FDA Executive Summary

Lead Reviewer:
Kathryn S. Daws-Kopp

Prepared May 25, 2005, for the
June 23, 2005, meeting of the
Obstetrics & Gynecology Devices Panel

Corrected, June 8, 2005

P020001/A011
Neoventa Medical AB
STAN® S31 Fetal Heart Monitoring System

Contents

I. Proposed Indications for use 3
II. Device Description 3
II. A. Principle of Operation (Fetal Pathophysiology) 4
II. B. Device Components 5
II. C. Display Output 6
II. D. Algorithm Characteristics 6
III. Pre-clinical Studies 7
III. A. Pre-clinical Bench Testing 7
III. B. Pre-clinical Animal Studies 7
IV. Regulatory History 10
V. Background Studies 11
V. A. Early Clinical Studies 11
V. A. 1. Early Observational Studies in Sweden and the UK 12
V. A. 2. Plymouth (United Kingdom) RCT 12
V. A. 3. EU Multicenter Study 13
V. B. Non-interventional, Observational Studies 13
V. B. 1. Nordic Study 13
V. B. 2. Canadian Study 14
V. B. 3. Utrecht (Netherlands) Study 14
V. C. Ancillary Clinical Studies 15
V. C. 1. Gothenburg (Centers of Excellence) 15
V. C. 2. EU Project (Centers of Excellence) 16
V. D. Swedish RCT (SRCT) 16
V. D. 1. SRCT Study Design 17
V. D. 2. SRCT Results: Overview and Primary Endpoints 18
V. E. Conclusions 21
VI. United States Studies 22
VI. A. United States Education Study (Education Study) 22
VI. A. 1. Education Study Design 23
VI. A. 2. Education Study Results 25
VI. A. 3. Additional Analyses 26
VI. B. United States Clinical Use Study (CUS) 28
VI. B. 1. CUS Study Design 29
VI. B. 2. CUS Results 31
VI. B. 3. Additional Analyses 35

This document addresses the PMA P020001, Amendment 11, for the STAN S31 fetal heart monitoring system that is currently under review. This document serves as a summary of the PMA including descriptions of the development of the product and the regulatory history that has brought us to this point in time. The regulatory history for this product includes the original submission of the PMA and the subsequent Panel Meeting in April 2002. The Swedish Randomized Clinical Trial (SRCT) which is summarized in this document was discussed at that Panel Meeting in 2002. That Clinical Trial, while an important milestone in the regulatory history of this PMA, is not the subject of our Panel Meeting on June 23, 2005. The upcoming Panel Meeting will focus on the Education and Clinical Use studies conducted in the United States since the time of the SRCT per FDA's recommendation to the sponsor.

I. Proposed Indications for use

The proposed indication for use for the STAN system is for use in pregnancies at term (>36 completed gestational weeks) to improve the assessment of the fetal condition during labor and as an adjunct to standard fetal heart monitoring.

Use of the STAN system is indicated when there is planned vaginal delivery and:

- there is need for close fetal surveillance during labor, or
- there are maternal disorders and/or utero-placental dysfunction with potential adverse influence on fetal oxygen and nutritional supply, or
- there is deviation from the normal course of labor including induction/augmentation of labor.

II. Device Description

This device is a perinatal monitoring system with ST analysis capabilities. It has many of the features of typical perinatal monitors, which are regulated as Class II devices.

The Class III, PMA, portion of this device is the fetal ECG (FECG) analysis portion. Previous fetal monitoring devices have collected FECG information, using a fetal spiral electrode. It is from this signal that the fetal heart rate (FHR) is derived. Conventional FHR monitoring devices merely record the time between two consecutive "R" portions of the QRS (contraction of the ventricles) segment of the ECG. This R-R interval determines the heart rate. The ECG information is not displayed or analyzed as part of these conventional monitoring devices.

The STAN system more fully realizes the potential data acquired by the scalp electrode by deriving the T/QRS amplitude and ST segment slope and comparing the most recent values to baseline values derived during the first 20 minutes of data collection. As output, it provides automatic identification of three types of ST events on the monitor and tracing: episodic T/QRS rise, baseline T/QRS rise, and 3 grades of biphasic ST.

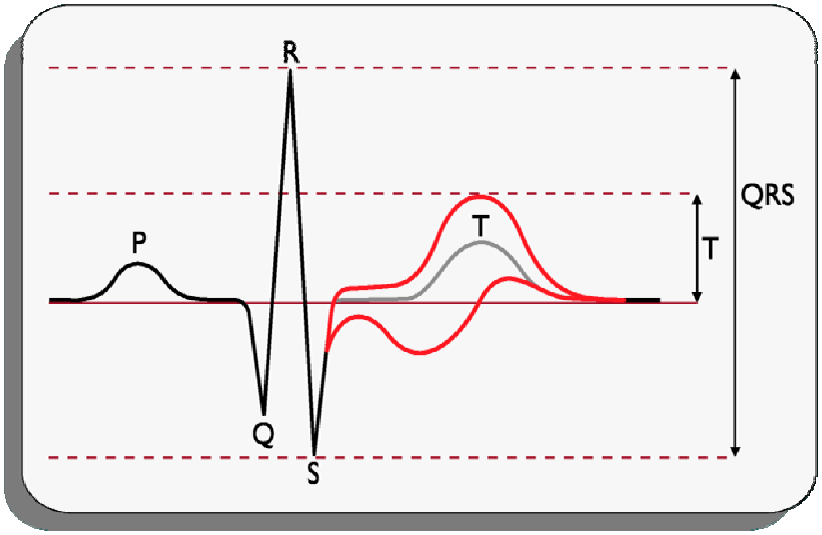
It should be noted that if a compromising event occurs prior to monitoring or during the first 20 minutes of data collection, these events may impact the effectiveness of the device.

II. A. Principle of Operation (Fetal Pathophysiology)

The clinical guiding principle for the STAN S31 system is that the physiology of ST waveform change is better understood than fetal heart rate change, and the combined use of ST waveform change with FHR pattern may improve clinical management of the complex physiological responses which occur in labor.

The theory underlying the operation of this device is that a fetus exposed to oxygen deprivation will develop tissue hypoxia. Hypoxic myocardium responds by changing from aerobic to anaerobic metabolism. The ST segment and T-wave portions of the ECG respond by showing changes from baseline dependent on the duration of hypoxia. In general, the amplitude of the T wave increases with continuing hypoxia, while the QRS amplitude remains unchanged. This results in an increase in the ratio of T/QRS. One caveat should be noted: Once a fetus has already developed significant hypoxia, the fetal myocardium might not respond to hypoxia in the same way and thus, may not show the same changes from baseline.

In a normally oxygenated fetus in labor, the ST portion of the ECG is flat and near the baseline. As hypoxia becomes severe and persistent, the slope of the ST segment may become negative and drop below the baseline.



It is the changes in the T/QRS ratio and ST segment that are the basis for the three types of ECG abnormalities identified automatically by the STAN S31 system.

An “episodic” T/QRS rise means that the T/QRS ratio rises and returns to baseline in < 10 minutes. This corresponds to short lasting hypoxia that nonetheless requires a shift to anaerobic metabolism.

A “baseline” T/QRS rise means that the rise in T/QRS lasts for > 10 minutes. This occurs in situations in which the fetus responds to hypoxia with anaerobic metabolism; there is persistent stress and no opportunity for recovery.

A “biphasic” ST is a type of abnormal ST segment in the FECG. There are three types of biphasic STs. In order of increasing clinical concern, these are: (1) “Grade 1” biphasic ST which is a downward-leaning segment that is above the baseline for the ECG; (2) “Grade 2” biphasic ST which is an ST segment that cuts across the baseline; and (3) “Grade 3” biphasic ST where the entire ST segment is below baseline. Grades 2 and 3 biphasic STs are abnormalities which are clinically significant, especially if repeated.

II. B. Device Components

The STAN System is made up of the following basic components:

- Base Unit
- Monitor
- Software
- Sensor

The **base unit** houses the main electrical components and serves as the conduit for input and output connections such as those for data storage or access and accessories to the device (including the sensors). The **monitor** is the display. The **software** includes all the software for the standard monitoring functions as well as the ST analysis features. The **sensor** is an existing spiral electrode currently cleared for use in internal fetal heart rate monitoring. The standard monitoring functions available for this device (which are all class II functions) are:

- External uterine activity (tocodynamometry)
- Internal uterine activity (intrauterine pressure, IUP, catheter)
- External fetal heart rate (ultrasound)
- Internal fetal heart rate (spiral electrode)
- Event marker

Note that while the STAN® S31 is the subject of this PMA, clinical testing (Swedish RCT, Education study, and US Clinical study) was conducted using a previous model of the device, the STAN® S21. The differences between the versions pertinent to the PMA functions are as follows:

- Human factor enhancements
- Minor hardware enhancements
- Support for connectivity options (USB and Ethernet)
- External monitoring (ultrasound FHR)

The S31 includes external FHR via ultrasound, making the S31 capable of monitoring twins. The data acquisition and signal processing devices for the two models (S21, S31) are very similar. The transducers and patient cables for FECG are the same. The S31 uses updated design and construction techniques and has been demonstrated to meet the specifications of the original unit. Testing shows that the S31 hardware and software perform equivalently to the model S21.

II. C. Display Output

The device displays the data by placing x's that represent the T/QRS ratio below the fetal heart rate and uterine activity tracing (both paper and electronic versions), corresponding to the specific time and annotated with the source of the event (episodic T/QRS, baseline T/QRS, or biphasic ST and grade.) One of the optional display windows is the Event log. The Event log displays these ST events, other automatically detected events, and user notes, all in chronological order.

The display screen is composed of several display areas or windows that provide various information including: data from other parameters collected, patient identifying information, status messages, signal quality indicators, and other information to help maximize the quality of the various signals received and displayed. Two additional windows are pertinent to the ST analysis functions. One is the ECG average window which displays the smoothed ECG complex. The second is the FECG Signal window which displays a real-time view of the raw FECG signal. This window is intended to facilitate establishment of a good FECG signal and disappears after a good signal is achieved (but may be recalled by the user when needed).

II. D. Algorithm Characteristics

The STAN algorithms use two signal characteristics derived from the FECG to determine events: ST segment slope differences and T/QRS amplitude differences. The algorithms generally operate by comparing the latest sample (either an ST segment or a T/QRS value) to an established baseline to see if there is a change.

