AOGS
Acta Obstetricia et Gynecologica Scandinavica

AOGS ORIGINAL RESEARCH ARTICLE

New FIGO and Swedish intrapartum cardiotocography classification systems incorporated in the fetal ECG ST analysis (STAN) interpretation algorithm: agreements and discrepancies in cardiotocography classification and evaluation of significant ST events

PER OLOFSSON[1], HÅKAN NORÉN[2] & ANN CARLSSON[2]

[1]*Institution of Clinical Sciences Malmö, Lund University, Malmö, and* [2]*Department of Obstetrics and Gynecology, Sahlgrenka University Hospital, Gothenburg, Sweden*

Key words
Cardiotocography, delivery, fetal ECG, fetal monitoring, fetal surveillance, FIGO, ST analysis

Correspondence
Per Olofsson, Cura Mödravård, Box 20037,
200 74 Malmö, Sweden.
E-mail: per.olofsson@med.lu.se

Conflict of interest
Per Olofsson is consulting Global Medical Adviser to Neoventa Medical AB. Håkan Norén and Ann Carlsson state explicitly that there are no conflicts of interest in connection with this article.

Please cite this article as: Olofsson P, Norén H, Carlsson A. New FIGO and Swedish intrapartum cardiotocography classification systems incorporated in the fetal ECG ST analysis (STAN) interpretation algorithm: agreements and discrepancies in cardiotocography classification and evaluation of significant ST events. Acta Obstet Gynecol Scand 2018; 97:219–228.

Received: 10 April 2017
Accepted: 30 November 2017

DOI: 10.1111/aogs.13277

Abstract

Introduction. The updated intrapartum cardiotocography (CTG) classification system by FIGO in 2015 (FIGO2015) and the FIGO2015-approached classification by the Swedish Society of Obstetricians and Gynecologist in 2017 (SSOG2017) are not harmonized with the fetal ECG ST analysis (STAN) algorithm from 2007 (STAN2007). The study aimed to reveal homogeneity and agreement between the systems in classifying CTG and ST events, and relate them to maternal and perinatal outcomes. *Material and methods.* Among CTG traces with ST events, 100 traces originally classified as normal, 100 as suspicious and 100 as pathological were randomly selected from a STAN database and classified by two experts in consensus. Homogeneity and agreement statistics between the CTG classifications were performed. Maternal and perinatal outcomes were evaluated in cases with clinically hidden ST data (n = 151). A two-tailed $p < 0.05$ was regarded as significant. *Results.* For CTG classes, the heterogeneity was significant between the old and new systems, and agreements were moderate to strong (proportion of agreement, kappa index 0.70–0.86). Between the new classifications, heterogeneity was significant and agreements strong (0.90, 0.92). For significant ST events, heterogeneities were significant and agreements moderate to almost perfect (STAN2007 vs. FIGO2015 0.86, 0.72; STAN2007 vs. SSOG2017 0.92, 0.84; FIGO2015 vs. SSOG2017 0.94, 0.87). Significant ST events occurred more often combined with STAN2007 than with FIGO2015 classification, but not with SSOG2017; correct identification of adverse outcomes was not significantly different between the systems. *Conclusion.* There are discrepancies in the classification of CTG patterns and significant ST events between the old and new systems. The clinical relevance of the findings remains to be shown.

Abbreviations: CI, confidence interval; CRF, case report form; CTG, cardiotocography; FHR, fetal heart rate; FIGO1987, FIGO CTG classification system from 1987; FIGO2015, FIGO CTG classification system from 2015; FIGO, International Federation of Gynecology and Obstetrics; MA, meta-analysis; PA, proportion of agreement; QW, quadratic weighted; RCT, randomized controlled trial; SSOG2017, SSOG CTG classification system from 2017; SSOG, Swedish Society of Obstetrics and Gynecology; STAN2007, CTG classification system from 2007; STAN, ST analysis; SwRCT, Swedish randomized controlled trial; T/QRS, T wave and QRS complex ratio in ECG.

Introduction

In 2015 the International Federation of Gynecology and Obstetrics (FIGO) presented an updated intrapartum cardiotocography (CTG) classification system (FIGO2015) (1), which replaces the system from 1987 (FIGO1987) (2). It is expected that the new system will be introduced worldwide. The CTG classification system used in the fetal ECG ST analysis (STAN) clinical guidelines from 2007 (STAN2007) (3), based on the FIGO1987 classification, has for a long time been used in clinical practice and in numerous randomized and clinical studies, but the STAN2007 and FIGO2015 systems are not harmonized. In comparisons between the two classification systems (Tables S1 and S2, Figure S1), the most conspicuous differences are that

- Baseline fetal heart rate (FHR) above 150 and 170 bpm are classified differently by the systems.
- FHR accelerations are not needed to classify a normal variability by FIGO2015.
- Absent variability (silent pattern) and preterminal pattern are not classified by FIGO2015 but constitute a fourth CTG class (preterminal CTG) in the STAN2007 system.
- Increased variability >25 bpm (saltatory pattern) and sinusoidal patterns are classified differently by the two systems, depending on the duration.
- The FIGO2015 system, but not the STAN2007 system, defines repetitiveness of decelerations.
- The depth of variable decelerations lasting <60 s and variable decelerations lasting 60–180 s are classified differently by the two systems.
- The uterine contraction frequency is not considered in the FIGO2015 system.
- Two of the element categories baseline FHR, variability, accelerations and decelerations are required to be classified as suspicious to judge a CTG trace suspicious in the STAN2007 system, whereas only one element is required in the FIGO2015 system.

