was the focus of our review. However, we did look for studies investigating long-term childhood outcomes (e.g. developmental) in comparative studies. In other words, while not of particular interest to the review, we can safely say that the autism was not investigated in comparative effectiveness primary studies investigating terbutaline pump therapy.
Peer Reviewer 2	Discussion/ Conclusion	Evidence does not support use outside of clinical studies. I am concerned that people will come away with the message that the therapy maybe beneficial and in the absense of more data continue to prescribe it. The biggest limitation is no comments or data regarding cost included.	Traditionally, AHRQ reports do not evaluate cost-effectiveness of interventions. During the topic refinement discussion, we agreed with the Key Informants that cost-effectiveness analysis may be a future recommendation if and when conclusive findings are generated in ensuing systematic review.
Peer Reviewer 2	Discussion/ Conclusion	I would like to see a comment added that its continued use should be restricted to women who are participating in appropriately designed research studies similar to the verbage used regarding multiple courses of corticosteroids.	Systematic reviews are not supposed to make recommendations.
Peer Reviewer 3	Discussion/ Conclusion	If fetal fibronectin status is not reported in any of the studies, this deserves specific mention, given the lower frequency of preterm birth in studies of preterm labor in which the definition of preterm labor does not require fetal fibronectin positivity. This lack of inclusion of sonographic cervical length or fibronectin among the inclusion criteria of a study of woenw ith preterm labor is a major limitation and ought to be emphasized in the Discussion. Inclusion of women at low risk for imminent preterm birth tends to bias towards the null in studies of	Included stronger wording

Source: www.effectivehealthcare.ahrq.gov

		tocolytic therapy.	
Peer Reviewer 3	Discussion/ Conclusion	The contribution of data from a company with a profit-motive behind demonstration of effectivenss is concerning. The report, in general, does a good job of distinguishing areas where Matria (Alere) data differ from other sources. The Discussion ought to emphasize this point more strongly.	Addressed.
Peer Reviewer 5	Discussion/ Conclusion	The main difference between this report and the existing systematic reviews is that this report is muddied by the inclusion of uncontrolled case series, papers with obvious patient selection bias, studies that likely enrolled patients who werent really in preterm labor, and reports that included patients who were already included in other studies. By considering data from these less that ideal studies, the conclusion of this review is made less emphatic than that of the other reviews; this review states that "Although there is some evidence favoring SQ terbutaline pump therapy, our confidence in the evidence is low" , while other reviews state that terbutaline pump therapy does not decrease the risk of preterm birth. Thus it is not clear how the current review will be used to inform policy or practice decisions.	These (methodological and clinical) limitations are incorporated in grading the strength of evidence and summarizing applicability of findings (see the methods subsection “Strength of Evidence and Applicability”. Uncontrolled case series were included only for the outcomes of pump failure.
Peer Reviewer 5	Discussion/ Conclusion	The data indicate what has been known for at least a decade- prolonged terbulatine ptherapy does not improve pregnancy outcome regardless of how it is administered, and is associated with very serious maternal and fetal morbidities. In view of this and the recent FDA bulletin, no further research on terbutaline as a tocolytic is warranted.	We think such a conclusion can only be made if we had very precise estimates around the null for most clinical efficacy outcomes with/without negative harms outcomes. We suspect type II error cannot be ruled out, at least in some subgroups, and as such, cannot conclude as proposed.
Peer Reviewer 4	Discussion/ Conclusion	The positive points can be underscored, however by using the efficacy studies previously mentioned in the non-comparator group. Whether one analyses them separately or adds them to the analysis is up to the authors. The limitations of the studies are described but after listing such issues as bias and other problems, the overall data are still there and that needs to be stated more strongly. As previously noted, the safety data should be bolstered by including the references previously mentioned (App. A), as I think they are important to the overall assessment by clinicians. Overall, the Discussion should make clear that using all the peer reviewed published data that is available, gestational age at birth, prolongation of pregnancy, birthweight, NICU admission, cost and other surrogate endpoints would lead to strong support to the concept that SCT is	The reviewer is urging inclusion of non-comparative data for all outcomes. Our response with rationale has been clearly stated above. As such, we stand by our conclusions and its associated strength.

