ST waveform analysis vs. cardiotocography alone for intrapartum fetal monitoring: A systematic review and meta-analysis of randomized trials

Running headline: ST analysis for intrapartum fetal monitoring

Ellen Blix[1], Kjetil Gundro Brurberg[2,3], Eirik Reierth[4], Liv Merete Reinar[2], Pål Øian[5,6]

[1]Faculty of Health Sciences, Oslo and Akershus University College of Applied Sciences, Oslo, Norway [2]The Norwegian Knowledge Centre for the Health Services, Oslo, Norway, [3]Centre for Evidence Based Practice, Bergen University College, Bergen, Norway [4]Science and Health Library, University Library, UiT The Arctic University of Norway, Tromsø, Norway, [5]Department of Obstetrics and Gynecology, University Hospital of North Norway, Tromsø, Norway, [6]Department of Clinical Medicine, UiT The Arctic University of Norway, Tromsø, Norway

Corresponding author:

Ellen Blix, Faculty of Health Sciences, Oslo and Akershus University College of Applied Sciences, PO box 4, St Olavs plass, NO-0130 Oslo, Norway. E-mail address: ellen.blix@hioa.no Telephone +47 67236504

Conflict of interest:

The authors state that there are no conflicts of interest in connection with this article.

Keywords: ST waveform analysis, cardiotocography, intrapartum fetal monitoring, fetal electrocardiography, meta-analysis

Abbreviations:

CTG, cardiotocography; MeSH, Medical Subject Headings; NICU, neonatal intensive care unit; GRADE, The Grading of Recommendations Assessment, Development and Evaluation; RCT, randomized controlled trial; STAN, ST waveform analysis; TSA, trial sequential analyses

Key message:

ST waveform analysis is purposed to be better for labor surveillance than conventional CTG. We quantified the efficacy of these two methods, assessing quality of the scientific evidence with the GRADE tool, and determined that evidence is insufficient to justify the use of ST waveform analysis as intrapartum monitoring.

Introduction. ST waveform analysis (STAN) was introduced to reduce metabolic acidosis at birth and avoid unnecessary operative deliveries relative to conventional cardiotocography (CTG). Our objective was to quantify the efficacy of STAN vs. CTG and assess the quality of the evidence by using the GRADE tool. *Material and Methods.* We identified randomized controlled trials (RCTs) through systematic literature searches and assessed included studies for risk of bias. Meta-analyses were performed, calculating pooled risk ratio (RR) or peto odds ratio (OR). We performed post hoc trial sequential analyses for selected outcomes to assess the risk of false-positive results and the need for additional studies. *Results.* Six RCTs were included in the meta-analysis. STAN was not associated with a reduction in operative deliveries due to fetal distress, but we observed a significantly lower rate of metabolic acidosis (peto OR 0.64; 95% confidence interval [CI] 0.46–0.88). Accordingly, 401 women need to be monitored with STAN to prevent one case of metabolic acidosis. No statistically significant effects were observed in other fetal or neonatal outcomes, except from fetal blood sampling (RR 0.59; 95% CI 0.45–0.79) and a minor reduction in the number of operative vaginal deliveries for all indications (RR 0.92; 95% CI 0.86–0.99). The quality of the evidence was high to moderate. *Conclusions.* Absolute effects of STAN were minor, and the clinical significance of the observed reduction in metabolic acidosis is questioned. There is not enough evidence to justify the use of STAN in contemporary obstetrics.

Introduction

Fetal monitoring should identify fetuses at risk of neonatal and long-term injury attributable to asphyxia. The aim is to identify and timely intervene in preventable cases of fetal damage.

Cardiotocography (CTG) was introduced in the 1960s and assumed to prevent fetal asphyxia. The CTG method has low specificity and high false-positive rates; therefore, its introduction into clinical practice was associated with increased incidences of cesarean section and operative vaginal delivery (1). Hence, a test with higher diagnostic accuracy is needed to identify truly hypoxic fetuses.

The ST waveform analysis (STAN) method was introduced after extensive experimental research (2). A lack of fetal oxygen produces changes in the fetal ECG waveform analysis. The method can be used after rupture of membranes in single pregnancies after 36 weeks' gestation, and it is purported that the STAN method (i.e., cardiotocography plus fetal STAN) can reduce metabolic acidosis at birth and avoid unnecessary operative deliveries (2).

Randomized controlled trials (RCTs) are designed to measure the efficacy of interventions and allow causal inferences between treatments and outcomes. By 2015, more than 15 000 women in labor had been randomized to receive either the STAN method or conventional CTG alone in attempts to estimate the efficacy of the STAN method (3-7). All five RCTs were performed in Europe, but they reported different outcomes and conflicting evidence. Five previous systematic reviews and one review article have compared STAN with CTG alone in meta-analyses (8-13). Three of the meta-analyses included all five RCTs (8, 9, 12); two (10, 11) excluded the Westgate trial (3) and one (13) excluded the Vayssière trial (6). A meta-analysis from 2012 showed no difference in perinatal outcomes between STAN and CTG alone, except a reduction in operative vaginal deliveries (8), whereas a second meta-analysis from 2012 reported a reduction in the need for fetal blood sampling and operative vaginal deliveries (9). Three systematic reviews (10-12) from 2013 all reported a reduction in the need for fetal blood sampling; additionally, two found reductions in operative vaginal deliveries (11, 13), and one reported a reduction in transfers to the neonatal intensive care unit (12). The most recent meta-analysis published in 2014 reported significantly reduced rates of metabolic acidosis, fetal blood sampling, and operative vaginal deliveries in the STAN group (13).

Recently, a large multicenter study from the United States including 11 108 patients showed that fetal ECG ST-segment analysis as an adjunct to conventional intrapartum electronic fetal monitoring neither improved perinatal outcomes nor decreased operative delivery rates (14). Because a new large trial has been published, and also because previous meta-analyses had different conclusions, a new systematic review to compare the effects of STAN vs. CTG alone (15) is warranted.

The aim of this review is to quantify the efficacy of the STAN method as an adjunct to conventional CTG compared with CTG alone. In addition to conventional quality assessments , we used the Grading of Recommendations Assessment, Development and Evaluation (GRADE) to assess the overall quality of evidence (16) and trial sequential analyses (TSA) to clarify the need for additional trials (17).

Material and methods

This systematic review was conducted based on a protocol published in the PROSPERO international prospective register of systematic reviews, registration no. CRD42015023563. We did some minor deviations from the protocol (Supplementary file S1). No ethical approval was needed.

We developed a search strategy, and a systematic literature search was performed in the following databases: Ovid MEDLINE® (In-Process & Other Non-Indexed Citations, Ovid MEDLINE®, Daily, Ovid MEDLINE® and Ovid OLDMEDLINE® 1946 to Present), EMBASE Classic+ (EMBASE 1947 to 2015 September 16) (Ovid), The Web of Science® (Thomson Reuters), Scopus® (Elsevier), The Cochrane Library (Wiley), and CINAHL Plus (EBSCO*host*).

An initial search, with the search terms and combinations shown in Supplementary file S2, was performed in May 2015. This search was followed up by a repetitive search in September 2015 (Supplementary file S2). The controlled vocabulary of Medical Subject Headings (MeSH) from MEDLINE, and the Emtree thesaurus from EMBASE, including sub-headings, were used when applicable. In addition, the search fields *title*, *abstract*, and *keywords*, were searched. In The Web of Science, the search fields *title* and *topic* were used. All references

were exported to Endnote™ (x7.4 – Thompson Reuters), where duplicates were removed. There were no restrictions regarding languages or publication year.

Study selection and data extraction procedures

First, the citations identified by the electronic searches were screened and potentially eligible studies were obtained in full text for further assessment. Two reviewers (EB and LMR) independently assessed eligibility of the studies. Persisting disagreements were resolved by consulting a third reviewer (PØ). The selection criteria were as follows:

- Population: Women in labor, $\geq$ 36 weeks of gestation with a singleton fetus in a cephalic presentation and a decision for continuous electronic fetal monitoring during labor;
- Intervention: CTG plus STAN;
- Comparator: CTG alone;
- Primary outcomes: Operative deliveries for fetal distress, metabolic acidosis in the newborn (pH <7.05 and $BD_{(ecf)}$ >12 mmol/L in umbilical artery). Secondary outcomes: neonatal and perinatal death, neonatal seizures, neonatal encephalopathy, transfers to the neonatal intensive care unit (NICU), fetal blood sampling, cesarean sections, operative vaginal delivery, Apgar score <7 at 5 min and a composite (i.e., either intrapartum death, neonatal death, Apgar score <4 at 5 min, neonatal seizures, metabolic acidosis, intubation at delivery for ventilation or neonatal encephalopathy);
- Study design: RCT.

Two of the reviewers (EB and LMR) extracted data from each study using a predesigned form.

Assessments and synthesis

All studies meeting the selection criteria were critically appraised using the Risk of Bias tool developed and recommended by the Cochrane Collaboration (18). Two reviewers (EB and LMR) performed the risk of bias assessments independently. Persisting disagreements were resolved by consulting a third reviewer (KGB).

Numbers of mothers or infants with the outcome of interest were extracted from all included studies. Outcomes occurring relatively frequently were analyzed by calculating the pooled

risk ratio (RR) with 95% confidence interval (CI) in accordance with a random-effect model. Rare outcomes with incidence less than 1% were combined using peto odds ratio and a fixed-effect model (19). All computations were performed using Review Manager (RevMan, Version 5.3. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration, 2014). Forest plots intended for publications were prepared using R software (Version3.1.2, Vienna: R Foundation for Statistical Computing, 2014) and the forest plot package (20, 21).