Hence, it is unclear whether the FIGO2015 CTG classification can be incorporated into the STAN clinical guidelines interpretation algorithm. In Sweden, the Swedish Society of Obstetrics and Gynecology (SSOG) and the Swedish Association of Midwives expert committee has introduced a FIGO2015-approached classification valid from 2017 (SSOG2017) (4) (Table S3). The similarities and differences between the old and the new systems are displayed in Figure S1.

The primary objective of the study was to investigate the agreements and homogeneities between the STAN2007, the FIGO2015 and the SSOG2017 systems in the classification of CTG, and the possible differences in significant and non-significant ST events read in the STAN clinical guidelines interpretation algorithm (3,5) by replacing the STAN2007 system with the FIGO2015 system or SSOG2017 system. No previous study has addressed this issue. Santo et al. (6) found that differences in CTG classification systems have profound effects on interobserver agreement and reliability, as well as on the sensitivity and specificity of CTG classification systems in predicting acidemia. To avoid the problem of interobserver differences in the present study, two senior experts of CTG and STAN interpretations made all classifications in consensus.

A further objective of the study was to evaluate the consequences on judging ST events as significant or not, i.e. whether an ST event should lead to a clinical intervention, and to investigate the associations with maternal and perinatal outcomes. Looking at the list of differences between the systems, we hypothesized that the different CTG classification systems will result in significantly discrepant classifications into normal/suspicious/pathological traces and significantly discrepant recommendations for action in cases of ST events (significant/non-significant ST events).

Material and methods

The material for the study was retrieved from the Swedish multicenter randomized controlled trial (SwRCT) (7) database. This database comprises a high-risk population, representing 31–36% of the total obstetrics population during an 18-month period from 1998 to 2000 at the three university hospitals involved in Gothenburg, Lund and Malmö.

In the total material (n = 4966), ST events were recorded in 2016 cases (40.6%). The CTGs were classified post hoc by the SwRCT authors as normal, suspicious, pathological and preterminal; among cases with these original classifications available (n = 4820), ST events were recorded in 1981/4820 (41.1%): in 1160/3146 (36.9%) of cases with a normal CTG pattern, in 606/1212 (50.0%) of cases with a suspicious pattern, and in 215/462 (46.5%) of cases with a pathological pattern. When preterminal patterns occur, possible ST events should be disregarded and preterminal CTGs were thus not included in the study. Among cases with ST events (n = 1981), 690 (34.8%) had only a single ST event recorded.

Key Message

The STAN2007, FIGO2015 and SSOG2017 CTG classification systems are not interchangeable in regard to CTG patterns and ST events classifications.

A sample of 100 CTG traces that were originally classified as normal by the SwRCT authors, 100 as suspicious and 100 as pathological, was randomly selected among cases with ST events, ignoring the number of ST events and when in labor they occurred (random selection process described below).

The 300 randomly selected CTG traces represented 15.1% (300/1981) of all traces with ST events: 8.6% (100/1160) of those originally classified as normal, 16.5% (100/606) of the suspicious, and 46.5% (100/215) of those classified pathological. In all, 151 cases were retrieved from the CTG arm (i.e. cases with hidden ST data) of the SwRCT and 149 from the STAN arm.

The selected traces should last for at least 30 min and be of adequate quality for reading the individual element categories for classification (baseline FHR, variability, accelerations, decelerations). There was no maximum time limit of monitoring. The number of 300 traces was chosen following Grant's recommendation (8).

Classification of CTG traces and ST events

The 300 CTG traces and the ST events were classified in consensus by two expert CTG and STAN users (H.N. and A.C.) with several decades of clinical experience. These experts were blinded to clinical information, from the original CTG classification by SwRCT authors, and from ST data when they classified the CTG traces. They classified each trace as normal/suspicious/pathological with the three classification systems, respectively, based on their detailed classification of the element categories of the trace (Figure S1). The fourth CTG class in the STAN2007 system, the preterminal pattern, has been omitted as an independent class in the FIGO2015 and SSOG2017 systems and was not evaluated in the present study.

In the comparisons of CTG classes, the intermediary CTG pattern in the STAN2007 system was compared with the suspicious pattern in the FIGO2015 and SSOG2017 systems, and the STAN2007 abnormal pattern was correspondingly compared with the pathological pattern. The terms used in this article are normal, suspicious and pathological.

Each CTG trace was first classified offline with the aid of the STAN Viewer software (Neoventa Medical, Mölndal, Sweden) at a paper speed of 1 cm/min by the two experts, without knowledge of the type of ST events, as the ST information was switched off; the ST event data were then switched on and the presence/absence of significant ST events determined.

Random selection of cases for the study

The random selection of cases was performed without knowledge of clinical outcomes. Since it was possible to switch on/off the ST information at the post hoc assessment, traces from both the STAN arm and CTG arm in the SwRCT were included. Cases in the SwRCT database (n = 4966) were first arranged in alphabetic order according to their original marking given at the enrollment in the SwRCT and then each case assigned a random number according to the Lehmer random number generator (9), starting with the "random number of the day" (10). Cases with inadequate CTG registrations and cases without ST events were omitted and the remaining cases were then pooled into groups of "normal," "suspicious" and "pathological" CTG traces. The 100 cases with the highest random numbers in each group were selected for the study. Finally, these 300 selected cases were arranged in a rising random-number sequence and presented to the two experts. It was not possible for the experts to identify the original classification of traces.