The STAN system calculates an average ECG waveform from the FECG channel (scalp-to-skin lead). The averaging is performed over 30 qualified beats. The device uses the average ECG waveform to process the T/QRS ratios – the amplitude between of the T wave versus a baseline value, and the QRS complex amplitude. A T/QRS baseline is computed every minute and monitored for multiple characteristics, contributing to a determination of a T/QRS difference and the identification of a significant event. The initial 20 minutes are used to collect T/QRS baseline data to allow for a robust determination of starting values used by the processing algorithms for event detection.

The ST analysis software conditions and analyzes the raw ECG signal to identify additional characteristic parameters (QRS complex duration, and P-QSR separation times) used in decision making. The decision algorithm evaluates these parameters looking for the 3 types of events: episodic T/QRS rise, baseline T/QRS rise, and biphasic ST. Biphasic is a term that indicates that the slope of the ST waveform has become negative, which the decision algorithm uses as an indicator of fetal abnormality. Biphasic events are further classified into category 1, 2, or 3 indicating that the slope is above baseline, crossing the baseline, or below baseline, respectively.

The device indicates when the signal quality is poor or when breech mode is activated (although it is specifically contraindicated for use in breech presentations). When the signal quality is bad and during the first 20 minutes of operation, the device will not register an event. Signal quality, as with internal FHR monitoring is dependent on the sensor making good contact with the fetal scalp.

III. Pre-clinical Studies

III. A. Pre-clinical Bench Testing

Electrical, Thermal and Mechanical Testing

The STAN system was evaluated for electric shock, fire and mechanical safety including, leakage current tests, rigidity and rough handling tests, and a normal temperature test. The STAN system was found to be compliant with IEC 60601-1.

EMC Testing

The STAN system was evaluated for electromagnetic compatibility including emissions and immunity tests performed by an independent laboratory. The STAN system was found to be compliant with IEC 60601-1-2.

Biocompatibility/Sterilization

The only component of the system to which biocompatibility and sterilization requirements apply is the sensor; a spiral electrode. The spiral electrode being used has already been cleared under 510(k). Biocompatibility and sterilization were evaluated as part of the 510(k) clearance process and the information provided was found to be adequate.

Software Validation

Software was validated to assure that it performed its intended functions. Validation consisted of testing, analysis, and inspection activities. System level software quality assurance (SQA) testing, unit level testing and integration testing were performed on the STAN software. Verification was performed using suites of simulated and recorded signals, using human-scored results as applicable to assess correctness. Extensive test suites were developed to assess the performance of the ECG analysis algorithm and its ability to reliably identify the salient waveform parameters. All testing was consistent with FDA guidance for software at a moderate level of concern.

Risk Management Activities

Risk Management was initiated in the Definition and Requirement phases of product development and conducted throughout the design and development activities for the STAN system. The purpose of the Risk Management activities were to identify and control potential hazards arising from any STAN system. Risk Management involved identifying potential hazards, estimating and evaluating the associated risks, and reducing these risks to acceptable levels. The criteria for risk acceptability were generated using the processes in ISO 14971, Application of risk management to medical devices.

III. B. Pre-clinical Animal Studies

The sponsor provided detailed summaries of approximately eleven pre-clinical (animal) studies that were conducted between 1971 and 1989 that led to the first large-scale clinical investigation using a prototype of the STAN S31 fetal heart monitoring system.

The sponsor included summaries of these studies in chart format. Each of these studies is published and references are provided.

Basically, two types of experiments were performed in guinea pigs, cats, and most commonly in lambs. In the first type of experiment, carried out primarily during the years 1971-1984, and one study from 1988-1989, anesthetized animal fetuses were exteriorized outside of the mother and then subjected to graded hypoxia. The numbers of animals involved in each of these studies was 14-40.

The objective of these early studies was to obtain a basic understand of metabolic and neurologic aspects of the physiology of fetal hypoxia and asphyxia, as described below:

- to investigate the vagal influence on the fetal hypoxic bradycardia and to identify changes in the FECG during graded hypoxia;
- to investigate how the depletion of glycogen, creatine phosphate and ATP in the fetal heart, liver and brain correlates with FECG changes during graded hypoxia;
- to analyze the metabolic background of hypoxic FECG changes;
- to investigate whether beta-adrenoreceptor stimulation and hypoxia exerted additive or potentiating effects of the FECG and several cardiovascular parameters, and whether hypoxic changes in the FECG could be blocked by beta-adrenoreceptor blocking agents;
- to test the influence of maternal terbutaline therapy on fetal reactions to hypoxia; and
- to study the ST waveform of the FECG, short-term heart rate variability and plasma catecholamines in normal and growth-retarded fetuses during hypoxia.

The following conclusions were drawn from these studies:

- Fetal bradycardia is due to an AV block, type II.
- Changes in FECG pattern can be regarded as a sign of myocardial glycogenolysis and early hypoxic stress.
- In lamb fetuses exposed to hypoxia, accumulation of lactate occurs in parallel to glycogen depletion.
- Mild hypoxia initiates the FECG changes via an activation of beta-adrenoreceptors while severe hypoxia induces FECG changes through a direct effect.
- Fetal myocardial beta receptors can be stimulated to their maximum capacity by high maternal doses of beta-agonist combined with hypoxia, and beta-agonist administration to the mother diminishes the fetal defense for hypoxia.
- Hypoxia-induced FECG waveform changes differ between intrauterine growth retarded guinea pig fetuses and their normal-sized litter mates.

In the second type of pre-clinical animal studies, pregnant ewes were anesthetized and catheterized so that ECG electrodes were implanted subcutaneously in the fetus. The number of ewes in each study ranged from 2-16. These studies were carried out from 1977-1987. The purpose of these studies was to evaluate FECG changes in response to hypoxia. The specific objectives of these studies were:

- to study the effect of hypoxia on the ECG waveform in the healthy but chronically instrumented fetal lamb and to explore the possibilities of automatic sampling and quantification of ST waveform changes;
- to test the hypothesized relationship between the concentration of circulating catecholamines and T-wave amplitude;
- to compare the different means of fetal surveillance with special emphasis on ST waveform analysis of the FECG during spontaneous labor in the chronically instrumented fetal lamb;
- to study long-term changes in the FECG and the relationship between FECG changes and spontaneous labor, hypoxia and intrauterine death; and
- to investigate the changes in the FECG and the release of catecholamines in fetal lambs of different maturity grades, during non-acidemic fetal hypoxia.

The conclusions reached from these studies can be summarized as follows:

- ST waveform elevation, expressed as the T/QRS ratio, identifies a shift to anaerobic myocardial metabolism.
- This shift to anaerobic myocardial metabolism is mediated by beta-adrenergic stimulation.
- Mature lamb fetuses maintain central nervous system integrity until there are signs of cardiovascular collapse.
- ST waveform changes are associated with intact fetal lamb cortical activity and precede fetal death by hours.
- Uterine contractions induce increases in T/QRS ratio. If ST elevation is normalized between contractions, the lamb fetus seems to compensate for this moderate hypoxia. When oxygenation is further reduced, T wave remains elevated between contractions.
- Lamb fetal death in utero, whatever the cause (e.g., bleeding, infection or spontaneous hypoxia) is always preceded by marked ST waveform changes.
- Basal catecholamine concentration does not differ between immature and mature lamb fetuses; however, during non-acidemic hypoxia catecholamine levels increase significantly more in the mature fetus.
- In mature lamb fetuses, a linear correlation is found between plasma epinephrine concentration and T/QRS ratio.

- Both immature and mature lamb fetuses show marked bradycardia during non-acidemic hypoxia, but differ in how rapidly they recover. Mature fetuses have a slower acceleration of the heart rate. Mature fetuses also respond with a significant rise in mean arterial blood pressure.

- Fetal maturity has a significant influence on:

    - release of catecholamines during non-acidemic hypoxia;

    - cardiovascular reaction pattern; and

    - changes in the ST waveform.

IV. Regulatory History

The company submitted the original PMA on January 2, 2002, using the Swedish Randomized Control Clinical Trial described below as the main data basis for approval. The PMA was presented to the Ob/Gyn Devices Panel on April 22, 2002, where the Panel voted 6-5 that the PMA was not approvable. FDA agreed with the Panel and sent a Not Approvable letter to the sponsor on June 27, 2002. The most important deficiencies in the original PMA were related to the fact that the STAN had never undergone clinical testing in the United States at the time of the original PMA submission; the Panel expressed its concern that differences in labor management between Sweden and the United States could influence results. FDA, in agreement with the Panel, indicated to the sponsor in our Not Approvable letter that the sponsor had not adequately demonstrated that device effectiveness could be extrapolated from the Swedish Randomized Controlled Trial (SRCT) to the US. In addition, both the Panel and FDA had concerns related to the adequacy of the STAN training program for US clinicians. FDA also requested that the sponsor Americanize the existing clinical management guidelines (a set of charts developed by the sponsor to guide the user in determining how to respond to the combination of FHR and ST information) and modify the printout chart speed.

The Panel had recommended that the sponsor conduct a new prospective, randomized clinical study in the US to overcome the deficiency regarding demonstration of device effectiveness. (The primary effectiveness endpoint of the SRCT was cord artery metabolic acidosis.) FDA considered the Panel's view but concluded that the approximately 5000-patient SRCT was sufficiently robust in design that a new RCT was not necessary. FDA worked with the sponsor on a plan to correct the deficiencies in the original PMA. As of June 2003, the two main elements of this plan were:

- US Education Study to test whether, following STAN training, US labor and delivery staff improves in their ability to decide whether and when to intervene compared to when only FHR data are available. Truth regarding whether and when to intervene will be based on (1) umbilical artery blood gases and (2) consensus of STAN experts.
- A prospective US Clinical Use Study (CUS) in which the performance of US clinicians regarding the decision to intervene will be compared with that of Swedish experts who will review the tracings retrospectively.

In a closed session held June 9, 2003, Neoventa and FDA presented the above plan for resolving the PMA deficiencies to the Panel. Five of the six Panel members (and four of the five perinatologists) who voted against the PMA at the April 2002 meeting were present. FDA posed several discussion questions to the Panel regarding the acceptability of the revised clinical management guidelines, and the two-part US clinical study outlined above. The Panel agreed with the revised guidelines. However, two Panel members questioned whether agreement between US investigators and STAN experts on the need to intervene was sufficient as a primary endpoint for the US clinical use study.

The Panel as a whole did not endorse the suggestion of one Panel member that the US study be re-designed to evaluate neonatal outcomes, as did the SRCT. Another Panel member suggested that an alternative to this suggestion would be to evaluate, in non-randomized fashion, the rates of acidosis and clinical interventions, e.g. cesarean and forceps delivery, prior to and following STAN training. Such an observational study might show whether such outcomes improved following introduction of STAN into US practice. Several Panel members recommended that the study size be increased from the planned 350 cases and three institutions to ensure adequate contribution from each investigator and to ensure representation for both academic and community hospitals. Panel members were also concerned that all US investigators be properly trained in STAN technology and clinical use prior to participating in the US clinical study. The general consensus was that the approach was acceptable, although no vote was taken.