To assess the risk of random errors and false-positive results, and to help clarify the need for additional trials by calculating an optimal information size (17), we performed post hoc TSA for selected outcomes in TSA viewer (Version 0.9 beta. Copenhagen: Copenhagen Trial Unit, 2011).

We did not perform any subgroup analysis, but conducted sensitivity analysis where we excluded one trial using old STAN technology (3), and one trial that used a different algorithm for interventions (14). Separate analyses were prepared to explore the impact of pooling data on neonatal and perinatal deaths.

We present overall assessment of the quality of evidence in a "summary of findings" table. The assessment includes the magnitude of effect of the STAN method vs. CTG alone, and a summary of available data on the most important outcomes (22). The quality of evidence was judged as either high, moderate, low, or very low (23).

Results

The electronic searches identified 921 unique records (Fig. 1) and one was identified by personal field knowledge. Ten records were assessed in full text; one was excluded because it evaluated the effect of CTG plus fetal ECG PR analyses, not ST analyses (24). The other nine records were included, six were original studies (3-7, 14) and three contributed additional data and corrections to already published studies (25-27).

Description of included studies

The included studies were performed in Sweden (4), USA (14), Finland (5), France (6), the Netherlands (7), and the UK (3) and included 26 554 women and their babies (Table 1). Most trials used the STAN S21 or S31 monitors (Neoventa AB), whereas the Westgate trial (3) used an older device without computerized assessment for the fetal ECG (STAN 8801, Cinventa AB). The Westgate study included women from 34 weeks gestation. We performed sensitivity analyses without the Westgate study. Moreover, the decision algorithm used in the Belfort study (14) implied that the fetal heart rate status was classified into three zones (green, red, yellow), which correspond closely to the 2008 National Institute of Child and Human Development criteria (28). If the fetal heart rate pattern is in the yellow zone, intervention is recommended if any ST event (either episodic or baseline increase) or two biphasic ST events occur. As this algorithm is different from the one used in other studies, we also conducted sensitivity analysis without the Belfort study.

We assessed the overall risk of bias as low in all the included trials (Table 1, Supplementary file S3).

The effect of STAN method vs. CTG alone

The six available trials included 26 554 women in labor, but a minority of the investigated outcomes reached statistical significance (Table 2, Supplementary file S4). Some of the investigated outcomes are rare, with incidence much lower than one percent, and it is difficult to gain statistical power for definite conclusions. Lack of power was not an issue for the investigated maternal outcomes, and our meta-analysis showed that the use of the STAN method is associated with little or no difference in the rate of cesarean sections (RR 0.93; 95% CI 0.78–1.12) or operative vaginal deliveries (RR 0.87; 95% CI 0.74–1.03) for fetal distress (Table 3). Conversion to absolute numbers suggests that the STAN method would probably result in five more to 10 fewer cesarean sections induced by fetal distress, and between two more and 14 fewer operative vaginal deliveries per 1000 cases of labor (Table 3).

Metabolic acidosis occurred with an incidence less than one percent in the group receiving CTG alone, and even lower in the STAN group (OR 0.64; 95% CI 0.46–0.88; Table 3). The difference corresponds to a number needed to treat to benefit of 401 (95% CI 232–1457), which means that one case of neonatal metabolic acidosis is avoided for every 401 women monitored with STAN. Given a higher baseline incidence of metabolic acidosis, e.g. 1.4% as

in the Amer-Wahlin trial (4), the NNT decreases to 198. All included trials reported data on deaths and three reported neonatal seizures (Fig. 2). Neither resulted in statistically significant differences between the STAN method vs. CTG alone: OR of 1.79 (95% CI 0.69–4.64) and 0.58 (95% CI 0.18–1.90), respectively. The CIs were wide when expressed in relative terms, but re-expressed in absolute terms they imply that STAN can be associated with one fewer to 17 more deaths per 10 000 births, and between nine fewer and nine more neonatal seizures per 10 000 births. Apgar scores <4 after 5 min seemed to occur more frequently with STAN (OR 2.63; 95% CI 1.16–5.96), but we found little or no difference with regard to the incidence of newborns with Apgar scores <7 after 5 min (RR 0.95; 95% CI 0.73–1.25; Table 3 and Fig. 2).

Of the other investigated maternal outcomes, only operative vaginal delivery for all indications reached statistical significance. The effect size suggests that the clinical relevance of the differences is limited (Table 2). Similarly, the other investigated neonatal outcomes pointed towards little or no difference between STAN vs. CTG alone. The only exception to this pattern was fetal blood sampling, which occurred less frequently in the STAN group. The magnitude of this effect was inconsistent across the available trials (Table 2, Supplementary file S4).

Sensitivity analysis

Analyses shows that the results are robust with regard to inclusion or exclusion of the Westgate (3) or Belfort (14) trials (Supplementary file S4). Because we decided to pool studies reporting neonatal and perinatal deaths, we also conducted a sensitivity analysis to explore the possible impact of this decision. The results were consistent among the studies reporting neonatal deaths (Supplementary file S4).

Trial sequential analysis

We determined that a relative risk reduction of 20% represent a clinically important difference in the number of operative deliveries for fetal distress (cesarean sections, vacuum or forceps). In this case, the TSA estimated the optimal information size at 29 940 participants, but suggested that the available sample size was sufficient to conclude that the two treatments are non-inferior (Supplementary file S5). Furthermore, as the majority of newborns with metabolic acidosis are without symptoms or elevated risk for adverse outcomes (29), we held 50% relative risk reduction as clinically important. The main analysis indicated that there was a statistically significant difference between STAN and CTG alone (Supplementary file S5),

but the conclusion depended on the choice of statistical methods. The significance was lost when we used peto OR in combination with a random-effect model rather than in combination with a fixed-effect model. With regard to perinatal- and neonatal deaths and neonatal seizures, the results were far from statistically significant, but the number of observed events was too small to allow firm conclusions about superiority or non-inferiority. For Apgar score <7 after 5 min, TSA showed signs of non-inferiority, i.e., no important difference between the groups.

Summary of findings

The application of GRADE showed that the quality of evidence was moderate or high for the most important outcomes (Table 3).

Discussion

With regard to our primary outcomes, the STAN method did not lead to a reduction in operative deliveries due to fetal distress, but was associated with a reduction in metabolic acidosis.

The meta-analysis showed no significant difference in perinatal or neonatal deaths, seizures, neonatal encephalopathy, pH < 7.05 in cord artery, Apgar score < 7 after 5 min, neonatal intubation, or transfers to the NICU. There was no difference in total operative delivery rate, but a small significant reduction in operative vaginal deliveries for all indications. We found a statistically significant reduction in fetal blood sampling in the STAN group.

Our review has several strengths. Our findings are based on a thorough and up-to-date literature search . We included a recently published trial and used corrected data from earlier trials. All trials are associated with a low risk of bias, and our findings seem robust with regard to inclusion or exclusion of two trials that prompted external validity. Other strengths are our application of GRADE to judge the quality of the evidence and the use of TSA to estimate statistical power and optimal information sizes for different outcomes. Even though GRADE and TSA can be seen as methodological strengths, it is important to note that both methods are based on subjective evaluations, and hence, some disagreement among readers is to be expected. Only RCTs were included in this systematic review. Some relevant outcomes occur at very low incidences, and it is possible that the inclusion of well-designed observational studies could lead to more decisive conclusions for these outcomes.

RCTs are considered the gold standard for clinical trials. They are typically conducted in ideal conditions and under the supervision of dedicated experts. Thus, the external validity to a normal clinical setting, i.e., the distinction between efficacy and effectiveness, can be questioned. The setting is almost never identical across all trials, and this is also the case for the six STAN trials. We believe the observed variation in settings is as can be expected to normal variation in practice, and therefore we decided to include all six trials in our meta-analysis. We are aware that the appropriateness of including some of the trials has been discussed (8-13, 30-33), and we therefore conducted sensitivity analyses to explore the robustness of our results. The overall conclusions of this review are robust with regard to the inclusion or exclusion of the oldest study that used non-computerized ST analysis (3) and the newest that used another decision algorithm (14).

The STAN method is widely used in Scandinavia and some other European countries. Earlier systematic reviews did not unequivocally recommend the technique (8-12), and the results from the Belfort study were eagerly awaited. However, this study used different decision algorithms from those of previous European trials, and the appropriateness of pooling can be debated. Importantly, the algorithms used by Belfort and coworkers were those recommended by the company that produces the STAN technology (Neoventa Medical AB) for their Food and Drug Administration approval and the ones that have been certified for use in the United States. The decision to change the algorithm was made by Neoventa and not the study investigators (personal communication, Michael Belfort, October 24th, 2015).

The included trials report numerous outcomes. We argue that the most important of these are perinatal and neonatal death, neonatal encephalopathy, seizures, and neurologic sequelae such as cerebral palsy was not reported since a long follow-up period is needed. Outcomes such as Apgar score < 7/5 min, transfers to NICU, and intubation for ventilation are less important. From a methodological perspective, it is interesting to note that all relevant neonatal outcomes occur with extremely low incidence (for example, less than 0.1% for death and 0.7% for metabolic acidosis). Under these circumstances, there is a risk that the use of relative effect sizes such as odds ratios inflates the reader's perception of the magnitude of a possible effect (34). Misconception can be avoided by presenting the relative effect sizes together with the corresponding difference in absolute terms (Table 3).

The rate of metabolic acidosis was significantly reduced, but there was a non-significant increase in the death rate in the STAN group. This finding may be the result of chance or a real effect. One theory could be that the STAN technology signals problems at a late stage of compensation, and deterioration is so rapid afterwards that that an intervention may not occur quickly enough in some cases.