Source: www.effectivehealthcare.ahrq.gov

		helpful in the appropriate patient as well as to payors and the patients. I do not agree (page 80) that data on clinical outcomes and maternal harms are sparse (if all the data is used).	
Peer Reviewer 4	Discussion/ Conclusion	As written the conclusions are not helpful to the clinician who will continue to use SCT in appropriate patients, because almost all private payors cover this as a benefit as it saves money on the neonatal side. However, the targeted audience also includes, State Medicaid officers will limit its use among this group of women who paradoxically have the highest risk to deliver preterm.	Our conclusions are based on evidence, its strength and applicability. It would have been helpful had the reviewer pointed to an specific error in our conclusions given the available evidence.
Peer Reviewer 4	Discussion/ Conclusion	....conclusions are flawed as it is currently written	We respectfully disagree. It would help to read exactly where and how the conclusions are erroneous given the available evidence. We have already explained why we think the RCT evidence should be included and not non-comparative for outcomes other than of pump failure.
Peer Reviewer 6	Discussion/ Conclusion	The discussion largely reiterates the findings presented in the results. More discussion of the limits of these data for making conclusions would be helpful. The conclusions lack a discussion of the low quality of studies from which the results were obtained. This significantly limits the ability even limited assertions regarding efficacy. I disagree with the authors' assertion that future observational studies are appropriate. If further studies are to be considered. Placebo controlled studies of adequate power to evaluate outcomes of primary importance (newborn morbidities and mortality) should be conducted.	The discussion section started with first outlining our approach: we first review the major findings pertaining to each key question and the strength of the evidence for the prespecified gradable outcomes .We then present our conclusions, make recommendations for future research, and offer clinical and public health perspectives. We crafted the discussion accordingly. However, we have now added a subsection discussing limitations of included evidence. In the conclusion section when we state the strength of evidence, subsumed within it is the low quality of primary studies – so describing the low strength of evidence and then again discussing the limitations/quality of included studies discussed at length in the results and now summarized in a subsection within the discussion seems redundant and repetitive. We, therefore, eschew modifications to the conclusions section. Where RCT evidence may be lacking for any possible reason, well conducted observational studies might fill the gap. AHRQ guidance upholds the importance of good quality observational studies .

Source: www.effectivehealthcare.ahrq.gov

Peer Reviewer 3	Applicability and Future Research	An important future research arena that ought to be emphasized is the inclusion of pharmacokinetic and pharmacodynamic measures within treatment trials. This will be critical to understand inter-individual differences in effectiveness and toxicity and can aid in avoidance of tachyphylaxis that occurs with beta-agonist therapy.	Thanks. Revised accordingly
Peer Reviewer 4	Applicability and Future Research	The section on future research is much too simplistic. It simply will not happen. No group will look at these outcomes in SCT patients compared to a control group and follow them long-term (2-6 years) to assess childhood outcomes. Such an appropriately powered study (n=1 to 2000) with long-term follow up should cost in the neighborhood of tens of millions if not hundreds million dollars and is not possible today or in the future. I think a better statement would be that there will be no future studies but that the available data supports the use of SCT in appropriate patients.	Feasibility and bias are not related. RCT is the best design and for interventions must always be a recommendation even when it may cost hundreds of millions. Before such RCTs may be conducted, we could recommend alternatives that may be less robust and accurate, but do have the potential to fill the gap with some reasonable level of confidence. Our future research recommendations observe these principles.
Peer Reviewer 5	Applicability and Future Research	Understanding the definition of preterm labor is essential to evaluating the papers chosen for inclusion in this review- according to my cursory review, many of them did not include demonstrated cervical change as a criterion for diagnosis, thus many patients enrolled in these studies (case series, mostly) had preterm contractions but were probably not in labor- hence their failure to deliver preterm was likely not the result of terbulatine pump therapy.	9 of 14 (64%) studies included patients judged to be in true labor with persistent uterine contractions and cervical changes. For others we cannot say that patients were probably not in labor, but rather that it is unclear how labor was determined.
Peer Reviewer 4	Implications	I do not agree that “evidence favoring in the SQ pump therapy over other treatment or no treatment shows low confidence that the evidence reflects a true effect”. This is simply not true if the randomized clinical trials are deleted and the other data is imported.	Firstly, we did not factor in the RCT evidence when we concluded that we have low confidence that the evidence reflects a true effect. This conclusion was based on observational studies only. Secondly, if the reviewer disagrees with our grading of the strength of evidence, could he please identify a domain (e.g., RoB of the body of evidence, consistency etc.) that was incorrectly categorized and why?
Peer Reviewer 3	General	The figures are a bit small.	Figures have been enlarged.