Case report form

A customized case report form (CRF) was used to simplify the classification procedure (Figure S1). Vague element category classifications in the FIGO2015 and SSOG2017 systems were clarified after contact with the principal authors of the respective publications (personal email communications with Prof. Diogo Ayres-de-Campos and Dr. Malin Holzmann). Each trace was classified according to the STAN2007, FIGO2015 and SSOG2017 systems (Tables S1–S3).

Main outcome measures

- The homogeneity and proportion of agreement (PA) between the element categories (baseline FHR, variability, accelerations and decelerations) in the CTG classification systems.
- The homogeneity and PA between CTG classification systems.
- The agreement/discrepancy between the classification systems in identifying significant ST events.
- The relations between significant ST events as judged in the three CTG classification systems and maternal and perinatal outcomes.

Outcome parameters

Since the ST data were available to the managing obstetricians and midwives in the STAN arm in the SwRCT, the relation between significant ST events and maternal and perinatal outcomes was analyzed only in the CTG arm of the trial (n = 151).

Operative delivery for fetal distress was regarded a maternal outcome parameter, and perinatal outcome parameters were Apgar score <4 at 1 min, score <7 at 5 or 10 min, umbilical cord artery pH <7.05, metabolic acidosis (pH <7.05 and base deficit in extracellular fluid >12.0 mmol/L) (11), and admission to the neonatal intensive care unit.

Statistical analyses

Statistics were performed with aid of StatView® computer software (SAS Institute, version 5.0.1, Cary, NC, USA). The PA with 95% CI was calculated and reported according to "Guidelines for reporting reliability and agreement studies" by Kottner et al. (12) with software available on the web (13). This statistics provide data on composite PA as well as specific PAs (which in the present study were the element categories of CTG classifications). A lower 95% confidence (CI) limit <0.50 was regarded as a poor agreement (8).

Agreement in 2 × 2 and 3 × 3 tables was calculated with quadratic weighted (QW) Cohen's kappa statistics (12) with software available on the web (13). The quadratic form of weight was chosen because the difference between nominal categories of CTG classification, for example bradycardia vs. normal FHR baseline, and normal baseline vs. tachycardia, has different clinical impacts. The same is valid for classification of variability (silent pattern vs. decreased variability vs. saltatory pattern) and decelerations (late vs. uncomplicated vs. no decelerations). We evaluated the level of agreement according to McHugh (14), where a kappa index of 0–0.20 represents no agreement (0–4% reliable data), 0.21–0.39 minimal agreement (4–15% reliable data), 0.40–0.59 weak agreement (15–35% reliable data), 0.60–0.79 moderate agreement (35–63% reliable data), 0.80–0.90 strong agreement (64–81% reliable data), and >0.90 almost perfect agreement (82–100% reliable data). According to McHugh, any kappa index <0.60 indicates inadequate agreement.

To test marginal homogeneity in matched-pairs dichotomous data, we used the McNemar test in 2 × 2 tables with software available on the web (13); for polytomous data of higher order than two, we used the Friedman test for STAN2007 vs. FIGO2015 vs. SSOG2017 and then the Wilcoxon matched-pairs signed-ranks test for two-group comparisons. For further explanations of these tests, see Appendix S1. To enable these two latter statistical tests, categorical data were transformed to continuous ordinal data and *p*-values adjusted for ties were used. A two-tailed *p*-value of <0.05 was regarded significant, i.e. a significant *p*-value indicated a significant discrepancy between the classifications.

Ethical approval

All enrolled women gave their oral consent to participate in the Swedish multicenter randomized controlled trial and ethical approvals were obtained from the Regional Ethical Review Boards in Lund (LU 305-98) and Gothenburg (Gbg M 66-98).

Results

Baseline FHR

The distribution of baseline FHR patterns and classifications are shown in Table S4. Due to the different cut-off limits for normal/suspicious/pathological baseline FHR, 46/300 (15.3%) of traces were classified discrepantly: 32 of suspicious STAN2007 traces were classified normal by FIGO2015 and SSOG2017, and 14 of pathological STAN2007 traces were classified suspicious by FIGO2015 and SSOG2017.

Tests for homogeneity showed that the STAN2007 and FIGO2015 systems were significantly heterogeneous for baseline FHR patterns (Table S5). The tests for agreement indicated moderate agreements between the systems, but the lower 95% CI limit (QW kappa index 0.54) indicated an inadequate agreement according to McHugh (14). Among the CTG classes, normal patterns showed a strong agreement but among suspicious and pathological patterns the agreements were poor. The SSOG2017 classification was identical to the FIGO2015 classification.

FHR variability

Among 16 CTG traces classified suspicious by STAN2007 due to increased FHR variability (saltatory pattern), 14 were classified pathological (lasting >30 min) and two normal (lasting <30 min) by the FIGO2015 and SSOG2017 systems (Table S4). Such a time aspect is not included in the STAN2007 system.

The STAN2007 vs. FIGO2015 and vs. SSOG2017 systems were significantly heterogeneous for variability (Table S5). The SSOG2017 classification was identical to the FIGO2015 classification.