Metabolic acidosis has long been viewed as a crucial outcome, but we regard it as a surrogate endpoint. It has been argued that surrogate outcomes with a higher incidence are necessary for efficient evaluation of intrapartum monitoring because the numbers of events for well-defined hard outcomes (such as death) are small (9). Methodological research has shown that surrogate endpoints are frequently associated with biased results (35). The appropriate use of surrogate endpoints requires accurate knowledge and direct correlation between the surrogate and the truly important outcome, and we argue that the relationship between metabolic acidosis and harder outcomes is questionable.

First, there is a known relationship between low cord artery pH values and serious outcome, but the threshold remains unknown (36). Few neonates with severe acidemia appear to have sequelae, and most neonates with adverse outcomes—even those with seizures—are not born acidemic. A study analyzing umbilical artery pH and serious neonatal outcomes in more than 50 000 validated samples from electronically monitored laboring women, showed that 2.2 % of the neonates had a cord pH < 7.00, and only 0.22 % had neonatal encephalopathy (grade 2–3). Only 22 % of all cases with encephalopathy and seizures or death had a pH < 7.00 (37). Another study from the same group showed that an arterial pH < 7.00 significantly predicted neonatal outcomes (38), but the addition of $BD_{(ecf)}$ did not improve the prediction. It is also important to know that similar cord gases may have very different outcomes. The relationship between cord pH values of 7.00–7.10 is much less clear (38).

Second, a Swedish study (29) showed that among 14 687 deliveries, 78 infants (0.5 %) had metabolic acidosis, but 45 % of these infants appeared healthy at birth and did not require transfer to the NICU. This group had no increased risk for neurologic or behavioral problems compared with control children at the age of 6.5 years.

Third, about 75 % of neonates with encephalopathy with seizures or death are born with a pH > 7.00. This may be owing to the “acidosis paradox,” where neonates without acidemia might

be hypoxic but unable to develop acidemia as a response (37, 39). However, the causes of severe long-term neurologic sequelae are probably more complex than previously believed and not simply due to hypoxia with metabolic acidosis (40). Thus, it seems obvious that metabolic acidosis is a surrogate endpoint, and should be interpreted with caution. We found a statistically significant difference in favor of STAN when comparing the incidence of metabolic acidosis, without observing similar effects in other important outcomes. We therefore discourage excessive emphasis on the positive results for metabolic acidosis.

In addition to conventional meta-analysis, we conducted TSA on selected endpoints to explore the possible impact of random errors and false-positive results on the conclusions of our meta-analysis. TSA can allow power analysis to clarify the need for additional trials (17). These analyses showed that the statistical power is too low to draw firm conclusions about superiority or non-inferiority of either STAN or CTG alone on deaths or neonatal seizures. In contrast, TSA showed adequate statistical power to conclude that the STAN method is probably not associated with important reductions in Apgar < 7 at 5 min or with operative deliveries for fetal distress.

Metabolic acidosis was associated with a statistically significant improvement in favor of STAN in the main analysis. Our protocol stated that this analysis should be performed using a random-effect model, but as RevMan does not enable the use of a random-effect model in combination with peto OR effect sizes, the main analysis was based on a fixed-effect model. Interestingly, the TSA analysis showed that the significant finding for metabolic acidosis disappeared in meta-analysis based on random-effect models, a result that underpins the need for caution in interpreting the statistically significant finding for metabolic acidosis.

Our meta-analysis includes the recent US study (14), and shows that the STAN technology does not reduce important outcomes such as perinatal or neonatal death, neonatal encephalopathy, seizures, or total operative delivery rate for fetal distress, but does reduce the fetal blood sampling and metabolic acidosis rates. A recent review of observational studies, reported similar outcomes (41). The reduction in fetal blood sampling is expected, as it is one main reason for introducing the STAN technology, although fetal blood sampling was optional in most of the RCTs. By applying GRADE, we assessed the quality of evidence for all important outcomes as either moderate or high, which implies that our results are close to the true effect of the intervention.

There is no clear indication that STAN causes harm. Some hospitals using STAN may therefore choose to continue, while others may not. In our opinion, hospitals not using STAN should not introduce it, as evidence for benefit is too scarce.

To date, more than 26 000 women and their babies have been included in RCTs to assess the effects of STAN compared with CTG alone. A modest reduction in the rate of metabolic acidosis has been reported, but no reduction in severe neonatal morbidity, mortality rates, or operative delivery rates for fetal distress. To conclude whether STAN performs better than CTG alone in reducing the rate of neonatal seizures, another 50 000 women need to be included in RCTs. For conclusions regarding perinatal or neonatal death, a minimum of 90 000 more women and their babies need to be included. The benefits or harms found will probably be marginal. It is time to conclude that STAN is not better than CTG alone.

There was no specific funding for this study.

References

1. Alfirevic Z, Devane D, Gyte GM. Continuous cardiotocography (CTG) as a form of electronic fetal monitoring (EFM) for fetal assessment during labour. Cochrane database of systematic reviews. 2013;5:CD006066.
2. Rosen KG. Fetal electrocardiogram waveform analysis in labour. Curr Opin Obstet Gynecol. 2005;17:147-50.
3. Westgate J, Harris M, Curnow JS, Greene KR. Randomised trial of cardiotocography alone or with ST waveform analysis for intrapartum monitoring. Lancet. 1992;340:194-8.
4. Amer-Wåhlin I, Hellsten C, Noren H, Hagberg H, Herbst A, Kjellmer I, et al. Cardiotocography only versus cardiotocography plus ST analysis of fetal electrocardiogram for intrapartum fetal monitoring: a Swedish randomised controlled trial. Lancet. 2001;358:534-8.
5. Ojala K, Vaarasmaki M, Makikallio K, Valkama M, Tekay A. A comparison of intrapartum automated fetal electrocardiography and conventional cardiotocography--a randomised controlled study. BJOG. 2006;113:419-23.
6. Vayssière C, David E, Meyer N, Haberstich R, Sebahoun V, Roth E, et al. A French randomized controlled trial of ST-segment analysis in a population with abnormal cardiotocograms during labor. Am J Obstet Gynecol. 2007;197:299.e1-6.
7. Westerhuis ME, Visser GH, Moons KG, van Beek E, Benders MJ, Bijvoet SM, et al. Cardiotocography plus ST analysis of fetal electrocardiogram compared with cardiotocography only for intrapartum monitoring: a randomized controlled trial. Obstet Gynecol. 2010;115:1173-80.
8. Potti S, Berghella V. ST waveform analysis versus cardiotocography alone for intrapartum fetal monitoring: a meta-analysis of randomized trials. Am J Perinatol. 2012;29:657-64.
9. Becker JH, Bax L, Amer-Wahlin I, Ojala K, Vayssiere C, Westerhuis ME, et al. ST analysis of the fetal electrocardiogram in intrapartum fetal monitoring: a meta-analysis. Obstetrics and gynecology. 2012;119:145-54.
10. Salmelin A, Wiklund I, Bottinga R, Brorsson B, Ekman-Ordeberg G, Grimfors EE, et al. Fetal monitoring with computerized ST analysis during labor: a systematic review and meta-analysis. Acta Obstet Gynecol Scand. 2013;92:28-39.
11. Schuit E, Amer-Wahlin I, Ojala K, Vayssiere C, Westerhuis ME, Marsal K, et al. Effectiveness of electronic fetal monitoring with additional ST analysis in vertex singleton pregnancies at >36 weeks of gestation: an individual participant data metaanalysis. Am J Obstet Gynecol. 2013;208:187 e1- e13.