FDA interpreted the Panel discussion to signal acceptance by the Panel majority that (1) the SRCT demonstrated the safety and effectiveness of the STAN S21; and (2) the US study could demonstrate that the STAN technology, with proper training, could be integrated into US clinical practice. The Panel left the details of both studies to the FDA and the sponsor to finalize.

V. Background Studies

In this Executive Summary, we are including discussions of a number of reference studies using the STAN system. We separated our discussion of these studies into 3 sections: Early Clinical Studies (this section); Non-interventional, Observational studies; and Ancillary Clinical studies. The Early Clinical Studies preceded the SRCT and were used in the development of both the device and clinical guidelines. With one exception, the Non-interventional, Observational studies were commenced after the SRCT was completed (one study overlapped the SRCT). The Ancillary Studies both occurred after the SRCT and report on experience with the device in real use. Discussions of these studies are included because they contribute to our knowledge of use of the device.

All of the studies discussed in the following 3 sections are also presented in more detail in tab 2 of the Sponsor's portion of the Panel Package. Some of these early studies used an older or prototype version of the STAN; the STAN 8801.

V. A. Early Clinical Studies

As noted above, the studies in this section preceded the SRCT. They were used in the development of the device and management concepts, such as the clinical guidelines the sponsor developed to guide interpretation of the information provided and

management of patients. These studies are all discussed in tab 2 of the sponsor's portion of the Panel Package

V. A. 1. Early Observational Studies in Sweden and the UK

Clinical studies using prototypes of the STAN S31 began in the late 1980s. The first three studies using the prototype STAN system were observational studies conducted in Sweden and in the United Kingdom. The first of these was a Swedish study of 201 fetuses at greater than 34 gestational weeks in vertex presentation. The FECG signal was obtained from a single spiral electrode on the fetal scalp with a reference electrode on the maternal thigh. The purpose of this study was to evaluate ST waveform changes during labor using the older STAN 8801 fetal heart monitoring system. Signal quality was a problem in this first clinical study of this STAN prototype. Available data showed that it was possible to use the system for on-line recording of the unfiltered FECG waveform and to analyze the ST waveform. Acute fetal hypoxia corresponded with a rapid rise in T/QRS ratio on the FECG. The study also suggested that T/QRS ratio is less sensitive to respiratory acidosis than to metabolic acidosis. Finally, the study suggested that a T/QRS ratio of < 0.25 identifies with 99.3% accuracy a fetus with *normal* buffering capacity, independent of FHR tracing.

The second observational study took place in the United Kingdom. The objective of this study was to compare the sensitivity and positive predictive value (PPV) between Cardiotocogram (CTG) and elevated T/QRS ratio, with fetal acidemia during labor and at delivery. Term fetuses of 113 women in labor were monitored using the STAN 8801 system. This study showed that the PPV for acidemia diagnosed solely from elevated T/QRS ratio was 71% and the sensitivity was 29%. T/QRS ratio did not correlate with scalp pH; however in combination with pathologic CTG, scalp pH values of <7.25 showed a strong correlation with increasing T/QRS.

The third observational study was also conducted in the United Kingdom. The objective of this study was to investigate the potential of ST waveform analysis during labor. Eighty-six term fetuses were monitored with the STAN 8801. The investigators concluded that the mean one-hour T/QRS ratio was not a good predictor of low Apgar scores and that if FECG analysis was to have a role in intrapartum fetal surveillance, it might be in conjunction with fetal heart rate data.

V. A. 2. Plymouth (United Kingdom) RCT

The first interventional study using the STAN prototype was conducted in Plymouth, Great Britain in the early 1990s. This was a randomized, controlled, prospective study to investigate if [FHR + ST analysis] could improve the predictive value of intrapartum surveillance compared to FHR only and thereby decrease interventions, without increased risk to the fetuses. Approximately 2400 fetuses at > 34 weeks gestation were randomized to monitoring with FHR only or [FHR + ST analysis]. The study demonstrated a 46% reduction ($p<0.001$) in operative deliveries for fetal distress, without an increase in operative deliveries for other reasons. There was no reduction in the rate of C-section for fetal distress. There was a trend towards less metabolic acidosis in the STAN arm of the study.

One of the things this study suggested was the significance of ST segment depression and fetal acidemia. As a result, the STAN 8801 was modified so as to provide automatic

identification of biphasic ST segments. This modification had been incorporated when the next clinical study took place.

V. A. 3. EU Multicenter Study

The European Community Multi-Center Trial was a multi-center, non-randomized *retrospective* assessment of changes in FECG waveform in cases of verified fetal hypoxia. Blinded reviewers evaluated the [FHR + ST] tracings for 320 labors and grouped them either into an intervention category or non-intervention category according to STAN guidelines.

This retrospective analysis showed that there were ST abnormalities evident in the tracings for 11/12 deliveries of infants with evidence of hypoxia or asphyxia. In one case, however, involving a newborn who showed signs of encephalopathy and later developed cerebral palsy, both the FHR and ST analysis were normal.
This study highlighted the potential value that automatic identification of biphasic ST events might contribute to [FHR + ST analysis] using the STAN system.
.

V. B. Non-interventional, Observational Studies

The following studies collected STAN data without providing it to the clinical user for use in patient management. Thus, this type of study provides information about what this device can contribute without compromising the outcome. This information can be predictive about where the device could contribute to useful intervention when the information is used.

It should be noted that one study included in this section, the Nordic Study, commenced prior to the SRCT. While it did not end until after commencement of the SRCT, it helped form the basis for the SRCT.

V. B. 1. Nordic Study

The article for this study is provided as tab 15 of the Sponsor's portion of the Panel Package, "Amer-Wahlin I, Bordahl P, Eikeland T, et al. ST analysis of the fetal electrocardiogram during labor: Nordic observational multicenter study. J Matern Fetal Neonatal Med 2002;12:260-6."

The objective of this study was to examine diagnostic power of CTG plus ST analysis to identify adverse labor outcomes such as neurological symptoms and/or metabolic acidosis. This multicenter observational study was conducted in 12 Nordic departments of obstetrics and gynecology and included a total of 573 women monitored during the period of June 1998 to January 1999. This period overlaps the conduct of the SRCT (December 1998 to June 2000). The study used an updated version of the STAN 8801 called the STAN ESST. The ESST model was essentially the same device as the STAN S21. The upgrade included an algorithm that automatically identified ST events.

As in the EC Observational Study, blinded reviewers evaluated the [FHR + ST] tracings and grouped them either into an intervention category or non-intervention category according to previously developed STAN guidelines. Fifteen infants were diagnosed as having been exposed to intrapartum hypoxia or asphyxia based on neurological symptoms and/or metabolic acidosis. The sensitivity of [FHR + ST] clinical guidelines to

recommend intervention in these cases was 100%. The specificity of the guidelines was 95%.

Also of interest in this study was that the projected rate of operative interventions, if intervention had been taken according to [FHR + ST] guidelines, would have been only 7.5%, compared to the actual rate of 15.3%. This observation reaffirmed the results of the Plymouth Trial.

V. B. 2. Canadian Study

The article for this study is provided as tab 16 of the Sponsor's portion of the Panel Package, "Dervaitis KL, Poole M, Schmidt G, et al. ST segment analysis of the fetal electrocardiogram plus electronic fetal heart rate monitoring in labor and its relationship to umbilical cord arterial blood gases. Am J Obstet Gynecol 2004;191:879-84."

The objective of this study was to assess the ability of electronic fetal monitoring (EFM) plus ECG to predict fetal metabolic acidemia (defined as umbilical cord pH<7.15 and BDecf of >=12mmol/L.) Recruitment took place from March 2002 to November 2003.

In this study, patients were managed with FHR only, as all ST-related information on the digital screen was covered. Retrospectively, trained and blinded observers after the delivery performed visual classification of the tracings following International Federation of Gynecology and Obstetrics (FIGO) guidelines. These classifications were combined with the automated ST analysis.

Based on the retrospective analysis of tracings of 143 women, for FHR+ST combined sensitivity was 43 %, specificity was 74%, negative predictive value was 96%, and positive predictive value was 8% for detecting metabolic acidemia at birth.

V. B. 3. Utrecht (Netherlands) Study

The article for this study is provided as tab 17 of the Sponsor's portion of the Panel Package, "Kwee A, van der Hoorn-van den Beld CW, Veerman J, et al. STAN® S21 fetal heart monitor for surveillance during labor: an observational study in 637 patients. J Matern Fetal Neonatal Med 2004;15:400-7."

This study was performed under the European Union Multicenter Project based on the center of excellence structure. This study was performed at the University Medical Center Utrecht (1800 deliveries per year) from 2000 – 2002.

The objective of this prospective observational study was to assess the value of the STAN fetal heart rate monitoring system for intrapartum fetal monitoring using CTG and ECG. A total of 637 high-risk pregnancies were monitored using a STAN. In 449 (70.5 %) all data for analysis were available. In 188 (29.5 %) the recordings were interrupted > 20 minutes before birth, there was a technical failure of recording, or umbilical cord gases were not available.

There were 18 cases with metabolic acidosis (4%) and three cases with low Apgar score only (0.7%). There were five cases with blood pH < 7.00 and in all of them the significant ST event occurred 18-31 min before birth. Among cases with blood pH of 7.00 -7.04, six of 13 had significant ST event (false negative rate of 7/449 or 1.6 %). However in those

cases without ST event there were no Apgar scores < 7 at 1 or 5 minutes and there were no neurological abnormalities. The three cases with low Apgar score only had no ST change, but recovered fully. There were a total of 41 interventions for fetal distress (20 in the presence of ST changes and 21 without it.) The mean blood pH was 7.09 in the ST change group and 7.16 in the group without it. ST changes occurred in all five cases of severe acidosis, 46% of cases with mild metabolic acidosis. The overall conclusion was that CTG and ST analysis were more specific in detecting fetal acidemia than CTG alone.

The false negative rate of 1.6 % (7/449) reported in the paper includes all cases in the denominator. The usual definition of false negative rate is the proportion of bad outcomes (pH < 7.05) for which the diagnostic result was negative (no significant ST event). Under this usual definition, the false negative rate is 38.9% (7/18).

V. C. Ancillary Clinical Studies

The following studies are Observational Interventional studies. They report on experience with the device in a Post-SRCT setting. Both studies were conducted as part of a program that the sponsor calls "Centers of Excellence." In this program, the sponsor is trying to foster a post-graduate training type of structure wherein knowledge gained in the use of the device will be shared with other clinicians and other facilities that use this device. The program is intended to cover many aspects of labor and postpartum neonatal care, some of which may not be associated with use of the device.