The STAN2007 system, but not the FIGO2015 and SSOG2017 systems, requires the presence of accelerations to classify a trace as normal. Accelerations were absent in 67/300 (22.3%) of traces; thus, 57/272 (21%) of these traces were re-classified from normal to suspicious in the STAN2007 system. Adjustments for lack of accelerations (Table S4) resulted in an increased heterogeneity with regard to both the FIGO2015 and SSOG2017 systems (statistical calculations not performed).

FHR decelerations

Decelerations were absent in 15 cases (Table S4). Among 285 cases with decelerations, 196 recordings (69%) showed decelerations during >50% of contractions. The lower part of Table S4 shows the distribution of decelerative patterns relative to repetitiveness of decelerations. Among 14 cases of pathological patterns in the FIGO2015 and SSOG2017 systems where decelerations occurred at ≤50% of uterine contractions, nine were single prolonged decelerations, where repetitiveness is disregarded, and five were variable decelerations lasting for >180 s. These latter five CTG traces were accordingly classified as normal in the FIGO2015 system. The number of discrepant cases between STAN2007 and FIGO2015 relative to repetitiveness and different classifications of variable decelerations was (31 + 38 + 5)/285 (26.0%) (Table S4).

Among decelerative patterns occurring at <50% of uterine contractions in the SSOG2017 system, six traces were classified as suspicious and 14 cases as pathological (Table S4). The 14 pathological cases were classified as for the FIGO2015 system (see above), and the six suspicious traces were all of type-variable decelerations lasting for 60–180 s with normal baseline and normal variability, i.e. they were accordingly classified as normal traces. The number of discrepant cases between STAN2007 and SSOG2017 relative to repetitiveness and different classifications of variable decelerations was (31 + 6+27 + 5)/285 (24.2%) (Table S4).

Of 31 suspicious patterns in STAN2007, all were classified normal by FIGO2015 because the FIGO2015 does not consider the depth of the variable decelerations (beat loss >60 bpm), in contrast to the STAN2007 system (Table S1). The same was true for the SSOG2017 system.

Among variable decelerations lasting 60–180 s, all 46 traces were classified as pathological in the STAN2007 system, whereas eight were pathological in the FIGO2015, and 13 in the SSOG2017 system (Table S4). In the FIGO2015 system, variable decelerations lasting 60–180 s are classified as normal unless they are U-shaped or the variability is abnormal (Figure S1, Table S2); in the SSOG2017 system these traces are classified as suspicious if there is normal baseline FHR and variability, but pathological if lasting for >20 min and with decreased variability or tachycardia (Figure S1, Table S3).

Testing for homogeneity showed significant heterogeneity STAN2007 vs. FIGO2015 and vs. SSOG2017 (Table S5). The tests for agreement showed an overall moderate agreement for STAN2007 vs. SSOG2017, though the PA was nil for the suspicious class.

Overall classification of CTG traces

Table 1 shows the distribution into the categories normal/suspicious/pathological classes relative to the different CTG classification systems. More CTG traces were classified as normal in the FIGO2015 system and more traces were classified as pathological in the STAN2007 system. When testing for homogeneity, there were significant discrepancies in all comparisons (Table 2). The classification systems showed moderate–strong degrees of agreement, with the highest agreement for FIGO2015 vs. SSOG2017 and the lowest for STAN2007 vs. FIGO2015 (Table 2). Among the CTG classes, normal and pathological CTG patterns showed moderate–strong agreements but suspicious patterns showed minimal or weak agreements between the three classification systems.

Classification of ST events

The most common ST event was a baseline rise of T/QRS ratio >0.05, occurring in 276/300 (92%). In 104/300 cases (35%) a single ST event occurred during the registration. The median number of ST events during a delivery was three, with a range from one to 109 events.

The distribution of significant and non-significant ST events is displayed in Table 3. Significant ST events were most common in the STAN2007 system and less common in the FIGO2015 system. There were 41 cases with discrepant ST event classification relative to the STAN2007 classification in the FIGO2015 system and 24 in the SSOG2017 system; between FIGO2015 and SSOG2017 there were 17 discrepancies. The agreements between the CTG classifications systems were moderate–strong, but the homogeneity tests showed significant discrepancies (Table 4).

Table 1. Overall classification of intrapartum cardiotocography (CTG) traces (n = 300).

	STAN2007		FIGO2015		SSOG2017	
CTG classification	*n*	%	*n*	%	*n*	%
Normal	151	51[a]	189	63	163	54
Suspicious	22	7	28	9	47	16
Pathological	127	42	83	28	90	30

STAN2007, CTG classification system used in the STAN interpretation algorithm from 2007 (3) (Table S1); FIGO2015, CTG classification system published by the International Federation of Gynecology and Obstetrics in 2015 (1) (Table S2); SSOG2017, CTG classification system introduced in Sweden by the Swedish Society of Obstetrics and Gynecology in 2017 (4) (Table S3).

[a] 50.33%, adjusted to 51% according to the largest remainder method.

Table 2. Tests for homogeneity, proportion of agreement (PA) and quadratic weighted (QW) Cohen's kappa index of CTG classes.