12. Neilson JP. Fetal electrocardiogram (ECG) for fetal monitoring during labour. Cochrane database of systematic reviews. 2013;5:CD000116.
13. Olofsson P, Ayres-de-Campos D, Kessler J, Tendal B, Yli BM, Devoe L. A critical appraisal of the evidence for using cardiotocography plus ECG ST interval analysis for fetal surveillance in labor. Part II: the meta-analyses. Acta Obstet Gynecol Scand. 2014;93:571-86; discussion 87-8.
14. Belfort MA, Saade GR, Thom E, Blackwell SC, Reddy UM, Thorp JM, Jr., et al. A Randomized Trial of Intrapartum Fetal ECG ST-Segment Analysis. N Engl J Med. 2015;373:632-41.
15. Bhide A, Acharya G. When is it appropriate to conduct a(nother) systematic review? Acta Obstet Gynecol Scand. 2015;94:1151-2.
16. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008;336:924-6.
17. Wetterslev J, Thorlund K, Brok J, Gluud C. Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis. J Clin Epidemiol. 2008;61:64-75.
18. Higgins JPT, Altman DG, Sterne JAC (eds.) Chapter 8: Assessing risk of bias in included studies. In: Higgins JPT, Green S (eds.) Cochrane Handbook for Systematic Review of Interventions. Version 5.1.0: The Cochrane Collaboration; 2011. Available from: www.cochrane-handbook.org.
19. Higgins JPT, Deeks JJ, Altmann DG (eds.) Chapter 16: Special topics in statistics. In: Higgins JPT, Green S (eds.) Cochrane Handbook for Systematic Review of Interventions. Version 5.1.0: The Cochrane Collaboration; 2011. Available from: www.cochrane-handbook.org.
20. Team RC. Team R: a language and environment for statistical computing. Vienna 2014. Available from: http://www.R-project.org.
21. Gordon M, Lumley T. Forest plot: Advanced Forest Plot Using "grid" Graphics. R package version 1.1 2015. Available from: http://CRAN.R-project.org/package=forestplot.
22. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. Journal of clinical epidemiology. 2011;64:383-94.
23. Balshem H, Helfand M, Schunemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol. 2011;64:401-6.
24. Strachan BK, van Wijngaarden WJ, Sahota D, Chang A, James DK. Cardiotocography only versus cardiotocography plus PR-interval analysis in intrapartum surveillance: a randomised, multicentre trial. FECG Study Group. Lancet. 2000;355:456-9.
25. Amer-Wåhlin I, Kjellmer I, Marsal K, Olofsson P, Rosen KG. Swedish randomized controlled trial of cardiotocography only versus cardiotocography plus ST analysis of fetal electrocardiogram revisited: analysis of data according to standard versus modified intention-to-treat principle. Acta Obstet Gynecol Scand. 2011;90:990-6.
26. Westerhuis ME, Visser GH, Moons KG, Zuithoff N, Mol BW, Kwee A. Cardiotocography plus ST analysis of fetal electrocardiogram compared with cardiotocography only for intrapartum monitoring: a randomized controlled trial. Obstet Gynecol. 2011;117:406-7.
27. Welin AK, Noren H, Odeback A, Andersson M, Andersson G, Rosen KG. STAN, a clinical audit: the outcome of 2 years of regular use in the city of Varberg, Sweden. Acta Obstet Gynecol Scand. 2007;86:827-32.
28. Macones GA, Hankins GD, Spong CY, Hauth J, Moore T. The 2008 National Institute of Child Health and Human Development workshop report on electronic fetal monitoring: update on definitions, interpretation, and research guidelines. Obstet Gynecol. 2008;112:661-6.
29. Hafström M, Ehnberg S, Blad S, Noren H, Renman C, Rosen KG, et al. Developmental outcome at 6.5 years after acidosis in term newborns: a population-based study. Pediatrics. 2012;129:e1501-7.
30. Olofsson P, Ayres-de-Campos D, Kessler J, Tendal B, Yli BM, Devoe L. A critical appraisal of the evidence for using cardiotocography plus ECG ST interval analysis for fetal surveillance in labor. Part I: the randomized controlled trials. Acta Obstet Gynecol Scand. 2014;93:556-68; discussion 68-9.
31. Øian P, Blix E. Scarce scientific evidence for the use of cardiotocography plus fetal ECG ST interval analysis (STAN). Acta Obstet Gynecol Scand. 2014;93:570.
32. Blix E, Øian P. Deviations from STAN guidelines are frequent but results cannot be excluded when the effectiveness of the method should be evaluated. Acta Obstet Gynecol Scand. 2014;93:589.
33. Steer PJ, Hvidman LE. Scientific and clinical evidence for the use of fetal ECG ST segment analysis (STAN). Acta Obstet Gynecol Scand. 2014;93:533-8.
34. Marshall KG. Prevention. How much harm? How much benefit? 1. Influence of reporting methods on perception of benefits. CMAJ. 1996;154:1493-9.
35. Ciani O, Buyse M, Garside R, Pavey T, Stein K, Sterne JA, et al. Comparison of treatment effect sizes associated with surrogate and final patient relevant outcomes in randomised controlled trials: meta-epidemiological study. BMJ. 2013;346:f457.

36. Malin GL, Morris RK, Khan KS. Strength of association between umbilical cord pH and perinatal and long term outcomes: systematic review and meta-analysis. Bmj. 2010;340:c1471.
37. Yeh P, Emary K, Impey L. The relationship between umbilical cord arterial pH and serious adverse neonatal outcome: analysis of 51,519 consecutive validated samples. BJOG. 2012;119:824-31.
38. Knutzen L, Svirko E, Impey L. The significance of base deficit in acidemic term neonates. Am J Obstet Gynecol. 2015;213:373 e1-7.38.
39. Hermansen MC. The acidosis paradox: asphyxial brain injury without coincident acidemia. Dev Med Child Neurol.. 2003;45:353-6.
40. Leviton A. Why the term neonatal encephalopathy should be preferred over neonatal hypoxic-ischemic encephalopathy. Am J Obstet Gynecol. 2013;208:176-80.
41. Amer-Wåhlin I, Kwee A. Combined cardiotocographic and ST event analysis: A review. Best Pract Res Clin Obstet Gynaecol. 2015; doi: 10.1016/j.bpobgyn.2015.05.007. [Epub ahead of print]

Table and figure legends

Table 1. Characteristics of included studies

Table 2. Outcome events and meta-analyses

Table 3. Summary of findings for selected outcomes

Figure 1. Flow diagram of the study selection process

Figure 2. Forest plot analyses for selected outcomes

Supplementary file S1. Deviations from protocol

Supplementary file S2. Search strategy and complete Medline search

Supplementary file S3. Detailed risk of bias assessments in included studies

Supplementary file S4. Neonatal and maternal outcomes in single studies, meta-analyses and sensitivity analyses

Supplementary file S5: Trial sequential analyses (TSA)

Table 1. Characteristics of included studies

Paper	Amer-Wåhlin, Sweden (4, 25)	Belfort, USA (14)	Ojala, Finland (5)	Vayssière, France (6)	Westerhuis, The Netherlands (7, 26)
Type of study	Multicenter (3 centers)	Multicenter (16 centers)	Single center	Multicenter (2 centers)	Multicenter (9 Centers)
No. included	5049 (revised ITT-analyses)	11 108	1483	799	5681
Inclusion criteria	Women in active labor ≥ 36 gestational w, with singleton fetuses in a cephalic presentation and where a clinical decision of continuous internal CTG	Women with a singleton fetus at more than 36 w of gestation who were attempting vaginal delivery and had cervical dilation of 2 to 7 cm	Consecutive women in active labor with term (≥36+0 gestational w) pregnancy, with a singleton fetus in a cephalic presentation and a decision about amniotomy	Women in labor with a term (≥ 36 gestational w) singleton fetus in cephalic presentation who met the following inclusion criteria: abnormal CTG or thick meconium-stained amniotic fluid	Laboring women aged 18 y or older with a singleton high-risk pregnancy, a fetus in cephalic presentation, gestational age >36 w, and an indication for internal electronic monitoring
Exclusion criteria		Noncephalic presentation, planned CS, a need for immediate delivery, absent fetal heart-rate variability or a sinusoidal pattern, minimal fetal heart-rate variability in the 20 minutes before randomization, or other fetal or maternal conditions that would preclude a trial of labor or the placement of a scalp electrode	Contraindications for scalp electrode or admitted to the labor ward in the second phase of labor	Gestational age <36 w , normal CTG during labor, maternal infection contraindicating placement of scalp electrodes (seropositive for HIV or hepatitis B or C) cardiac malformation, severe decelerations with variability reduced immediately on entry into the delivery room	
Intervention	STAN S21 (Neoventa AB)	STAN S31 (Neoventa AB)	STAN S21 (Neoventa AB)	STAN S21 (Neoventa AB)	STAN S21 or S31 (Neoventa AB)
Algorithm for interventions in STAN group	Table 1, Amer-Wåhlin (4)	Supplementary appendices and trial protocol, referred to in Belfort (14)	As Amer-Wåhlin (4)	As Amer-Wåhlin (4)	FIGO guidelines and STAN clinical guidelines, see appendices 1 & 2, referred to in Westerhuis (7)
Control	Masked STAN S21	Masked STAN S31	CTG (Hewlett-Packard 8030A	CTG (Hewlett-Packard 8030A	Conventional FHR monitor
Proportion primiparas	62% in both arms[1]	42.6% in both arms	51.0% in CTG+STAN arm, 52.4% in CTG arm	72.2% in CTG+STAN arm, 71.8% in CTG arm	57.2% in CTG+STAN arm, 57.0% in CTG arm
Previous CS	-	-	-	6.3% in CTG+STAN arm, 6.0% in CTG arm	12.2% in CTG+STAN arm, 13.1% in CTG arm
Induction of labor	17% in both arms[1]	59.2% in CTG+STAN arm, 58.6 % in CTG arm	20.0% in CTG+STAN arm, 17.5% in CTG arm	8.5% in CTG+STAN arm, 8.8% in CTG arm	40.9% in CTG+STAN arm, 41.8% in CTG arm
Epidural analgesia	37% in CTG+STAN arm, 40 % in CTG arm*	-	54.6% in CTG+STAN arm, 54.0% in CTG arm	91.2% in CTG+STAN arm, 90.3% in CTG arm	41.7% in CTG+STAN arm, 42.6% in CTG arm
Overall risk of bias[2]	Low risk of bias	Low risk of bias	Low risk of bias	Low risk of bias	Low risk of bias

[1]Originally, 5049 women were included and randomized to the study. 83 were excluded for technical reasons, leaving 4966 women for the analyses. In 2011 (24) the authors published analyses according to intention to treat including the 83 previous excluded cases. The estimates are based on the publication from 2001 (4).
[2]See Supplemetary file S3 for detailed risk of bias assessment.