V. C. 1. Gothenburg (Centers of Excellence)

The article for this study is not yet published, but the study is discussed in section VI.A.3 under tab 2 in the sponsors portion of the Panel Package.

Upon completion of the Swedish Randomized Controlled Trial (SRCT) the evaluation of the STAN Fetal Heart Rate Monitoring System was continued in the city of Gothenburg, Sweden. This study was the first in a series of studies that were conducted under the auspices of the European Union Project and supported by the EU commission.

The objective of this prospective observational study was to evaluate the value of the STAN system for intrapartum fetal monitoring using CTG + ECG in a 24-month period of time in the total population of term deliveries in the city of Gothenburg and surrounding areas.

A total of 14687 term deliveries were included in the study, 4830 (32.9%) of which were monitored by STAN S21. Assessment of the condition of a newborn was based on cord artery and vein acid-base status (pH, PCO_2 and base deficit (BDecf), mmol/L) and Apgar scores. Metabolic acidosis was defined as cord artery pH< 7.05 and BDecf > 12.0 mmol/l. The analysis plan included the comparison of the first and second years of the study with regard to the use of STAN devices, metabolic acidosis rates, operative interventions for fetal distress, the rates of clinical compliance and rates of moderate/severe hypoxic neonatal encephalopathy.

A comparison between the first and the second years revealed an increase in STAN usage from 21.8% to 37%. A significant reduction in number of cases of metabolic acidosis from 1.12 to 0.56% was observed comparing the first and second years of

STAN use; OR=0.49; (95% CI 0.25 – 0.98). In cases monitored with STAN the reduction of metabolic acidosis decreased from 0.77% to 0.44% with an odds ratio (OR) of 0.58 (95 % CI 0.37-0.93). The rates of moderate/severe hypoxic neonatal encephalopathy were 0.49 and 0.36 per 1000 deliveries with STAN, respectively. A measure of clinical compliance was identification of the number of cases with prolonged response time measured by the time from the STAN indication to the time to intervene and reduction from 0.75% to 0.28%, OR=0.37 (95 % CI 0.15 -0.92) was observed. At the same time, total operative deliveries for fetal distress decreased from 6.6% to 6.1%.

The conclusion of this study was that the increased use the STAN system provides improvement in fetal outcome.

V. C. 2. EU Project (Centers of Excellence)

There are 2 published articles for this study and one manuscript. The study is discussed along with references to the papers in section VI.A.4 under tab 2 in the sponsor's portion of the Panel Package.

Following the Gothenburg study, the European Center for Excellence Project studies continued. The objective of this project was to develop and validate the model in which the user education and training programs played a crucial role. Ten academic centers participated with a total of 6303 cases from 2000 – 2002. Sixty-three cases had umbilical cord metabolic acidosis. Of these cases there were nine fetuses born with cord artery metabolic acidosis and Apgar score <5. The rate of newborn babies being affected ranged from 0.30% at the first six months to 0.05% toward the end of the study.

V. D. Swedish RCT (SRCT)

The article for this study is provided as tab 14 of the Sponsor's portion of the Panel Package, "Amer-Wahlin I, Hellsten C, Haken N, et al. Cardiotocography only verses cardiotocography plus ST analysis of fetal electrocardiogram for intrapartum fetal monitoring: a Swedish randomized controlled trial. Lancet 2001;358:534-38."

The Swedish Randomized Clinical Trial was the second prospective, randomized controlled clinical trial to test the STAN system. Unlike the Plymouth study, the SRCT was a multi-center study that enrolled approximately 5000 women with term >36 weeks gestations. The STAN S21 monitor was used in both arms of the study; however investigators were blinded to the ST data in the control arm of the study. Umbilical cord arterial and venous blood samples were obtained from all infants in both arms of this study in order to determine how many infants in the respective arms met predetermined criteria for metabolic acidosis, i.e. cord artery pH of <7.05 and BDecf of >12.0.

The study demonstrated a significant reduction in cases of confirmed cord artery metabolic acidosis in infants monitored with the unblinded ST data compared with those in the control arm of the study. The study also showed a significant reduction in the rate of cesarean sections for fetal distress in the STAN arm of the study, after approximately 12% of the cases were excluded from the analysis due to problems with the quality of the tracing. The rate of excluded cases was approximately equal in both arms of the study.

V. D. 1. SRCT Study Design

The study was conducted at three university-based labor wards in Sweden. The STAN system was used in both arms of the study. In the STAN arm, clinicians used T/QRS event data and ST segment analysis adjunctively with FHR tracing to manage patients according to clinical guidelines. In the control arm of the study, the only data from the scalp electrode that was available to the clinician was continuous fetal heart rate tracing. ST segment data for the control arm was unveiled for retrospective analysis of tracings with clinical outcomes data.

For all subjects, both umbilical cord arterial and venous blood was to be sampled according to the following strict guidelines: The cord was to be double-clamped as soon as possible after birth of the baby (before the placenta was delivered) and using pre-heparinized syringes or capillary tubes the blood was to be obtained from both an artery and the vein. An alternative acceptable procedure was to puncture both vessels directly at delivery.

The primary objective of the study was:

> To reduce perinatal morbidity as identified by a significant cord artery metabolic acidemia (pH<7.05 and BDecf >12.0 mmol/L). The statistical endpoint was at least a 50% reduction in the number of cases with metabolic acidemia, with a power of 80% and a test performed on the 5% level.

The secondary objectives were:

- to evaluate the use of FHR and ST waveform protocols and guidelines in clinical practice;
- to reduce operative interventions; and
- to undertake cost-benefit analysis.

Study endpoints were:

- umbilical cord acid-base status and condition of the neonate;
- change in neonatal morbidity as identified by Apgar scores at 5 minutes, assisted ventilation, admission to special care baby unit, and neonatal seizures or other neurological abnormalities;
- change in frequency of operative delivery;
- test for gender differences;
- validation of how event log file based information in the FHR-only arm retrospectively would affect clinical activity; and
- change in relative frequency of normal to abnormal scalp pH.

Inclusion Criteria

- term gestation (>36 weeks) in cephalic presentation in active labor
- intrapartum monitoring with fetal scalp electrode

Exclusion Criteria

- recording time during first stage of labor <30 minutes
- recording only represents second stage of labor

- lag time between end of recording and delivery exceeds 20 minutes
- trans-cutaneous neuro-stimulation (TENS) for analgesia
- gross fetal abnormality diagnosed prior to labor

In the control arm of the study, subjects were managed according to standard FHR-based practice. The sponsor prepared a chart describing possible fetal heart rate patterns during the first and second stages of labor, and a recommended action for the clinician to take depending on the stage of labor and the heart rate pattern (ref. Tab 4, page 10 of Sponsor's portion of the Panel Package). The definitions of fetal heart rate patterns for this chart were based on terminology proposed by the International Federation of Gynecology and Obstetrics (FIGO) (Int. J. Gynaecol. Obstet., 1987, 25:159-167).

The sponsor also developed a set of clinical guidelines to be used in the [FHR + ST] arm of the study. For these "STAN" clinical guidelines (ref. Tab 4, page 11 of Sponsor's portion of the Panel Package), the terminology and criteria for the FHR portion of the tracing was based on, but differed slightly from FIGO terminology. The most significant aspect of the STAN clinical guidelines is that they recommend clinical intervention on the basis of *both* the ECG event data and FHR pattern. (The current versions of the Clinical Guidelines and FHR (or CTG) Classification are included in the discussion of the United States Clinical Use Study in section VI, below).

V. D. 2. SRCT Results: Overview and Primary Endpoints

Effectiveness Endpoints

The SRCT took place between December 1, 1998 and June 4, 2000. Three university-based hospital labor wards contributed approximately equally to the total number of patients in this trial: University Hospital Lund (36%); University Hospital Malmo (31%); and Sahlgrenska University Hospital, Gothenburg (33%).

The study protocol called for an interim analysis of the data after treatment of 1600 subjects to assess the true incidence of metabolic acidemia in the study population and whether 3200 cases was sufficient to demonstrate a 70% reduction in the number of newborns with cord artery metabolic acidosis, assuming an incidence of 1.3%, $\beta = 0.20$ and $\alpha = 0.05$. At interim analysis, the actual incidence turned out to be 0.65%. Therefore, a second power analysis was done to calculate the number of additional cases needed. The number of additional cases was 2160. A new study deadline was set on the basis of the previous rate of recruitment.

A second result of the interim analysis was that the sponsor decided to retrain clinicians in how to use the STAN system. This came about because of protocol violations that occurred during the first 1600 cases. The study results were then presented for "before retraining" (n=2583) and "after retraining" (n=2383). Because the decision to retrain came well after the planned interim analysis of the first 1600 cases, the number of subjects treated before retraining does not correspond to 1600.

Patient accounting for the SRCT is summarized in the patient tree in Figure 1. The total number of subjects treated was 4966. The FHR + ST arm enrolled 2519, and the FHR-only arm enrolled 2447. The sponsor evaluated the two main clinical endpoints for this population, and referred to it as Analysis I. (The sponsor's "Analysis II" will be explained

below.)