	Tests for homogeneity		Tests for agreement			
			Composite PA		QW kappa index	
	Friedman test (*p*)	Wilcoxon matched-pairs test (*p*)	PA	95% CI	Kappa	95% CI
STAN2007 vs. FIGO2015 vs. SSOG2017	<0.0001	–	–	–	–	–
STAN2007 vs. FIGO2015	–	<0.0001	0.80	0.75–0.85[a]	0.70	0.64–0.77
STAN2007 vs. SSOG2017	–	<0.0001	0.83	0.78–0.87[b]	0.86	0.83–0.89
FIGO2015 vs. SSOG2017	–	<0.0001	0.90	0.86–0.93[c]	0.92	0.91–0.93

CI, confidence interval.

[a] Category agreements: for normal patterns 0.78 (0.71–0.84), for suspicious 0.32 (0.18–0.49) and for pathological 0.61 (0.53–0.70).

[b] Category agreements: for normal 0.90 (0.84–0.94), suspicious 0.21 (0.12–0.34) and pathological patterns 0.67 (0.58–0.75).

[c] Category agreements: for normal 0.86 (0.80–0.91), suspicious 0.44 (0.31–0.59) and pathological patterns 0.92 (0.84–0.97).

Table 3. Classification of ST events into significant ST events (indicating action) and non-significant ST events according to STAN clinical guidelines (3,5) (*n* = 300).

	STAN2007		FIGO2015		SSOG2017	
	n	%	*n*	%	*n*	%
Significant ST events	135	45	94	31	111	37
Non-significant ST events	165	55	206	69	189	63

Table 4. Tests for homogeneity and agreement of significant ST events.

	Test for homogeneity	Tests for agreement			
		Composite PA		QW kappa index	
	McNemar test (*p*)	PA	95% CI	Kappa	95% CI
STAN2007 vs. FIGO2015	<0.000001	0.86	0.82–0.90	0.72	0.65–0.78
STAN2007 vs. SSOG2017	<0.000001	0.92	0.88–0.95	0.84	0.78–0.89
FIGO2015 vs. SSOG2017	0.000015	0.94	0.91–0.97	0.87	0.83–0.92

The CTG and ST event data of the 41 cases of discrepant classifications of ST events for STAN2007 vs. FIGO2015 are displayed in Table 5. In eight cases the different classifications of tachycardia had a decisive influence, and in 33 cases, the different classifications of variable decelerations lasting 60–180 s. To classify a CTG trace as suspicious in the STAN2007 system, two suspicious element categories are required: in two cases, these were tachycardia of 150–160 bpm plus deep variable decelerations with a beat loss of >60 bpm.

Evaluation of ST events relative to outcome variables in the CTG arm

Significant ST events occurred most often with STAN2007 (66/151), second most often with SSOG2017 (59/151) and less often with FIGO2015 (49/151) (McNemar's test: STAN2007 vs. FIGO2015, $p = 0.016$; STAN2007 vs. SSOG2017, $p = 0.5$; FIGO2015 vs. SSOG2017, $p = 0.062$). There were no significant differences between the classification systems in outcome variables (Table 6). One newborn had a low Apgar score at both one and five minutes (and metabolic acidosis), one newborn at one minute only, and one at five minutes only. No metabolic acidosis occurred in the latter two newborns. Thus, among the three newborns with metabolic acidosis, one had low Apgar scores. The newborn with metabolic acidosis not indicated by FIGO2015 had normal Apgar scores.

Discussion

This study showed that in comparisons with the STAN2007 classification system, the CTG patterns were classified discrepantly in the FIGO2015 and SSOG2017 systems. Although the tests for agreement indicated an overall moderate to good agreement between the systems, the tests for homogeneity all indicated significant heterogeneity. Overall, fewer CTG traces were classified normal in the STAN2007 system compared with the FIGO2015 and SSOG2017 systems.

When analyzing the individual CTG element categories, i.e. baseline FHR, variability, accelerations and decelerations, we found significant heterogeneity for all categories in the comparisons between STAN2007 and the other two systems. This was due to different cut-off limits of tachycardia, variability classification with or without presence of accelerations, and differences in repetitiveness and algorithms in classifying decelerations. A majority of

Table 5. Cases with discrepant classification of significant ST events STAN2007 vs. FIGO2015 (n = 41).

Type of CTG changes	*n*	CTG classification → classification of ST event		
		STAN2007	FIGO2015	SSOG2017
Tachycardia 150–160 bpm + deep variable decelerations	2	Suspicious → significant ST (n = 2)	Normal → non-significant ST (n = 2)	Normal → non-significant ST (n = 2)
Tachycardia >170 bpm	6	Pathological → significant ST (n = 6)	Suspicious → non-significant ST (n = 6)	Suspicious → non-significant ST (n = 6)
Variable decelerations 60–180 s + normal baseline, variability	28	Pathological → significant ST (n = 28)	Normal → non-significant ST (n = 28)	Suspicious → non-significant ST (n = 19), significant ST (n = 9)
Variable decelerations 60–180 s + tachycardia >160 bpm	5	Pathological → significant ST (n = 5)	Suspicious → non-significant ST (n = 5)	Pathological → significant ST (n = 5)

bpm, beats per minute.

Table 6. Maternal and perinatal outcomes relative to significant ST events in the CTG arm (n = 151). A case may show more than one outcome.