Table 2. Outcome events and meta-analyses

Outcome	No. of studies	Events, n/N	Effect measure[1]	Effect size (95% CI)	I^2 (%)
Metabolic acidosis	6	151/26493	Peto OR	0.64 (0.46-0.88)	31
Total operative deliveries[2] for fetal distress	6	2514/26446	RR	0.88 (0.75-1.03)	74
Perinatal and neonatal deaths	6	17/26446	Peto OR	1.79 (0.69-4.64)	0
Neonatal seizures	3	11/13343	Peto OR	0.58 (0.18-1.90)	0
Apgar <4 at 5 min	1	23/11108	Peto OR	2.63 (1.16-5.96)	-
Apgar <7 at 5 min	5	211/15303	RR	0.95 (0.73-1.25)	0
Neonatal encephalopathy	4	25/23177	Peto OR	0.66 (0.30-1.46)	14
Neonatal intubation	2	85/12544	Peto OR	1.37 (0.90-2.10)	34
Fetal blood sampling	5	2103/15338	RR	0.59 (0.45-0.79)	91
Admittance to NICU	5	1521/26410	RR	1.00 (0.90-1.11)	0
Cord pH <7.05	4	216/10336	RR	1.05 (0.63-1.76)	66
Composite endpoint[3]	1	92/11108	Peto OR	1.31 (0.87-1.98)	-
Total operative deliveries[2] for all indications	6	6451/26446	RR	0.96 (0.91-1.02)	39
Cesarean delivery for fetal distress	6	1124/26446	RR	0.93 (0.78-1.12)	47
Cesarean delivery for all indications	6	3589/26446	RR	1.02 (0.96-1.08)	0
Assisted vaginal delivery for fetal distress	6	1390/26446	RR	0.87 (0.74-1.03)	59
Assisted vaginal delivery for all indications	6	2862/26446	RR	0.92 (0.86-0.99)	0

[1] Peto odds ratio (OR) for outcomes with incidence less than 1%, else risk ratio (RR).
[2]Total operative deliveries = cesarean sections + assisted vaginal deliveries
[3]Composite of intrapartum death, neonatal death, Apgar < 4 at 5 minutes, neonatal seizures, metabolic acidosis, intubation at birth, or neonatal encephalopathy.

Table 3 Summary of findings for selected outcomes

Outcomes	Anticipated absolute effects[1] (95% CI)		Relative effect (95% CI)	№ of participants (studies)	Quality of the evidence (GRADE)
	Risk with CTG alone	**Risk with STAN**			
Metabolic acidosis Cord pH<7.05 + $BD_{(ecf)}$>12 mmol/L	7 per 1000	**4 per 1000** (3 to 6)	**OR 0.64** (0.46 to 0.88)	26493 (6 RCTs)	⊕⊕⊕◯ MODERATE [3,4]
Cesarean delivery for fetal distress	44 per 1000	**40 per 1000** (34 to 49)	**RR 0.93** (0.78 to 1.12)	26446 (6 RCTs)	⊕⊕⊕⊕ HIGH [2]
Operative vaginal delivery for fetal distress	55 per 1000	**47 per 1000** (40 to 56)	**RR 0.87** (0.74 to 1.03)	26446 (6 RCTs)	⊕⊕⊕⊕ HIGH [2]
Neonatal and perinatal death	0 per 1000	**1 per 1000** (0 to 2)	**OR 1.79** (0.69 to 4.64)	26446 (6 RCTs)	⊕⊕⊕◯ MODERATE [5]
Neonatal seizures	1 per 1000	**1 per 1000** (0 to 2)	**OR 0.58** (0.18 to 1.90)	13343 (3 RCTs)	⊕⊕⊕◯ MODERATE [5]
Apgar score <7 at 5 min	14 per 1000	**13 per 1000** (10 to 18)	**RR 0.95** (0.73 to 1.25)	15303 (5 RCTs)	⊕⊕⊕◯ MODERATE [6]

1. Calculations based on mean incidence in study populations
2. OIS achieved. RRR probably less than 20%
3. A surrogate estimate with questionable clinical importance. Choose not to downgrade
4. Both estimate and the conclusion of TSA unstable with regard to choice of analytic methods (e.g fixed vs random model)
5. Current information size much lower than OIS for detecting a 50% reduction with 80% certainty
6. Wide confidence interval – imprecise data

Supplementary file S1. Deviations from the protocol

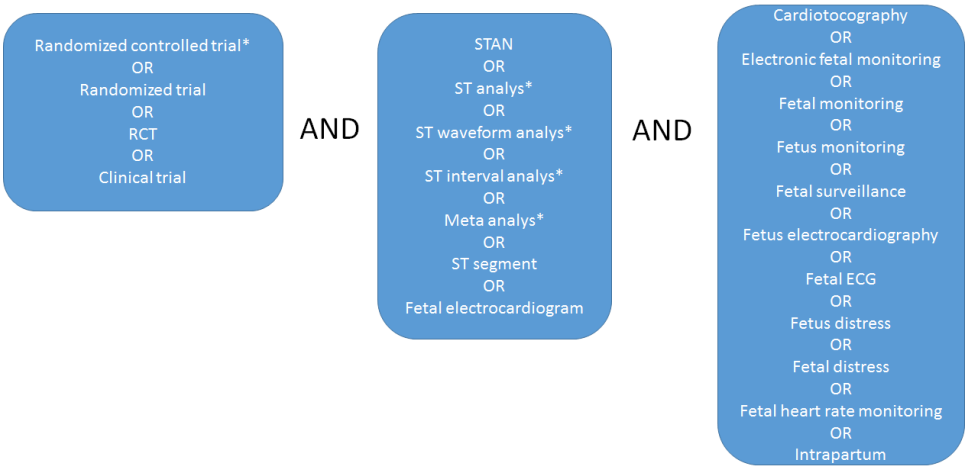
1. In the protocol, we wrote: "One person will extract data from included studies by using a data extraction form. Another person will check the data extraction." When performing the data extraction, we found it more effective that two persons extracted data independently into the extraction forms, and afterwards we compared results.

2. In the protocol, we wrote : "In case of very rare events (<1%) we will calculate Peto OR. Whenever appropriate, estimates will be pooled across trials in meta-analysis by using random effect models in the RevMan Software." We used Peto OR and RR in accordance with the protocol, but Peto OR was used in combination with a fixed effect model rather than a random effect model. This is in accordance with recommendations for meta-analysis of rare events in the Cochrane Handbook.

3. In the protocol, we defined three main endpoints: 1) Operative deliveries for fetal distress 2) Metabolic acidosis in the newborn and 3) Apgar score <4 and 7 at five minutes. During the process, we observed that only Belfort et al reported Apgar score <4 at 5 min and therefore decided to use 1) and 2) as primary endpoints.

4. In the protocol, we wrote: "We will do subgroup analyses for old vs. new STAN technology." We had extensive discussions in the author group about doing subgroup analyses or not after we had collected the results from the included trials. Two of the included studies differed from the other four: the Westgate trial used old technology without computerized assessment for the fetal ECG and the Belfort trial used a different decision algorithm. As our aim was to assess the efficacy of STAN in clinical settings, and we believe that the observed variation in settings is as close as can be expected to normal variations in practice, we included al six RCTs in the main analyses. We decided to do sensitivity analyses instead of subgroup analyses to assess the influence on the overall results from the Westgate and Belfort trials.

Supplementary file S2. Search strategy and complete Medline search



Full search strategy. Example from Medline September 2015

Ovid MEDLINE(R) In-Process & Other Non-Indexed Citations, Ovid MEDLINE(R) Daily, Ovid MEDLINE(R) and Ovid OLDMEDLINE(R) 1946 to Present

#	Searches	Results	Search Type
1	exp Clinical Trial/	845068	Advanced
2	exp Clinical Trials as Topic/	300913	Advanced
3	randomized trial.ti,ab,kf.	32629	Advanced
4	RCT.ti,ab,kf.	11251	Advanced
5	clinical trial.ti,ab,kf.	93686	Advanced
6	randomized control trial.ti,ab,kf.	1877	Advanced
7	or/1-6	1103356	Advanced
8	STAN.ti,ab,kf.	251	Advanced
9	ST analys*.ti,ab,kf.	189	Advanced
10	ST waveform analys*.ti,ab,kf.	45	Advanced
11	ST interval analys*.ti,ab,kf.	7	Advanced
12	exp Meta-Analysis/	60313	Advanced
13	Meta analys*.ti,ab,kf.	81091	Advanced
14	ST segment.ti,ab,kf.	18243	Advanced
15	Fetal electrocardiogram.ti,ab,kf.	389	Advanced
16	or/8-15	116726	Advanced
17	exp Cardiotocography/	1657	Advanced
18	Cardiotocography.ti,ab,kf.	1062	Advanced
19	exp Fetal Monitoring/	8012	Advanced
20	Fetal monitoring.ti,ab,kf.	1816	Advanced
21	Fetal surveillance.ti,ab,kf.	614	Advanced
22	Fetus electrocardiography.ti,ab,kf.	2	Advanced
23	Fetal ECG.ti,ab,kf.	298	Advanced
24	Fetus distress.ti,ab,kf.	0	Advanced
25	exp Fetal Distress/	3037	Advanced
26	Fetal distress.ti,ab,kf.	3816	Advanced
27	Fetal heart rate monitoring.ti,ab,kf.	748	Advanced
28	Intrapartum.ti,ab,kf.	6631	Advanced
29	Electronic fetal monitoring.ti,ab,kf.	508	Advanced
30	or/17-29	19054	Advanced
31	7 and 16 and 30	169	Advanced