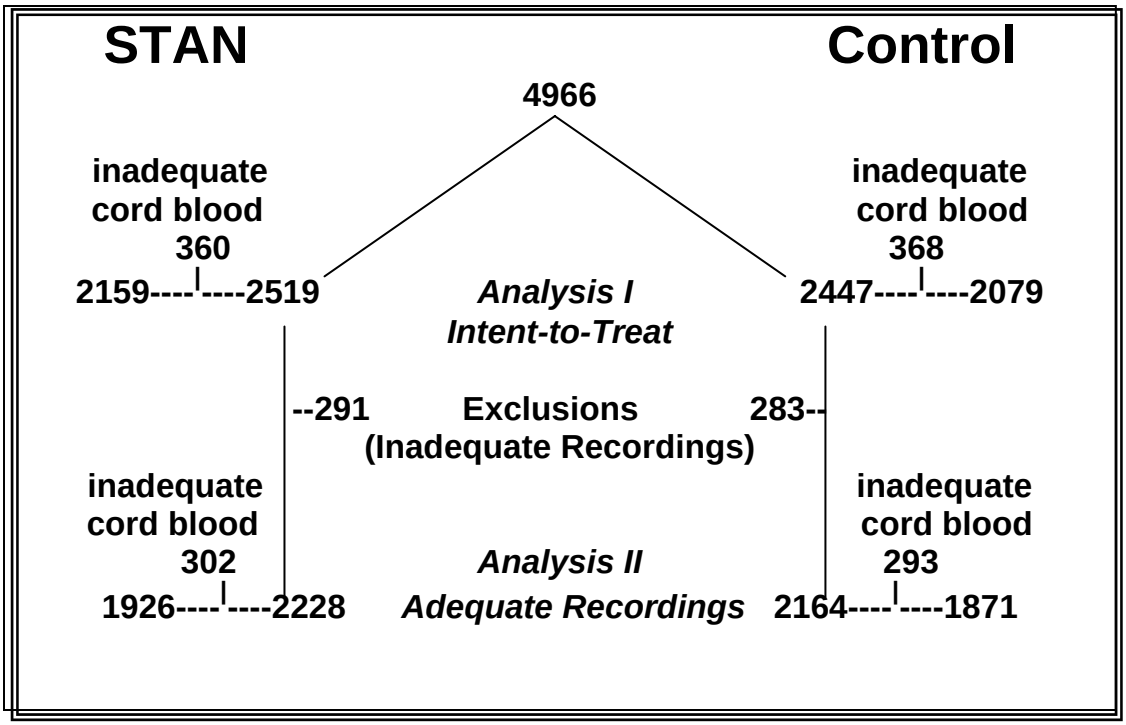


Figure 1. Swedish RCT Patient Tree

For the “intent-to-treat” population (Table 1), there were 31/2079 (1.5%) cases of metabolic acidosis in the FHR-only arm and only 15/2159 (0.7%) of cases in the [FHR + ST] arm (p=0.02). Therefore, there was a statistically significant reduction in cases of metabolic acidosis in the patients monitored with the STAN S21. The reason that the denominator numbers in these fractions are different from the actual numbers of subjects randomized in each arm is that 368 subjects in the control arm and 360 subjects in the STAN arm did not have adequate cord blood data. For example, blood was not sampled from both an umbilical artery and the umbilical vein or there was evidence from the blood gas values that cord clamping had occurred too late for the samples to be representative of true fetal hypoxic status in utero.

A secondary endpoint was operative deliveries for fetal distress (ODFD), a category that includes instrumented vaginal deliveries as well as cesarean section. For the intent-to-treat population, the rate was 227/2447 (9.3%) for the FHR-only arm and 193/2519 (7.7%) for the [FHR + ST] arm (p=0.047). This difference was statistically significant, in favor of the STAN system.

The rate for cesarean section for fetal distress (CSFD) in the intent-to-treat population was 97/2447 (4.0%) in the FHR-only arm and 87/2519 (3.5%) in the [FHR + ST] arm (p=0.38). This difference was not statistically significant.

Table 1. Results, Intent to Treat (Analysis I)

	FHR	FHR +ST
Cases of metabolic acidosis (p=0.02)	31/2079 (1.5%)	15/2159 (0.7%)
Total ODFD (p=0.047)	227/2447 (9.3%)	193/2519 (7.7%)
CSFD (p=0.38)	972447 (4.0%)	87/2519 (3.5%)

The sponsor performed an additional analysis of the clinical study endpoints (Analysis II) after excluding 291 cases from the [FHR + ST] arm and excluding 283 cases from the FHR-only arm. (Refer again to "patient tree" above.) The reason given by the sponsor to justify these exclusions was that the ECG data was inadequate due to:

- less than 20 minutes of ECG data (37.5%)
- more than 20 minutes elapsed between removal of scalp electrode (56.1%);
- congenital malformations (1.3%); and
- "other" (5.1%)

For Analysis II, the results for metabolic acidosis (Table 2) confirmed that there was a decrease in cases in the [FHR + ST] arm. Also, the "p" value for the metabolic acidosis endpoint implies greater statistical significance for Analysis II than for Analysis I.

The results for Analysis II for the endpoint ODFD also reinforced the finding from Analysis I that there was a statistically significant reduction in ODFD in subjects monitored with [FHR + ST] compared with FHR-only. Again, the statistical significance of this finding improved in Analysis II.

For CSFD, in Analysis II the trend towards reduction in the [FHR + ST] arm became statistically significant (p=0.04).

Table 2. Results, Adequate Recordings (Analysis II)

	FHR	FHR +ST
Cases of metabolic acidosis (p=0.01)	27/1871 (1.44%)	11/1926 (0.57%)
Total ODFD (p=0.009)	173/2164 (7.99%)	132/2228 (5.92%)
CSFD (p=0.04)	63/2164 (2.91%)	43/2228 (1.93%)

Safety Results

There were 29 adverse outcomes in the study:

- intrapartum death (3)
- mild/moderate/severe encephalopathy, with or without metabolic acidosis (11)
- neonates admitted to Sick Care Baby Unit (SCBU) with metabolic acidosis and other symptoms (15)

There is only partial overlap between the above cases and the total number of cases in the study in which metabolic acidosis was documented (46). The incomplete overlap is

explained by the fact that some of the infants with umbilical cord metabolic acidosis were clinically perfectly normal, vigorous neonates with high Apgar scores. These infants were not included in the count of "adverse outcomes." Alternately, some infants with normal cord blood or unavailable cord blood showed evidence of intrauterine hypoxia and/or asphyxia. These made it into the category of "adverse outcomes."

Table 3. Adverse Events

	FHR	FHR + ST
Death (intrapartum)	1	2
Severe Encephalopathy	3	0
Moderate Encephalopathy	4	0
Mild Encephalopathy	1	3
SCBU Admissions	10	5

Of the 3 intrapartum deaths in the study, 2 were in the FHR + ST arm and 1 was in the FHR only arm. One of the FHR + ST deaths occurred before retraining of clinicians, and the other occurred after retraining of clinicians. All 7 of the cases of moderate and severe encephalopathy occurred in the FHR-only arm of the study. There were no cases of moderate or severe encephalopathy in the FHR + ST arm. Of the 15 babies admitted to the SCBU with metabolic acidosis and "other symptoms" not neurologic in nature (e.g., respiratory problems), all were discharged in good condition without evidence of sequelae.

V. E. Conclusions

Observational Studies

The results of the early clinical studies varied substantially. For example, sensitivity and specificity were reported, respectively, as 100% and 95% for the Nordic study (Amer-Wahlin et al., 2002) and 43% and 74% for the Canadian study (Dervaitis et al., 2004).

The varied results were likely due largely to differing study designs. Some studies were randomized, and some were observational (non-randomized). Among the observational studies, some were interventional (patients managed using ST) and others were not (patients managed without using ST). The Nordic and Canadian studies were both non-interventional observational studies, but the attending physicians were blinded to the ST segment in the Canadian study, but not in the Nordic study. The lack of blinding in the Nordic study could have affected results, even though the attending physicians were instructed to not use the ST segment. Furthermore, different definitions were used in the two studies for metabolic acidemia (standard cord blood BDecf in Canadian study vs. BDecf in Nordic study).

A limitation with all of the studies is that unbiased estimation of sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) is difficult. These quantities are accuracy measures comparing for each delivery the baby outcome (e.g., metabolic acidemia or not) with the diagnostic result (intervention or non-intervention). The problem is that once intervention has taken place, the outcome for the baby *had the intervention not taken place* is missing. As a result, only NPV can be estimated in an unbiased fashion as the proportion of non-interventions for which a normal outcome was

obtained. In general, for all of the studies, sensitivity, specificity, PPV, and NPV were computed inappropriately by combining the outcomes for non-interventions (standard vaginal deliveries) with the outcomes for interventions (operational deliveries) at the time of intervention. The rate of intervention thus affects the estimates and probably contributed to the study-to-study variation in results.

Swedish RCT

For the "intent-to-treat" population, metabolic acidosis rate was 1.5% (31/2079) in the FHR-only arm while only 0.7% (15/2159) in the [FHR + ST] arm, a statistically significant reduction (p=0.02). The patients were monitored with the STAN S21.

Based on this study alone, one cannot conclude definitively that the same reduction in metabolic acidosis rate can be achieved in the United States. As discussed above, the reasons are that labor management practices between Sweden and the United States are different and that the training program needed to be Americanized. A main aim of the US Education and US Clinical Use Studies was to demonstrate that US attending staff can be trained satisfactorily to use STAN to manage neonates in the clinical setting. This extra information provides a bridge between Sweden and US experiences with STAN such that extrapolation of the Swedish RCT results to the US could be considered tenable.

A limitation of the Swedish randomized study is that the very presence of the ST segment in the FHR+ST arm could have affected the subjective reading of the FHR tracing by the attending physician. If the FHR tracing was over- or under-read in the FHR+ST arm relative to the FHR-only arm, then a comparison between the arms, for example, in metabolic acidosis rate, confounds ST effectiveness with the differential between arms in FHR reading. For example, over-reading of FHR in the FHR+ST arm relative to the FHR-only arm could induce a higher intervention rate in that arm and consequently a lower metabolic acidosis rate. (However, the ODFD rate was actually lower in the FHR+ST (7.7%) than in the FHR-only arm (9.3%)). The potential difference in FHR reading between arms and therefore, the potential bias, could also vary by attending physician.

VI. United States Studies

While viewed by many Panel members as being a robust study to show that the STAN system had significant impact in use in Sweden, most of the Panel members expressed concern about the transfer of the technology to the United States. As previously mentioned this was the most significant deficiency cited for the Swedish study. FDA worked with the company to develop a plan to resolve this issue, which resulted in the sponsor conducting a set of two studies in the United States: the Education Study and the United States Clinical Study.

VI. A. United States Education Study (Education Study)

The Education study and summary results are included in tab 5 of the sponsor's portion of the Panel Package.

The objective of this study was to test whether, in a class-room setting, US obstetrical personnel could be trained to understand and apply the ST analysis. If successful, the

Education Study would be used to support the sponsor's proposal for an interventional clinical study using [ST + FHR] data in laboring patients on US labor wards. The Education study was conducted in May of 2003 at three US sites; 13 clinicians participated.

VI. A. 1. Education Study Design

The study consisted of the reading of 51 tracings in random order by 13 raters (clinicians) at three exam times. The 1st exam was before training on STAN and therefore raters used FHR-only. The 2nd was after completion of training on STAN but raters used FHR-only. The 3rd exam immediately followed the 2nd exam for that case but allowed raters to use FHR+ST.

Intervention on a case was defined as being warranted if the case had cord artery pH < 7.15. Otherwise, intervention was not warranted. Among the 51 tracings, 9, 10, and 32 had cord artery pH < 7.05, 7.06-7.14, and >= 7.15, respectively.