Outcome variables	Total number of outcomes[a]	Number of outcomes identified by significant ST events		
		STAN2007	FIGO2015	SSOG2017
Operative delivery for fetal distress	23	18	14	16
Apgar score <4 at 1 min	2	2	2	2
Apgar score <7 at 5 min	2	2	2	2
Apgar score <7 at 10 min	0	–	–	–
Umbilical cord artery pH <7.05	11	7	5	7
Metabolic acidosis	3	3	2	3
Neonatal intensive care admission	13	8	5	7

[a] For individual outcome variables, there were no significant differences between the systems (McNemar test, $p \geq 0.12$).

discrepant classifications of decelerations were attributed to differences in classifying variable decelerations.

The homogeneity and agreement statistics require that the same categories are represented in all compared groups. Since the SSOG2017 system lacks the category suspicious in classifying variability, and the FIGO2015 system in classifying variability and decelerations, some of the statistical analyses could not be performed.

More ST events were classified significant by the STAN2007 system (45%) than by the FIGO2015 system (31%) or SSOG2017 system (37%). Even though the agreements between the systems were moderate–strong, the tests for homogeneity constantly showed significant discrepancies. The causes of discrepancy were due to different classifications of variable decelerations and tachycardia. For both CTG classes and ST events, the agreements with the STAN2007 system was slightly better with the SSOG2017 system than with the FIGO2015 system.

The focus of the study was the evaluation of the new CTG classification systems relative to the old system. However, it is also important to mention that in the comparisons of FIGO2015 vs. SSOG2017 there were strong to almost perfect agreements in overall CTG classification and ST event classification, though here also the homogeneity statistics showed significant discrepancies.

The present study was a simulation study and did not show the in vivo impact of different CTG classification systems on STAN interpretation and perinatal outcome, as it would have done had the study been performed in reality. In another simulation study including three different CTG classification systems, Santo et al. (6) demonstrated that CTG classification by different systems may result in highly different sensitivity and specificity figures in indicating cord blood acidemia.

We are not aware of any prospective randomized study evaluating the impact of different CTG classification systems on perinatal outcome. In the discussion about STAN RCTs, the 3-tier CTG classification system used in the RCT on CTG only vs. STAN in the USA (15) has been criticized for not being discriminating enough (16–18). The American 3-tier system differed from the 4-tier STAN2007 system used in the European RCTs mainly in that the “yellow zone” (category II) in the American classification covered many of the CTG traces classified as either suspicious or pathological in the STAN2007 system. Yellow zone CTGs are “indeterminate” in that they are neither clearly normal or abnormal (19) and these traces are inconsistently associated with fetal acidemia, making the clinical management uncertain (6). When incorporated in the STAN interpretation algorithm, the lack of discriminatory ability between mild and more severe CTG abnormalities seems to be a weakness in the American RCT.

Of the 300 CTG traces selected for the study, 100 were originally classified as normal by the SwRCT authors, 100 as suspicious and 100 as pathological in the STAN2007 system, but in the present study the classification performed by two experts in consensus, the corresponding distribution was 151, 22 and 127. It was beyond the scope of the study to address the problem of interobserver agreement in CTG interpretation, but these figures give a composite PA of only 0.38 (95% CI 0.32–0.44) and a QW kappa of only 0.17 (95% CI 0.09–0.26) between original and expert observers. Such poor figures is no novelty, since a high intra- and interobserver disagreement in CTG classification has repeatedly been reported in the literature, as summarized by Santo et al. (6,20). In particular, there are discrepancies in the classifications of variability, decelerations and the overall classification. It was from this starting point of poor reproducibility that the updating of the old FIGO1987 to FIGO2015 evolved, with the aim to keep CTG guidelines as simple and objective as possible (1,20).

Since information about ST events was not available to the managing midwives and obstetricians in the CTG arm of the SwRCT, we aimed to evaluate the associations between significant ST events and maternal and perinatal outcomes in this part of the series. Although there were differences between the systems in occurrence of significant ST events, there were no statistically significant differences in the correct identification of maternal and perinatal outcome parameters.

It may seem odd that the statistics for homogeneity showed significant discrepancies when agreement tests showed good agreements. However, the kappa index as well as other measures of reliability and diagnostic accuracy are dependent on the prevalence of different observations and thus reflect the population characteristics (12). A low kappa value might reflect the inability of a diagnostic measure to identify rare conditions (12) and we therefore tried to simulate a high prevalence of adverse outcomes by compiling the study series with an over-representation of pathological CTG traces. The 100 pathological CTG traces in the study series represented 46.5% of all pathological traces with ST events in the material of 4820 cases available for the study, indicating that the study series represented a true high-risk population.

The labels no/minimal/weak/moderate/strong/perfect agreement do not alone indicate the clinical relevance of an agreement. Kottner et al. (12), who have outlined guidelines for performing reliability and agreement studies, have stated that even when a high agreement is obtained, the agreement might be clinically unacceptable due to a too high level of disagreement. That was the circumstance in the present study. The magnitude of acceptable differences is not only a statistical decision but, above all, a clinical decision; if important decisions are made upon these estimates, the level of agreement should be at least 0.90–0.95 according to Kottner et al. (12). Intrapartum electronic fetal surveillance leads to vital decisions with far-reaching consequences for mother and neonate, but in the present study no agreement or reliability statistics, i.e. composite PA or QW kappa indices, reached that high level for STAN2007 vs. FIGO2015 or vs. SSOG2017.