Supplementary file S3. Detailed risk of bias assessments

Paper	Amer-Wåhlin, Sweden (4, 25)	Belfort, USA (14)	Ojala, Finland (5)	Vayssière, France (6)	Westerhuis, The Netherlands (7, 26)	Westgate, UK (3)
Adequate sequence generation	Low risk of bias	Low risk of bias	Low risk of bias	Low risk of bias	Low risk of bias	Low risk of bias
Allocation concealment	Low risk of bias	Low risk of bias	Low risk of bias	Low risk of bias	Low risk of bias	Not described
Blinding of participants and personnel	Not possible	Not possible	Not possible	Not possible	Not possible	Not possible
Blinding of outcome assessor	Not described	Low risk of bias	Not described	Not described	Not described	Not described
Incomplete outcome data addressed	Low risk of bias	Low risk of bias	Low risk of bias	Low risk of bias	Low risk of bias	Low risk of bias
Free of selective reporting	Low risk of bias	Low risk of bias	Low risk of bias	Low risk of bias	Low risk of bias	Low risk of bias
Free of other bias*	Low risk of bias	Low risk of bias	Low risk of bias	Low risk of bias	Low risk of bias	Low risk of bias
Total quality judgement	Low risk of bias	Low risk of bias	Low risk of bias	Low risk of bias	Low risk of bias	Low risk of bias

*Other bias: comparable groups at baseline, the groups received the same treatment and care apart from the method of fetal surveillance,

Supplementary file S4. Neonatal and maternal outcomes in single studies, meta-analyses and sensitivity analyses

	CTG+STAN	CTG		
	Events/total	Events/total		95% CI
Metaboloc acidosis (cord pH < 7.05 + $BD_{(ecf)}$ >12mmol/l			Peto OR, fixed effect	
Amer-Wåhlin[1] (4)	18/2565	35/2484	0.51	(0.29-0.87)
Belfort[2] (14)	3/5532	8/5576	0.41	(0.12-1.32)
Ojala[3] (5)	6/714	4/722	1.51	(0.44-5.24)
Vayssière (6)	8/399	5/400	1.60	(0.54. 4.79)
Westerhuis[4] (7)	19/2827	27/2840	0.71	(0.40-1.26)
Westgate (3)	5/1219	13/1215	0.41	(0.16-1.03)
Total	59/13256	92/13237	0.64	(0.46-0.88)
Test for heterogeneity: I^2=27%				
Sensitivity analysis. Westgate (3) not included:				
Total	54/12037	79/12022	0.68	(0.48-0.95)
Test for heterogeneity: I^2=31%				
Sensitivity analysis. Belfort (14) not included:				
Total	56/7724	84/7661	0.66	(0.47-0.92)
Test for heterogeneity: I^2=36%				
Total operative deliveries[5] for fetal distress			RR, random effect	
Amer-Wåhlin (4)	193/2519	227/2447	0.83	(0.69-0.99)
Belfort (14)	512/5532	516/5576	1.00	(0.89-1.12)
Ojala (5)	51/733	63/739	0.82	(0.57-1.16)
Vayssière (6)	134/399	148/400	0.91	(0.75-1.10)
Westerhuis (7)	261/2827	237/2840	1.11	(0.94-1.31)
Westgate (3)	61/1219	111/1215	0.55	(0.40-0.74)
Total	1212/13229	1302/13217	0.88	(0.75-1.03)
Test for heterogeneity: I^2=74%				
Sensitivity analysis. Westgate (3) not included:				
Total	1151/12010	1191/12002	0.95	(0.86-1.06)
Test for heterogeneity: I^2=42%				
Sensitivity analysis. Belfort (14) not included				
Total	700/7697	786/7641	0.84	(0.68-1.03)
Test for heterogeneity: I^2=77%				
Perinatal and neonatal death			Peto OR, fixed effect	
Amer-Wåhlin (4)	3/2519	2/2447	1.45	(0.25-8.38)
Belfort (14)	3/5532	1/5576	2.74	(0.39-19.46)
Ojala (5)	0/733	0/739	-	
Vayssière (6)	0/399	1/400	0.14	(0.00-6.84)
Westerhuis (7)	3/2827	2/2840	1.50	(0.26-8.66)
Westgate (3)	2/1219	0/1215	7.37	(0.46-117.91)
Total	11/13229	6/13217	1.93	(0.62-5.98)
Sensitivity analysis. Westgate (3) not included:				
Total	9/12010	6/12002	1.48	(0.54-4.08)
Test for heterogeneity: I^2=0%				
Sensitivity analysis. Belfort (14) not included:				
Total	8/7697	4/7641	1.57	(0.53-4.66)
Test for heterogeneity: I^2=0%				
Perinatal death			Peto OR, fixed effect	

Amer-Wåhlin (4)	3/2519	2/2447	1.45	(0.25-8.38)
Ojala (5)	0/733	0/739	-	-
Westerhuis (7)	3/2827	2/2840	1.50	(0.26-8.66)
Westgate (3)	2/1219	0/1215	7.37	(0.62.5.98)
Total	8/7298	4/7241	1.93	(0.62-5.98)
Test for heterogeneity: I^2=0%				
Sensitivity analysis. Westgate (3) not included:				
Total	6/6079	4/6026	1.47	(0.43-5.09)
Test for heterogeneity: I^2=0%				
Neonatal death			Peto OR, fixed effect	
Belfort (14)	3/5532	1/5576	2.74	(0.39-19.46)
Vayssière (6)	0/399	1/400	0.14	(0-6.84)
Total	3/5931	2/5976	1.50	(0.26-8.67)
Test for heterogeneity: I^2=45%				
Neonatal seizures			Peto OR, fixed effect	
Belfort (14)	3/5532	4/5576	0.76	(0.17-3.33)
Ojala (5)	0/714	2/722	0.14	(0.01-2.19)
Vayssière (6)	1/399	1/400	1.00	(0.06-16.06)
Total	4/6645	7/6698	0.58	(0.18-1.90)
Test for heterogeneity: I^2=0%				
Sensitivity analysis. Belfort (14) not included:				
Total	1/1113	3/1122	0.37	(0.05-2.63)
Test for heterogeneity: I^2=0%				
Apgar score <4 at 5 min			Peto OR, fixed effect	
Belfort (14)	17/5532	6/5576	2.86	(1.13-7.24)
Apgar score <7 at 5 min			RR, random effect	
Amer-Wåhlin (4)	26/2519	28/2447	0.90	(0.53-1.53)
Ojala (5)	9/714	8/722	1.14	(0.44-2.93)
Vayssière (6)	6/399	6/400	1.00	(0.33-3.08)
Westerhuis (7)	42/2827	34/2840	1.24	(0.79-1.94)
Westgate (3)	20/1219	32/1215	0.62	(0.36-1.08)
Total	103/7678	108/7624	0.95	(0.72-1.24)
Test for heterogeneity: I^2=0%				
Sensitivity analysis. Westgate (3) not included:				
Total	83/6459	76/6409	1.09	(0.80-1.48)
Test for heterogeneity: I^2=0%				
Neonatal encephalopathy			Peto OR, fixed effect	
Amer-Wåhlin (4)	3/2519	8/2447	0.39	(0.12-1.28)
Belfort (14)	4/5532	5/5576	0.81	(0.22-2.98)
Ojala (5)	0/714	1/722	0.14	(0-6.90)
Westerhuis (7)	3/2827	1/2840	2.73	(0.38-19.41)
Total	10/11592	15/11585	0.66	(0.30-1.46)
Test for heterogeneity: I^2=14%				
Sensitivity analysis. Belfort (14) not included:				
Total	6/6060	10/6009	0.60	(0.22-1.59)
Test for heterogeneity: I^2=40%				
Neonatal intubation			Peto OR, fixed effect	
Belfort (14)	42/5532	27/5576	1.56	(0.97-2.51)
Ojala (5)	7/714	9/722	0.79	(0.29-2.10)

Total	49/6246	36/6298	1.37	(0.90-2.10)
Test for heterogeneity: I^2=34%				
Fetal blood sampling			RR, random effect	
Amer-Wåhlin (4)	234/2519	261/2447	0.87	(0.74-1.03)
Ojala (5)	51/733	115/739	0.45	(0.33-0.61)
Vayssière (6)	108/399	248/400	0.44	(0.37-0.52)
Westerhuis (7)	301/2827	578/2840	0.52	(0.46-0.60)
Westgate (3)	93/1219	114/1215	0.81	(0.63-1.06)
Total	787/7697	1316/7641	0.59	(0.45-0.79)
Test for heterogeneity: I^2=91%				
Sensitivity analysis. Westgate (3) not included:				
Total	694/6478	1202/6426	0.55	(0.40-0.76)
Test for heterogeneity: I^2=92%				
Admittance to NICU			RR, random effect	
Amer-Wåhlin (4)	169/2519	181/2447	0.91	(0.74-1.11)
Belfort (14)	498/5532	470/5576	1.07	(0.95-1.20)
Ojala (5)	26/714	26/722	1.01	(0.59-1.72)
Vayssière (6)	5/399	6/400	0.84	(0.26-2.72)
Westerhuis[6] (7)	40/2827	45/2840	0.89	(0.59-1.36)
Westgate (3)	24/1219	31/1215	0.77	(0.46-1.31)
Total	767/13210	759/13200	1.00	(0.91-1.11)
Test for heterogeneity: I^2=0%				
Sensitivity analysis. Westgate (3) not included:			RR, random effect	
Total	738/11991	728/11985	1.01	(0.92-1.12)
Test for heterogeneity: I^2=0%				
Sensitivity analysis. Belfort (14) not included:				
Total	264/7678	289/7624	0.90	(0.76-1.06)
Test for heterogeneity: I^2=0%				
Cord pH <7.05			RR, random effect	
Ojala (5)	20/714	8/722	2.53	(1.12-5.70)
Vayssière (6)	12/399	11/400	1.09	(0.49-2.45)
Westerhuis (7)	47/2827	70/2840	0.67	(0.47-0.97)
Westgate (3)	23/1219	25/1215	0.92	(0.63-1.76)
Total	102/5159	114/5177	1.05	(0.63-1.76)
Test for heterogeneity: I^2=66%				
Sensitivity analysis. Westgate (3) not included				
Total	76/3940	89/3962	1.16	(0.53-2.55)
Test for heterogeneity. I^2=77%				
Composite endpoint[7]			Peto OR, fixed effect	
Belfort (14)	52/5532	40/5576	1.31	(0.87-1.98)
Total operative deliveries[5] for all indications			RR, random effect	
Amer-Wåhlin (4)	454/2519	500/2447	0.88	(0.79-0.99)
Belfort (14)	1263/5532	1228/5576	1.04	(0.97-1.11)
Ojala (5)	117/733	114/739	1.03	(0.82-1.31)
Vayssière (6)	216/399	221/400	0.98	(0.86-1.11)
Westerhuis (7)	789/2827	822/2840	0.96	(0.89-1.05)
Westgate (3)	344/1219	383/1215	0.90	(0.79-1.01)
Total	3183/13229	3268/13217	0.96	(0.91-1.02)
Test for heterogeneity: I^2=39%				