A panel of seven Swedish clinicians, considered by the sponsor to be experts in FHR+ST analysis, had previously examined the same group of 51 tracings blinded to the pH data. Agreement of the US raters with the Swedish experts is an aspect of some secondary hypotheses of the study.

Primary Hypothesis

The primary endpoint was the average consensus of US raters. Consensus on a case was defined as "the proportion of US raters who agree to intervene on cases that (sic) intervention is warranted [cord artery pH level < 7.15] plus the proportion of US raters who agree to not intervene on cases that (sic) intervention is not warranted [cord artery pH level >= 7.15]". In other words, consensus was the percent agreement of US raters with true intervention status (as determined by cord artery pH level). The average consensus was defined as the average percent agreement with true intervention status across cases. The primary hypothesis was that the average consensus of US raters is statistically significantly greater for Exam 3 (FHR+ST) than Exam 2 (training completed, FHR-only allowed). That is, on average, US raters are more likely to make the correct decision using FHR+ST than using FHR-only.

Key Secondary Hypotheses

Three secondary hypotheses are tied to the performance of the Swedish experts who reviewed the same set of 51 cases:

1. This hypothesis was the same as the primary hypothesis except that *time of intervention* is incorporated into the definition of consensus. Consensus on a case is redefined as the proportion of US raters who make the correct decision to intervene and whose intervention time is within 20 minutes of the median intervention time given by the Swedish experts.
2. This hypothesis compares US raters with Swedish experts on the use of FHR+ST among cases for which the Swedish expert consensus was to intervene. The hypothesis was that the mean absolute difference between the Swedish experts and the US raters in the median time of intervention is less than 20 minutes.

3. This hypothesis was that the average percent agreement with true intervention status across cases (average consensus) was about the same for US raters as for EU experts when both groups used FHR+ST data to read the tracings.

The primary and secondary hypotheses were defined to be demonstrated if the difference between the two groups being studied (e.g. US investigators using FHR-only vs. FHR+ST, or US investigators vs. Swedish experts) is statistically significant.

Additional Other Analyses

Some key additional analyses were on endpoints related to, but not the same, as the primary and secondary endpoints.

1. For the primary hypothesis and key secondary hypothesis 1, consensus was specially defined as percent agreement of the raters with true intervention status. More traditionally, consensus on a case is the majority decision by the raters. In an important additional analysis, the majority was used to determine the decision to intervene or not for each case. The intervention rate of the majority was computed for US raters and for EU experts and stratified by exam (1, 2, or 3) and cord artery pH level group (< 7.05, 7.06-7.14, > 7.14). From these majority vote data, the proportion of correct interventions (sensitivity) and the proportion of correct non-interventions (specificity) were obtained, where, as before, correctness to intervene was defined by pH level < 7.15. Agreement of the majority of US raters with the majority of EU experts in both decision and timing (within 20 minutes) was also considered.
2. Intervention rates were stratified by rater (US rater or EU expert) and pH level.

Concept

Although the guidelines may be slightly different because Europeans follow FIGO conventions for terminology to describe intrapartum fetal status, participants received the same training and took the same certification test as do labor and delivery personnel in countries where the STAN system is approved. This is also the same training and certification test completed by investigators in the SRCT. The training and certification test consist of the following:

- Self study (textbook and interactive CD):
    - Material covered in the textbook and CD covers basis physiology related to fetal hypoxia and fetal surveillance during labor including electronic fetal monitoring (EFM) interpretation, FECG physiology, and FECG interpretation.
    - The interactive CD includes FHR+ST interpretation exercises and a quiz.
- Interactive on-site tutorial (including case discussions; cases are not part of the set of 51 cases evaluated during the Education Study) scheduled following self study;
- Written certification test consisting of 18 multiple choice questions.

This was a “virtual” clinical study which used a library of 51 tracings from European patients whose labors were monitored with the STAN S21. This library of cases was selected on the basis of the cord artery pH values in order to challenge clinicians with a range of pathology. All of these cases had previously been evaluated in a similar

fashion by Swedish experts in use of the STAN S21. The distribution of cases stratified by pH follows:

- pH <7.05 (9 cases)
- pH 7.05-7.14 (10 cases)
- pH ≥7.15 (32 cases)

The US investigators examined each tracing three different times. First before STAN training and without ST segment analysis (i.e. only FHR data were available); next following STAN training and again without ST segment analysis; last, after STAN training but this time both the FHR data as well as ST segment analysis were available. At each examination, the US investigator was to indicate whether or not an intervention was needed, and if so, the time at which the tracing indicated an intervention.

Effectiveness Endpoints

The primary endpoint of this study was the average consensus of US investigators who (1) agree to intervene on cases that warrant intervention based on pH; plus (2) agree not to intervene on cases that do not warrant intervention. Study "success" was defined as improvement in consensus after the addition of ST segment analysis to the FHR data. Success for this study was therefore determined using the US investigators as their own controls. Additional analyses were performed comparing performance of US investigators against Swedish experts in the use of the STAN S21 system.

VI. A. 2. Education Study Results

Three of the 13 raters were not certified (did not pass certification exam) but still rated the tracings during exams 2 and 3. Analyses were made excluding as well as including these three raters.

Primary Hypothesis

For the US investigators average % agreement with true intervention status (average consensus) was 47% for Exam 1 (FHR-only, before training), 53% for Exam 2 (FHR-only, after training), and 69% for Exam 3 (FHR+ST, after training) (Table 4). From a repeated measures analysis, the difference in average % agreement between Exams 3 and 2 was statistically significant ($p = 0.0001$). The small p value and positive difference indicates that the primary hypothesis that the average % agreement would increase from Exam 2 to Exam 3 was met.

The repeated measures analysis included random effects for US raters. However, the analysis did not allow for interaction between US rater and exam time, that is, for differences in % agreement over exam times to depend on US rater. Inclusion of these interactions would have introduced more variability into the analysis. This extra variability would have increased the p value in the comparison between Exams 3 and 2.

When the three US raters who were not certified were excluded, the remaining 10 certified US raters had average % agreement with true intervention status (average consensus) of 55% for Exam 2 (FHR-only, after training), and 69% for Exam 3 (FHR+ST, after training). These average % agreements were very similar to those given above that include the three uncertified US raters.

As a benchmark for comparison, the Swedish experts as a group improved from 59% average % agreement when reading with FHR-only to 85% when reading with FHR+ST (Table 4).

TABLE 4. % Agreement with True Intervention Status, as Defined by Cord Artery pH level (< 7.15 threshold), N=51 cases.

	FHR-only, before training		**FHR-only, after training**		**FHR+ST, after training**	
	Average	**Range**	**Average**	**Range**	**Average**	**Range**
US Raters (13)	47	37-64	53	41-67	69	43-88
EU Experts (7)	NA	NA	59	51-63	85	75-90

Secondary Hypotheses

The first secondary hypothesis is the % agreement (consensus) not only on the decision to intervene but on the decision to intervene being within ± 20 minutes of the median time of intervention given by the EU experts. Because of the need to meet both conditions, we would expect the % agreement figures to be lower than those observed in the discussion of the primary hypothesis. This was the case. Among US investigators who correctly decided to intervene in cases where pH<7.15, % agreement within ±20 minutes of the median intervention time of the EU experts improved from 43% when FHR-only data were available to 59% when FHR+ST data were available. From a paired t-test, the difference was statistically significant ($p = 0.005$), indicating that the secondary hypothesis of a statistically significant increase was met.

Two additional secondary hypotheses compared performance of US investigators against Swedish experts when both FHR +ST data were available. Regarding the decision to intervene only (and not including timing), the Swedish experts as a group had an average % agreement (average consensus) of 85% when reading with FHR+ST, compared with 69% for the US investigators (Table 1). The statistical significance of this difference was not assessed. The hypothesis was that these two figures would be similar.

When the timing of the US investigators decision to intervene was compared to that of the Swedish experts, the hypothesis was that the mean absolute difference between the Swedish experts and the US raters in the median time of intervention is less than 20 minutes. The sponsor did not provide an assessment of this hypothesis. Instead of the mean absolute difference, the sponsor analyzed the proportion of cases in which in the median times of intervention for the two groups differed by less than 20 minutes. The proportion of cases was 89% (8/9) for cases with pH level < 7.05 (the Swedish experts did not intervene on one of these cases), 60% (6/10) for cases with pH 7.05-7.14, for a total of 74% (14/19) for cases with pH <7.15.

VI. A. 3. Additional Analyses

In a key additional analysis, the US raters and the EU experts were compared on their majority decision to intervene or not in each case (Table 2). Time of intervention was not considered in this analysis. For the US rater majority, the % agreement with intervention status (labeled total accuracy) increased over the exam times, from 40%

using FHR-only before training to 56% after training, and to 77% using FHR+ST after training. The sensitivity (% agreement with the correct decision to intervene) decreased only slightly from 95%, to 89%, to 84% over the three exams. The results were similar for the EU expert majority when comparing the use of FHR-only data with FHR+ST data, although the EU expert majority had lower sensitivity and higher specificity (% agreement with the correct decision to not intervene) than the US rater majority. The statistical significance of these results was not assessed.

TABLE 5. Majority Decision Intervention Rates (%) by Cord Artery pH.

	Intervention Rate (n/N) by Cord Artery pH level			Accuracy (% Agreement with Intervention Status)		
	< 7.05	7.05-7.14	> 7.14	Sensitivity (pH<=7.14)	Specificity (pH>7.14)	Total
US Raters						
FHR-only, Exam 1	100% (9/9)	90% (9/10)	91% (29/32)	95% (18/19)	9% (3/32)	40% (21/52)
FHR-only, Exam 2	89% (8/9)	90% (9/10)	63% (20/32)	89% (17/19)	38% (12/32)	56% (29/52)
FHR+ST, Exam 3	100% (9/9)	70% (7/10)	25% (8/32)	84% (16/19)	75% (24/32)	77% (40/52)
EU Experts						
FHR-only	78% (7/9)	70% (7/10)	50% (16/32)	74% (14/19)	50% (16/32)	58% (30/52)
FHR+ST	89% (8/9)	60% (6/10)	6% (2/32)	74% (14/19)	94% (30/32)	85% (44/52)

Another additional analysis stratified intervention rate by rater (US rater or EU expert) and pH level (Table 5). For the US raters, the average intervention rate decreased over increasing pH level, from 90% to 75% to 39% for pH level groups < 7.05, 7.05-7.14, and > 7.14. However, the spread of the rater-specific intervention rates increased over the pH level groups, from 67-100% to 40-100% to 3-88%. Excluding the three uncertified raters, the spread was still 6-84% for pH level > 7.15. For the EU experts, the spread of the intervention rates for pH level > 7.15 was much smaller, 6-19%. For the US raters, the large spread of the intervention rates for pH level > 7.15 is evidently due to the effectiveness of the training varying by reader. Indeed, some certified raters substantially increased their % agreement with the true intervention when going from Exam 2 to Exam 3 while others did not. For example, rater 1 increased from 51% to 88%, while rater 10 decreased from 53% to 45%.