The FIGO2015 and SSOG2017 systems have been introduced without prior clinical or simulated studies and it remains to be shown whether they are safe in the clinical setting. The present study indicates that the new classification systems have a higher threshold for classifying CTG traces as abnormal; the new systems were introduced to make CTG interpretation as simple and objective as possible (1,4), but Santo et al. (6) have notified that a classification system with a comparably lower proportion of pathological traces might imply a higher neonatal risk. On the other hand, a high proportion of pathological traces might imply a higher rate of obstetric interventions (7).

A strength of the study is that the evaluations of CTG and ST events were made by two expert clinicians with long experience of CTG and STAN monitoring. To avoid bias by interobserver discrepancies in this inter-guidelines study, the CTG classifications were performed in consensus between the experts. However, although they were experienced, they had no previous experience in using the FIGO2015 and SSOG2017 systems. In terms of that, we believe that the detailed colored CRF substantially simplified the classification.

To date, STAN has been introduced in 30 countries in Asia, Australia, Europe and North America (A. Mårtendal, personal communication). However, the clinical bearing of STAN is debated. Ten meta-analyses (MAs) (21–25, and references 8–13 in Blix et al., 2016) have been performed on four or more of the six RCTs on CTG only vs. STAN (7,15, and references 3, 5–7, 25 and 26 in Blix et al., 2016), including more than 26 000 randomized women, and more than 20 clinical studies have been published (for details, see 26), but there is still confusion as to how to validate the results and there is no general endorsement of STAN. Adding to the confusion, both the RCTs and MAs have been criticized for methodological inconsistencies (21,27). Three MAs contain all six RCTs (22–24). In two of them, one by Saccone et al. (24) and one by Neilson in a Cochrane systematic review (22), no significantly improved perinatal outcome was found by using STAN. However, these authors mixed up original and revised RCT data and they reported metabolic acidosis in blood and extracellular fluid as being equivalent. Only the MA by Blix and co-workers (23) handled original RCT data correctly (28). The Blix MA was a trial sequential MA (29), which is a more conservative statistical method where conclusions have the potential to be more reliable than those using the traditional aggregate MA technique (30,31). However, Blix et al. (23) do not consider their finding of a 36% significant reduction in neonatal metabolic acidosis rate enough to justify the use of STAN, arguing that metabolic acidosis is a surrogate outcome measure and that only “hard endpoints” such as mortality and long-time outcome are valid outcomes, that only a >50% reduction in metabolic acidosis is clinically important, and that the absolute risk reduction was only 0.25% in the population (23,31,32).

In summary, the study showed that the STAN2007, FIGO2015 and SSOG2017 classification systems were not exchangeable in the STAN interpretation algorithm with regard to classification of CTG patterns and ST events. The discrepancies between the new and old classifications were attributed to differences in definitions of tachycardia and classifications of variable decelerations. Hence, the accuracy and safety of the FIGO2015 and SSOG2017 systems compared with the STAN2007 system remain to be evaluated, as well as the benefits of the new systems. A retrospective simulation study such as the present study is only a proxy for how abnormal CTG patterns and differences in classifying significant ST events would relate to adverse outcomes had the study been performed in vivo, but we found no solid support for an innocuous incorporation of the FIGO2015 or SSOG2017 CTG classification systems into the STAN interpretation algorithm. Future studies should address the effectiveness of the new CTG classification systems relative to the old system in identifying fetuses at risk of developing severe hypoxia, alone or combined with ST analysis, and the impact on maternal outcome as well.

Acknowledgments

Neoventa Medical AB is gratefully acknowledged for providing the study material from the SwRCT database and the STAN Viewer computer software for offline analyses.

Funding

This study was not supported by grants.

References

1. Ayres-de-Campos D, Spong CY, Chandraharan E, for the FIGO Intrapartum Fetal Monitoring Expert Consensus Panel. FIGO consensus guidelines on intrapartum fetal monitoring: cardiotocography. Int J Gynecol Obstet. 2015;131:13–24.
2. FIGO Subcommittee on Standards in Perinatal Medicine. Guidelines for the use of fetal monitoring. Int J Gynecol Obstet. 1987;25:159–67.
3. Amer-Wåhlin I, Arulkumaran S, Hagberg H, Marsál K, Visser G. Fetal electrocardiogram: ST waveform analysis in intrapartum surveillance. BJOG. 2007;114:1191–3.
4. Holzman M, Jonsson M, Weichelbaum M, Herbst A, Ladfors L, Nordström L. Nya svenska riktlinjer för CTG-tolkning under förlossning [New Swedish guidelines for intrapartum CTG interpretation] (in Swedish). Swedish Soc Obstet Gynecol. Medlemsbladet. 2016;4:33–4.
5. Rosén KG, Mårtendal A. The physiology of fetal surveillance. The green book of Neoventa part I. Gothenburg: Neoventa Medical AB, 2014.
6. Santo S, Ayres-de-Campos D, Costa-Santos C, Schnettler W, Ugwumadu A, Da Graca LM, for the FM-Compare Collaboration. Agreement and accuracy using the FIGO, ACOG and NICE cardiotocography interpretation guidelines. Acta Obstet Gynecol Scand. 2017;96:166–75.
7. Amer-Wåhlin I, Hellsten C, Norén H, Hagberg H, Herbst A, Lilja H, et al. Intrapartum fetal monitoring: cardiotocography versus cardiotocography plus ST analysis of the fetal ECG. A Swedish randomized controlled trial. Lancet. 2001;358:534–8.
8. Grant JM. The fetal heart trace is normal, isn’t it? Lancet. 1991;337:215–8.
9. Wikipedia. Lehmer random number generator. Available online at: https://en.wikipedia.org/wiki/Lehmer_random_number_generator (accessed 25 November, 2017).
10. The MathsLinks network. Number of the day. Available online at: https://mathsstarters.net/numoftheday (accessed 25 November, 2017).
11. Wiberg N, Källén K, Olofsson P. Base deficit estimation in umbilical cord blood is influenced by gestational age, choice of fetal fluid compartment, and algorithm for calculation. Am J Obstet Gynecol. 2006;195:1651–6.
12. Kottner J, Audigé L, Brorson S, Donner A, Gajeweski BJ, Hróbjartsson A, et al. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. J Clin Epidemiol. 2011;64:96–106.
13. VassarStats: website for statistical computation. Available online at: http://vassarstats.net/ (accessed 25 November, 2017).
14. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22:276–82. Available online at: http://www.biochemia-medica.com/2012/22/276 (accessed 25 November, 2017).
15. 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.
16. Gucciardo L, Blavier F, Faron G. Intrapartum fetal ECG ST-segment analysis. N Engl J Med. 2015;373:2480.