Sensitivity analysis. Westgate (3) not included:				
Total	2839/12010	2885/12002	0.98	(0.92-1.04)
Test for heterogeneity: I^2=35%				
Sensitivity analysis. Belfort (14) not included:				
Total	1920/7697	2040/7641	0.94	(0.89-0.99)
Test for heterogeneity: I^2=0%				
Cesarean delivery for fetal distress			RR, random effect	
Amer-Wåhlin (4)	87/2519	97/2447	0.87	(0.66-1.16)
Belfort (14)	287/5532	298/5576	0.97	(0.83-1.14)
Ojala (5)	15/733	15/739	1.01	(0.50-2.05)
Vayssière (6)	54/399	65/400	0.83	(0.60-1.16)
Westerhuis (7)	91/2827	70/2840	1.31	(0.96-1.78)
Westgate (3)	15/1219	30/1215	0.50	(0.27-0.92)
Total	549/13229	575/13217	0.93	(0.78-1.12)
Test for heterogeneity: I^2=47%				
Sensitivity analysis. Westgate (3) not included:				
Total	534/12010	545/12002	0.98	(0.85-1.13)
Test for heterogeneity: I^2=19%				
Sensitivity analysis. Belfort (14) not included:				
Total:	262/7697	277/2641	0.90	(0.69-1.19)
Test for heterogeneity: I^2=57%				
Cesarean delivery for all indications			RR, random effect	
Amer-Wåhlin (4)	210/2519	222/2447	0.92	(0.77-1.10)
Belfort (14)	934/5532	901/5576	1.04	(0.96-1.14)
Ojala (5)	47/733	35/739	1.35	(0.88-2.07)
Vayssière (6)	99/399	109/400	0.91	(0.72-1.15)
Westerhuis (7)	405/2827	391/2840	1.04	(0.91-1.18)
Westgate (3)	115/1219	121/1215	0.95	(0.74-1.21)
Total	1810/13229	1779/13217	1.02	(0.96-1.08)
Test for heterogeneity: I^2=0%				
Sensitivity analysis. Westgate (3) not included:				
Total	76/3940	89/3962	1.16	(0.53-2.55)
Test for heterogeneity: I^2=7%				
Sensitivity analysis. Belfort (14) not included:				
Total:	876/7697	878/7641	0.99	(0.91-1.08)
Test for heterogeneity: I^2=0%				
Operative vaginal delivery for fetal distress			RR, random effect	
Amer-Wåhlin (4)	106/2519	130/2447	0.79	(0.62-1.02)
Belfort (14)	255/5532	218/5576	1.04	(0.87-1.25)
Ojala (5)	36/733	48/739	0.76	(0.50-1.15)
Vayssière (6)	80/399	83/400	0.97	(0.73-1.27)
Westerhuis (7)	170/2827	167/2940	1.06	(0.86-1.30)
Westgate (3)	46/1219	81/1215	0.57	(0.40-0.81)
Total	663/13229	727/13217	0.87	(0.74-1.03)
Test for heterogeneity: I^2=59%				
Sensitivity analysis. Westgate (3) not included:				
Total	617/12010	646/12002	0.95	(0.85-1.07)
Test for heterogeneity: I^2=22%				

Sensitivity analysis. Belfort (14) not included:				
Total	438/7697	509/7641	0.83	(0.68-1.01)
Test for heterogeneity: I^2=58%				
Operative vaginal delivery for all indications			RR, random effect	
Amer-Wåhlin (4)	244/2519	278/2447	0.85	(0.72-1.00)
Belfort (14)	329/5532	327/5576	1.01	(0.87-1.18)
Ojala (5)	70/733	79/739	0.89	(0.66-1.21)
Vayssière (6)	117/399	112/400	1.05	(0.84-1.30)
Westerhuis (7)	384/2827	431/2840	0.90	(0.79-1.02)
Westgate (3)	229/1219	262/1215	0.87	(0.74-1.02)
Total	1373/13229	1489/13217	0.92	(0.86-0.99)
Test for heterogeneity: I^2=0%				
Sensitivity analysis. Westgate (3) not included:				
Total	1147/1210	1227/12002	0.93	(0.86-1.00)
Test for heterogeneity: I^2=0%				
Sensitivity analysis. Belfort (14) not included:				
Total:	1044/7697	1162/7641	0.90	(0.83-0.97)
Test for heterogeneity: I^2=0%				

[1]We used corrected data published by Amer-Wåhlin et al. 2011 (25).
[2]Belfort et al. (14) defined metabolic acidosis as pH $\leq$ 7.05 and BD $\geq$ 12.0, the other studies as pH < 7.05 and BD > 12.0.
[3]Ojala et al. (5) used data with BD (blood), these have been recalculated to $BD_{(ecf)}$ and published by Welin et al (27). We used the recalculated data in the meta-analysis.
[4]We used corrected data published by Westerhuis et al. 2011 (26).
[5]Total operative deliveries = cesarean sections + operative vaginal deliveries.
[6]We used corrected data published by Westerhuis et al. 2011 (26).
[7]A composite of intrapartum death, neonatal death, Apgar score < 4 at 5 minutes, neonatal seizures, metabolic acidosis (pH $\leq$ 7.05 and $BD_{(ecf)}$ $\geq$ 12 mmol/l in the umbilical artery), intubation at birth, or neonatal encephalopathy.

Supplementary S5: Trial sequential analysis (TSA)

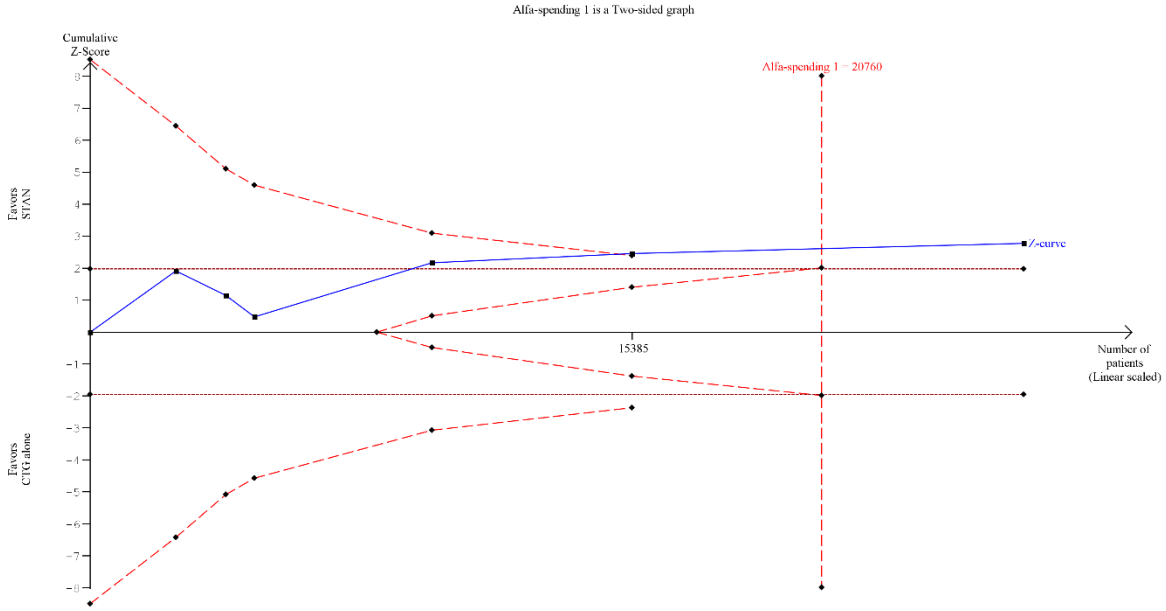
All trials are associated with a risk of arriving at erroneous conclusion. There is a risk of false positive findings (Type I errors), but there is also a risk that a trial erroneously arrive at negative conclusions (Type II errors). A reduction of the risk of type I and/or type II errors in a trial comes at the cost of a need to recruit more participants. Statistical power calculations are about finding an acceptable balance between the risk of arriving at erroneous conclusions and the need to limit the number of participants.

Meta-analyses are also associated with certain risks of spurious findings due to type I and type II errors, but the pooled effect estimate can be expected to converge towards the truth as the number of available trials and the number of events accumulate. TSA can be used to evaluate if a meta-analysis is prone to spurious findings, or whether the pooled estimate is likely to represent the truth. TSA calculation resembles traditional power analysis, and enquires the user to define a set of input variable, i.e. an estimate for baseline incidence, an estimate for minimal important difference, a limit for the largest acceptable risk of type I error, and a limit for the largest acceptable risk of type II errors.