TABLE 6. Intervention Rate (%) Using FHR+ST data, by Cord Artery pH level (N=51 cases).

	Cord Artery pH level					
	< 7.05		**7.05-7.14**		**> 7.15**	
	Average	**Range**	**Average**	**Range**	**Average**	**Range**
US Raters (13)	90	67-100	75	40-100	39	3-88
EU Experts (7)	90	78-100	64	50-80	9	6-19

In a repeated measures analysis of US rater intervention rate, a interaction between exam time (Exams 1, 2, 3) and pH level group (< 7.05, 7.05-7.14, > 7.14) was significant (p=0.0001), indicating the difference in intervention rate between exam times depended on pH group. Specifically, when going from Exam 2 (FHR-only read after training) to Exam 3 (FHR+ST read after training), the intervention rate increased only slightly and insignificantly for cases with pH level < 7.05 while the intervention rate dropped sharply and significantly for cases with pH level > 7.14. These findings are consistent with the results in Table 6 for the US rater majority.

The repeated measures analysis included random effects for US raters. However, the analysis did not allow for interaction between US rater and exam time, that is, for differences in intervention rate over exam times to depend on US rater. Inclusion of these interactions would have introduced more variability into the analysis.

Summary

Overall, using pH as the benchmark for the need to intervene, the US investigators showed improvement in their ability to discern correctly that intervention was not warranted after the ST data were added to the FHR data. The study was a success in showing the primary hypothesis that the average % agreement with the decision to intervene, as determined by cord artery pH level, improved significantly from using FHR-only data to FHR+ST data (p = 0.0001). These results demonstrate that following STAN training and availability of FHR+ST data, the US investigators improved both in the correctness and timing of their decisions to intervene, and in the decision not to intervene when intervention was unwarranted. These findings validated the adequacy of the training program and supported the sponsor's request to proceed with a prospective, interventional study of the STAN in the US; the United States Clinical Use Study.

Compared to the Swedish experts, however, the US investigators still tended to intervene unnecessarily at a higher rate although the US investigators vastly improved their performance after ST data were available. Variability over investigators in the success of training was evident. The rater-specific intervention rates for cases in which intervention was not warranted (pH level > 7.14) ranged from 3-88% for all 13 US raters and from 6-84% for the ten certified raters, very large spreads that were not seen for the EU experts (6-19%).

VI. B. United States Clinical Use Study (CUS)

The United States Clinical Use study and summary results are included in tab 6 of the sponsor's portion of the Panel Package.

The US Clinical Use Study (CUS) was a prospective, non-randomized, un-controlled multi-center trial. Thirty-nine clinicians participated at 6 sites, 3 of which were community hospitals and 3 of which were teaching hospitals.

The purpose of the CUS was to demonstrate that US obstetrical staff can appropriately use the STAN system, correctly interpret FHR+ST data and apply the STAN clinical use guidelines on a par with STAN experts. Success for the CUS would validate the training program and allow the results of the SRCT to be extrapolated to the US.

VI. B. 1. CUS Study Design

As previously noted, the clinical guidelines (Table 7 and 8) used in the SRCT were modified for use in the United States:

Table 7 Classification of FHR patterns

FHR-classification	Baseline heart frequency	Variability Reactivity	Decelerations
Reassuring	• 110 - 150 bpm	• 6 – 25 bpm • Accelerations present	• Early decelerations • Variable decelerations. with a duration of <60 sec and depth <60 beats
Non Reassuring, grade 1	Bradycardia: • Rate <110 bpm (without accelerations) • Episode >2 min duration regardless of reactivity or variability Tachycardia: • Rate >150-170 bpm and absent/minimal variability • Rate >170 bpm	• ≤5 bpm for >40 min • ≥25 bpm for >40 min • Accelerations absent	• Variable decelerations with a duration >60 sec and depth >60 beats • Repetitive late decelerations
Non Reassuring, grade2 (Preterminal)	▪ Absent variability and reactivity with or without decelerations or bradycardia ▪ Sinusoidal pattern		

Table 8 STAN simplified clinical guidelines
These guidelines are applicable to a term pregnancy of 36 completed gestational weeks or more. They indicate situations in which intervention is **required**. This means delivery or alleviation of a cause of fetal distress such as over-stimulation or maternal hypotension. During the second stage of labor with active pushing, **immediate** delivery is recommended.

	Reassuring FHR	Non-reassuring FHR, grade 1	Non-reassuring FHR, grade 2 – Preterminal FHR
No ST change	"Routine Management" Continued Observation	Expectant Management During 2nd stage you may wait no longer than 90'	Intervention required regardless of any ST changes
Episodic T/QRS rise (duration <10 min)	"Routine Management" Continued Observation	If > 0.10* delivery is required within 30 minutes in 1st stage and immediately if recorded during active pushing.	Delivery to be expedited
Baseline T/QRS rise (duration >10 min)	"Routine Management" Continued Observation	If > 0.05* delivery is required within 30 minutes in 1st stage and immediately if recorded during active pushing.	Delivery to be expedited
Biphasic ST (Three degrees of Biphasic ST changes are possible, BP1, BP2 and BP3, only BPs grade 2 and 3 are regarded as siginificant)	Closer Observation	If continuous > 2 min or > 1 episode of BP2 or BP3 delivery is required within 30 minutes in 1st stage and immediately if recorded during active pushing.	Delivery to be expedited

*compared to baseline T/QRS.
Recommendations for Intervention using FHR Patterns and ST waveform changes
Non-reassuring, grade 2 (pre-terminal) would prompt an expedited delivery without the need to consider ST data.
Non-reassuring grade 1 with episodic T/QRS rise, baseline rise, or persistent biphasic ST pattern would mandate delivery within 30 minutes or less.
At start-up and when there is a decrease in signal quality with discontinuous T/QRS ratios, manual data analysis is required.

Primary Effectiveness Endpoints

- Probability that non-intervention results in a normal outcome in the cohort of infants with non-reassuring FHR (NRFHR) tracings. (A normal outcome is defined as cord artery pH>7.12 and a NRFHR pattern as defined by the Clinical Guidelines.) This quantity is referred to as the Negative Predictive Value (NPV). It is defined in the protocol as follows:

$$\frac{\text{\# cases w/NRFHR, STAN allows continued labor, no intervention taken, pH>7.12}}{\text{\# cases w/NRFHR, STAN allows continued labor}}$$

- Agreement of US clinicians with the majority decision of 3 STAN experts on the cause and time of a decision to intervene based on the information available from the STAN recording. Agreement on timing of the decision to intervene must be within ±20 minutes of the majority of STAN experts in the second stage of labor, and within ±30 minutes within the first stage of labor.

For the second co-primary endpoint, agreement was measured in terms of positive percent agreement (PPA) and negative percent agreement (NPA) with the STAN experts. PPA is the proportion of cases warranting intervention according to the STAN experts for which the US clinician decided to intervene at about the same time, where timing is defined as above. NPA is the proportion of cases not warranting intervention according to the STAN experts for which the US clinician decided not to intervene. The protocol pre-specified that since the US clinicians have clinical information available to them that is not available to the STAN experts, who only have the tracing, some failures to meet the above definition of agreement may be "overruled." Disagreements will each be reviewed, and if the disagreement is due to lack of relevant information by the STAN experts, those cases may be considered agreement.

Secondary Effectiveness Endpoints

- Case-based analysis of all cases with cord artery pH <7.14;
- Comparisons of operative intervention rates of clinical investigators before and during the study;
- Analysis of all cases with NRFHR including periods of normal, Grade 1 and Grade 2 non-reassuring. This analysis will provide the distribution of cases with FHR changes, their duration, and whether intervention was indicated according to STAN guidelines;
- Analysis of pH ranging from 7.10 to 7.15 and corresponding NPV in cases with non-reassuring FHR not intervened on;
- Comparison of US clinicians vs. STAN experts with regard to "true" intervention or no intervention and pH;
- Analysis of primary endpoints stratified by clinical investigator to assess what proportion of experts would meet the primary endpoints; and
- Analysis of the second primary endpoint, excluding those cases that were accepted to have met with the standards of agreement according to the Cause of Disagreement protocol.

VI. B. 2. CUS Results

A total of 530 subjects were monitored.

Primary Effectiveness Endpoints

The hypothesis for the first primary endpoint was that the NPV as defined above would be >75%, with the pH threshold set at >7.12 for cases that do not require intervention. For the protocol threshold of > 7.12 for a normal outcome, NPV was 95.2% (180/189) with 95% CI (91.2, 97.8%), based on 189 cases with NRFHR tracing. Because this pH

threshold is somewhat arbitrary (and was different from the pH threshold for normal pH of ≥7.15 in the US Education Study), FDA requested that the sponsor stratify the results for this endpoint over a range of pH values from 7.10 to 7.15. For each of these cut off points, the NPV was ≥ 92.6% and the lowest bound on the 95% CI was 87.9%. Therefore, the sponsor easily met the target for success for this primary endpoint. This endpoint relates only to patients who experienced a period of non-reassuring fetal heart rate.

Co-primary endpoints were the positive percent agreement (PPA) and negative percent agreement (NPA) between the US investigators and the majority consensus of three STAN experts. PPA reflects the level of agreement in cases where **intervention is necessary according to STAN experts**, and NPA reflects agreement in cases where **intervention is not indicated according to STAN experts**. Unlike the first primary endpoint (NPV), cord artery pH was not part of the basis for determining PPA and NPA. In addition, all 530 cases were considered, not just those with an NRFHR tracing. The gold standard that established the need for intervention was conformance with the STAN guidelines, as interpreted by the STAN experts. For PPA, agreement with STAN experts to intervene was based not just on the decision to intervene or not, but also on the timing of intervention being within 30 minutes during 1st stage of labor and 20 minutes during 2nd stage of labor.

Two cases were excluded from PPA/NPA analysis. In one case, the STAN experts could not agree. In another case, the US clinician expressed both FTP and FHR+ST as cause for intervention and at the same time. Thus, 528 of the 530 cases were analyzed.