17. Yli B, Hahn T, Kessler J, Lie HK, Martinussen M. Bigger is not always better . . . The validity of the US randomized trial of STAN for Norway. Available online at: http://www.neoventa.com/2015/11/bigger-is-not-always-better/ (accessed 25 November, 2017).
18. Xodo S, Saccone G, Schuit E, Amer-Wåhlin I, Berghella V. Why STAN might not be dead. J Matern Fetal Neonatal Med. 2017;30:2306–8.
19. Macones GA, Hankins GDV, 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.
20. Santo S, Ayres-de-Campos D. Human factors affecting the interpretation of fetal heart rate tracings: an update. Curr Opin Obstet Gynecol. 2012;24:84–8.
21. 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.
22. Neilson JP. Fetal electrocardiogram (ECG) for fetalmonitoring during labour. Cochrane Database Syst Rev. 2015;12:CD000116.
23. Blix E, Brurberg KG, Reierth E, Reinar LM, Øian P. ST waveform analysis versus cardiotocography alone for intrapartum fetal monitoring: a systematic review and meta-analysis of randomized trials. Acta Obstet Gynecol Scand. 2016;95:16–27.
24. Saccone G, Schuit E, Amer-Wåhlin I, Xodo S, Berghella V. Electrocardiogram ST analysis during labor. A systematic review and meta-analysis of randomized controlled trials. Obstet Gynecol. 2016;127:127–35.
25. Vayssière C, Ehlinger V, Paret L, Arnaud C. Is STAN monitoring associated with a significant decrease in metabolic acidosis at birth compared with cardiotocography alone? Review of the three meta-analyses that included the recent US trial. Acta Obstet Gynecol Scand. 2016;95:1190–1.
26. Amer-Wahlin I, Kwee A. Combined cardiotocographic and ST event analysis: a review. Best Pract Res Clin Obstet Gynaecol. 2016;30:48–61.
27. 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;93:556–69.
28. Olofsson P. Belittling of a significant decline in neonatal metabolic acidosis rate achieved by STAN monitoring. Acta Obstet Gynecol Scand. 2016;95:604–5.
29. Brok J, Thorlund K, Gluud C, Wetterslev J. Trial sequential analysis reveals insufficient information size and potentially false positive results in many meta-analyses. J Clin Epidemiol. 2008;61:763–9.
30. Thorlund K, Engstrøm J, Wetterslev J, Brok J, Imberger G, Gluud C. User manual for Trial Sequential Analysis (TSA). Copenhagen Trial Unit 2017. Available online at: http://www.ctu.dk/tsa/files/TSA%20manual.pdf (accessed 25 November, 2017).
31. Blix E, Brurberg KG, Reierth E, Reinar LM, Øian P. STAN technology, surrogate outcomes and possible sources of bias. Acta Obstet Gynecol Scand. 2016;95:608–9.
32. Blix E, Brurberg KG, Reierth E, Reinar LM, Øian P. Statistical significance is not necessarily equal to clinical significance. Acta Obstet Gynecol Scand. 2016;95:1192.

Supporting information

Additional Supporting Information may be found in the online version of this article:

Figure S1. Case report form used for classification of cardiotocography (CTG) patterns and fetal ECG ST events. STAN2007 denotes the CTG classification system used in STAN clinical guidelines from 2007 (3) (Table S1), FIGO2015 the system published in 2015 by the International Federation of Gynecology and Obstetrics (1) (Table S2), and SSOG2017 the system published in 2017 by the Swedish Society of Obstetrics and Gynecology (4) (Table S3).

Table S1. The STAN2007 cardiotocography (CTG) classification system, modified from Amer-Wåhlin et al. (3).

Table S2. The FIGO2015 CTG classification system (1).

Table S3. The SSOG2017 CTG classification system (4), translated from the Swedish.

Table S4. Distribution of element fetal heart rate (FHR) patterns and classifications. Values are number of cases.

Table S5. Tests for homogeneity and agreement of CTG element categories.

Appendix S1. Description of statistical tests used.