All TSA were conducted *post hoc* in TSA viewer (Version 0.9 beta. Copenhagen: Copenhagen Trial Unit, 2011). We used the same statistical methods in the TSA as in the main analysis in RevMan. Alfa-spending boundaries were calculated using two-sided tests with accepted risks of type I and type II errors at 5% and 20%, respectively. An estimate for baseline incidence was obtained by calculating the mean incidence across all available trials.

Incidence of metabolic acidosis

Six included trials. The blue line shows the cumulative Z-score of the meta-analysis. The outer red line represents the alfa-spending level of significance, and the black dotted lines represents the conventional significance 5% significance borders. The inner red lines represent borders for non-inferiority. Optimal information size (OIS) estimated to 20760 participants.



Meta-analysis

Effect measure: peto odds ratio

Effect model: Fixed effect

Pooled effect: 0.64 (95% CI 0.46 to 0.88)

Heterogeneity (Q), p-value: 0.23

Inconsistency (I^2): 27% (95% CI 0.00 to 0.53%)

Diversity (D^2): 36%

Boundaries

Name: Alpha-spending

Type: Two-sided

Type I error: 5%

Power: 80%

Minimal important difference: RRR 50%

Incidence in control arm: 0.7%

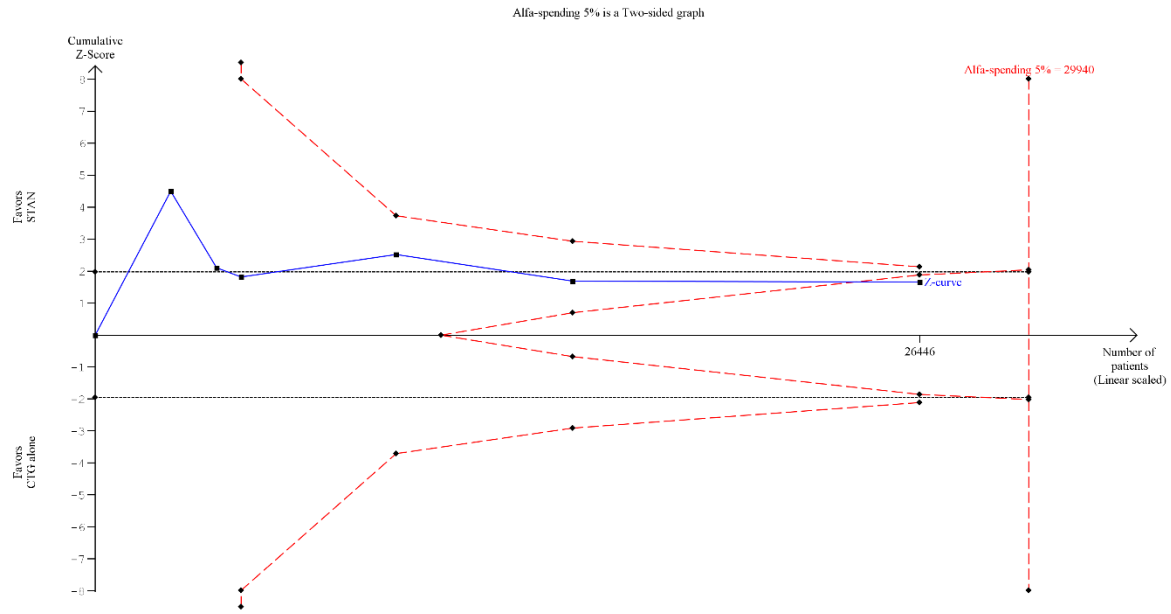
Name: Conventional

Type: Two-sided

Type I error: 5%

Operative deliveries for fetal distress (cesarean sections, vacuum or forceps)

Six included trials. The blue line shows the cumulative Z-score of the meta-analysis. The outer red line represents the alfa-spending level of significance, and the black dotted lines represents the conventional significance 5% significance borders. The inner red lines represent borders for non-inferiority. Optimal information size (OIS) estimated to 29940 participants.



Meta-analysis

Effect measure: Risk ratio

Effect model: Random effect (DL)

Pooled effect: 0.88 (95% CI 0.75 to 1.03)

Heterogeneity (Q), p-value: 0.002

Inconsistency (I^2): 74% (95% CI 61 to 83%)

Diversity (D^2): 79%

Boundaries

Name: Alpha-spending

Type: Two-sided

Type I error: 5%

Power: 80%

Minimal important difference: RRR 20%

Incidence in control arm: 10%

Name: Conventional

Type: Two-sided

Type I error: 5%

Incidence of neonatal or perinatal death

Six included trials. The blue line shows the cumulative Z-score of the meta-analysis. The outer red line represents the alfa-spending level of significance, and the black dotted lines represents the conventional significance 5% significance borders. The inner red lines represent borders for non-inferiority. Optimal information size (OIS) estimated to 117663 participants.

Meta-analysis

Effect measure: peto odds ratio

Effect model: Fixed effect

Pooled effect: 1.79 (95% CI 0.69 to 4.63)

Heterogeneity (Q), p-value: 0.57

Inconsistency (I^2): 0% (95% CI 0.00 to 0.54%)

Diversity (D^2): 0%

Boundaries

Name: Alpha-spending

Type: Two-sided

Type I error: 5%

Power: 80%

Minimal important difference: RRR 50%

Incidence in control arm: 0.08%

Name: Conventional

Type: Two-sided

Type I error: 5%

Incidence of neonatal seizures

Three included trials. The blue line shows the cumulative Z-score of the meta-analysis. The outer red line represents the alfa-spending level of significance, and the black dotted lines represents the conventional significance 5% significance borders. The inner red lines represent borders for non-inferiority. Optimal information size (OIS) estimated to 94116 participants.

Meta-analysis

Effect measure: peto odds ratio

Effect model: Fixed effect

Pooled effect: 0.59 (95% CI 0.18 to 1.90)

Heterogeneity (Q), p-value: 0.66

Inconsistency (I^2): 0% (95% CI 0.00 to 0.60%)

Diversity (D^2): 0%

Boundaries

Name: Alpha-spending

Type: Two-sided

Type I error: 5%

Power: 80%

Minimal important difference: RRR 50%

Incidence in control arm: 0.1%

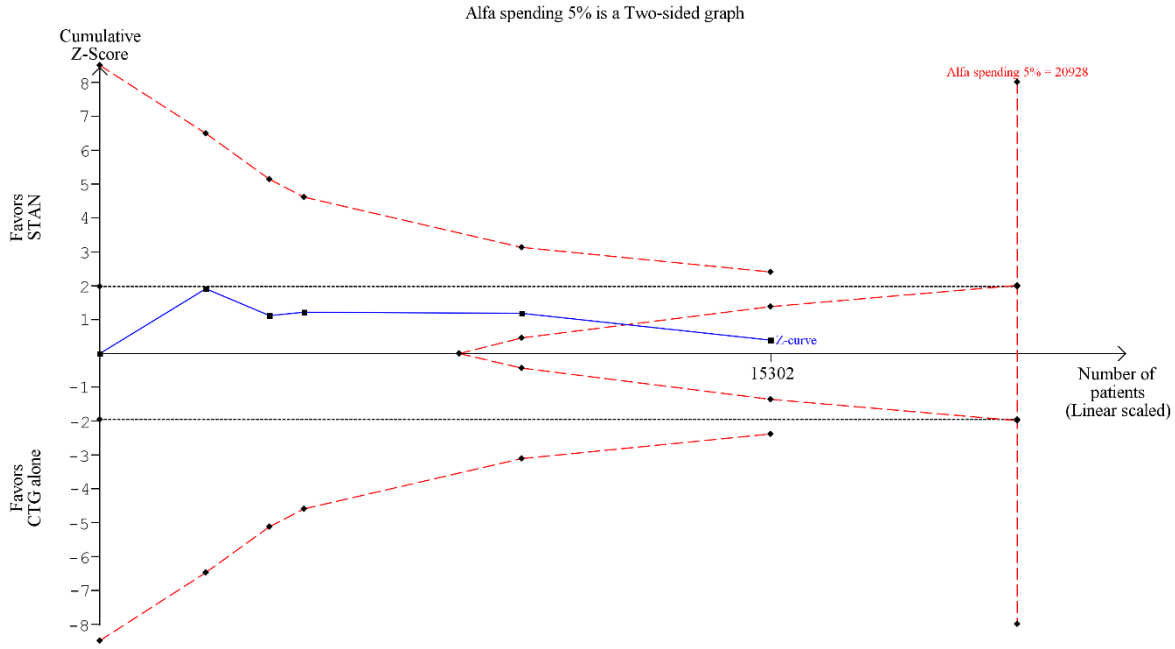
Name: Conventional

Type: Two-sided

Type I error: 5%

Incidence of Apgar scores less than seven after five minutes

Four included trials. The blue line shows the cumulative Z-score of the meta-analysis. The outer red line represents the alfa-spending level of significance, and the black dotted lines represents the conventional significance 5% significance borders. The inner red lines represent borders for non-inferiority. Optimal information size (OIS) estimated to 20928 participants.



Meta-analysis

Effect measure: Risk ratio

Effect model: Random effect (DL)

Pooled effect: 0.95 (95% CI 0.73 to 1.25)

Heterogeneity (Q), p-value: 0.44

Inconsistency (I^2): 0% (95% CI 0.00 to 0.54%)

Diversity (D^2): 0%

Boundaries

Name: Alpha-spending

Type: Two-sided

Type I error: 5%

Power: 80%

Minimal important difference: RRR 30%

Incidence in control arm: 1.4%

Name: Conventional

Type: Two-sided

Type I error: 5%