Table 9 Data Table for Calculation of NPA and PPA

N=528	**STAN Expert Majority Decision**			
US Clinician Decision	**No intervention**	**Intervention 2° to FHR concern outside STAN guidelines**	**Intervention 2° to FHR concern within STAN guidelines**	**Intervention 2° to FHR+ST concern within STAN guidelines**
No intervention or intervention for failure to progress	444 a	b	c	5 d
Intervention 2° to FHR concern outside of STAN guidelines	31 e	f	g	h
Intervention 2° to FHR concern within STAN guidelines	6 i	j	k	l
Intervention 2° to FHR+ST concern within STAN guidelines	10 m	n	1 o	31 p

*or intervention for failure to progress

NPA = a/(a + e + i + m) = 444/(444 + 31 + 6 + 10) = 444/491 = 90.4%

PPA = (k + p)/(b + c + d + f + g + h + j + k + l + n + o + p) = 31/37 = 83.8%

The STAN experts saw a need to intervene in 37 tracings. The US investigators saw a need to intervene in 79 tracings. These figures are not directly comparable because, as stated earlier, the US clinicians were responding to clinical events in labor and delivery to which the STAN experts were not privy. These included 47 cases in which the STAN experts would not have intervened. Thirty-one of these cases were ones in which events known to the US clinician and **not** within the purview of the STAN guidelines were the basis for the intervention. Therefore, there were 16 cases in which US clinicians correctly intervened according to STAN guidelines and STAN experts did not, and 5 cases in which STAN experts correctly intervened according to the guidelines and US investigators did not.

The protocol states that the "target" for PPA will be 75% and the "target" for NPA will be 75% (p. 16 of 18). The study protocol provides that 500 cases with a prevalence rate of 9% (for cases that truly require intervention) would have **80% power to show PPA>75%,** assuming the true PPA is 90%. The protocol also provides that for negative

percent agreement (NPA), 500 total cases would be sufficient to show that **with 80% power NPA >75%,** assuming the true NPA is 80%. In its conditional approval letter of January 30, 2004, FDA gave permission to enroll 650 subjects in order to have 500 treated subjects.

The PPA was 83.8% (31/37) with 95% CI (68.0%, 94.7%). The 95% CI lower bound indicates that the target value of > 75% was not met. The sponsor has pointed out (p.32 of 54 of Volume VII of the PMA), that the study was sized for showing PPA > 75% on the assumption that the true PPA is 90%. The 90% PPA was based on the rate of agreement to intervene in the Education Study. This rate did not take into account timing of intervention. When the Education Study data are re-calculated for **both** decision **and** timing of intervention, the new rate is **only** 74% (14/19). Another thing to consider in this context is that for the Education Study, study both the US and STAN experts had the same information available to them. In the CUS, the STAN experts were at a disadvantage compared to the US clinicians. Therefore, the 90% target rate for PPA based on the Education Study may have been set too high.

In the calculation of PPA, some technical disagreements were considered to be agreements because, as discussed above, US clinicians had clinical information not available to the STAN experts. Specifically, 17 standard vaginal deliveries (SVD) were considered to be in positive agreement with the STAN expert decision to intervene on these same cases because the SVD occurred within 20 minutes of the time of the decision to intervene. Agreement on such cases was pre-specified in the protocol, but not whether it would be considered a positive agreement (for inclusion in PPA) or a negative agreement (for inclusion in NPA).

The NPA was 90.4% (444/491) with 95% CI (87.8%, 97.0%). The 95% CI lower bound indicates that the target value of > 75% was easily met. In the calculation of NPA, some technical disagreements were considered to be agreements because US clinicians had clinical information not available to the STAN experts. Specifically, 82 interventions for failure to progress (FTP) were considered to be in negative agreement with the STAN expert decision to not intervene on these same cases. The first row of the exam matrix above is labeled to include this situation.

Secondary Effectiveness Endpoints

- Case-based analysis of all cases with cord artery pH <7.14;

  Five percent (27/528) of cases had cord artery pH<7.15. There was one case of metabolic acidosis (pH 6.88, BDecf 14.4 mmol/L) but this infant had reasonable Apgar scores and was not one of the 8 infants admitted to the NICU.

  One case had a poor outcome during the neonatal period. Of 4 cases with pH of ≤7.05 (one of the criteria for metabolic acidosis in the Swedish RCT), one infant (pH 7.05, BDecf 10.3mmol/L, Apgars 6 and 8) was admitted to the NICU for observation due to grunting, maternal fever and possible sepsis. At 28 hours of age, the infant displayed seizure activity lasting for 1 minute. The initial EEG was abnormal. MRI two days later showed “acute infarction involving left anterior brain stem and basal ganglia + portions of left frontal and temporal lobe.” The sponsor attributed this event to maternal heparin therapy 2° history of deep vein thrombosis.

There was an important protocol deviation in that this subject had a private physician (not an investigator) who attended the labor. The study investigator consented the patient and placed the STAN system, but then left the room. Labor and delivery proceeded under the management of the private physician.

VI. B. 3. Additional Analyses

The additional analyses and charts mentioned in the following discussion are contained in Tab 25 of the sponsor's portion of the Panel Package.

Alternative NPV calculation

For the per protocol threshold > 7.12 for a normal outcome, NPV was originally reported as 95.2% (180/189). This calculation was based on the per protocol definition of NPV, which was (from Section VI.B.2)

$$\frac{\text{\# cases w/NRFHR, STAN allows continued labor, no intervention taken, pH>7.12}}{\text{\# cases w/NRFHR, STAN allows continued labor}}$$

The NPV denominator (189) are cases in which STAN allows continued labor. These cases include not only non-interventions but interventions outside of STAN guidelines. A problem with including interventions outside of STAN guidelines is that for an intervention, the outcome for the baby *had the intervention not taken place* is technically missing. Only the outcome for the baby at the time of intervention is available.

Likewise, the NPV numerator (180) includes not only non-interventions but interventions outside of STAN guidelines. From the per protocol definition given above, the interventional cases should not have been included.

Per FDA's request, the primary endpoint of NPV was recalculated. In keeping with its statistical definition of NPV, NPV was redefined as

$$\frac{\text{\# cases w/NRFHR, STAN allows continued labor, no intervention taken, pH>7.12}}{\text{\# cases w/NRFHR, STAN allows continued labor, no intervention taken}}$$

Using this definition, the NPV was 95.5% (127/133) with 95% CI (90.4%, 98.3%). From the lower bound on the 95% CI, the primary hypothesis of NPV > 75% was still met.

Alternative PPA calculations

Per protocol, a spontaneous vaginal delivery (SVD) was considered to be in agreement with the STAN expert decision to intervene if the SVD occurred within 20 minutes of the time of the decision to intervene. The rationale is that the attending staff member did not indicate intervention because delivery was imminent. These cases were included as positive agreements in the calculation of PPA. A problem with considering these cases as positive agreements is that it presumes that the attending staff member would have decided to intervene *had delivery not been imminent*. We do not know this. Because the decision to intervene or not is missing for the attending staff member in these cases, FDA requested recalculation of PPA in two ways: (1) with these cases excluded to reflect the missing data, and (2) with these cases considered disagreements to reflect the worst case scenario.

When these cases are excluded, PPA was 70.0% (14/20) with 95% CI (45.7%, 88.1%). The 95% CI lower bounds indicates that the hypothesis PPA > 75% target value for PPA was still not met. When these cases are considered to be disagreements, PPA was 37.8% (14/37) with 95% CI (22.5%, 55.2%), indicating that the hypothesis PPA > 75% target value for PPA was still not met.

As noted above, the > 75% target value for PPA was derived from the results of the Education Study using just decision and not including timing. When PPA is calculated without timing and with the 17 SVD agreement cases excluded PPA becomes 75.0% (15/20) with 95% CI (50.9%, 91.3%). When the SVDs are considered disagreements, PPA without the time factor is 40.5% (15/37) with 95% CI (24.8%, 57.9%). Thus, without timing, the target value for PPA was still not met.

It should also be noted that evidence for agreement of the SVD cases may be derived from review of the case report forms. While these forms may not inform on all cases and do not necessarily indicate if the US clinician would have intervened, they do provide some insight.

Alternative NPA calculation

In the calculation of NPA, 82 operative deliveries for failure to progress (FTP) were considered to be in negative agreement with the STAN expert decision to not intervene on the same case. These cases were considered to be agreements anyway because the attending staff member did not intervene either within or outside STAN guidelines except for FTP. If these cases are excluded, the NPA becomes 88.5% (362/409) with 95% CI (85.0%, 91.4%). The 95% CI lower bound indicates that the hypothesis that NPA > 75% is still easily met. If these cases are considered disagreements to reflect the worst case scenario, the NPA becomes 73.7% (362/491) with 95% CI (69.6%, 77.6%). From these results, the hypothesis that NPA > 75% is no longer met.

Alternative NPA and PPA calculations: swapping technical disagreements

In the calculation of PPA, 17 non-interventions on SVDs were considered to be in positive agreement with the STAN expert decision to intervene on the same case because SVD occurred within 20 minutes of the time of the decision to intervene. Per protocol these were considered agreements even though technically they are disagreements (the US clinician did not intervene). Going beyond the exclusion of these cases as discussed above, consider that a reasonable assumption might be that the STAN experts would not have intervened on these cases had they known that an SVD was imminent. However, from that perspective these cases should then be considered as agreements to not intervene and therefore be included in the calculation of NPA, not PPA. Note, however, that timing of intervention is not considered in this calculation.

In the calculation of NPA, the 82 interventions for FTP were considered to be in negative agreement with the STAN expert decision to non-intervene on the same case. As previously indicated these are, technically, disagreements. Going beyond the exclusion of these cases as discussed above, consider that a reasonable assumption might be that the STAN experts would have intervened for FTP in these cases had they been clinically managing these cases. However, from that perspective these cases should

then be considered as agreements to intervene (outside of STAN guidelines) and therefore be included in the calculation of PPA, not NPA.

When the 82 FTP cases included as agreements for NPA and the 17 SVD cases included as agreements for PPA are swapped, the hypotheses of NPA > 75% and PPA > 75% are both met. PPA becomes 94.1% (96/102) with 95% CI (87.6%, 97.8%), which has 95% lower bound > 75%. NPA becomes 88.9% (379/426) with 95% CI (85.6%, 91.8%), which has 95% lower bound > 75%. Note that time of intervention is not considered for the 82 FTP cases folded into PPA.

VI. D. 3. Conclusions

Among these primary hypotheses that NPV > 75%, PPA > 75%, and NPA > 75%, only PPA > 75% was not met, according to the sponsor's analyses in the PMA.

FDA asked for alternative analyses of NPV and PPA. Neither of these alternative analyses changed conclusions. The hypothesis NPV > 75% was still met and the hypothesis PPA > 75% was still not met. The initial NPV analysis included interventions not under STAN guidelines. The alternative NPV analysis restricted to non-interventions, in keeping with the statistical definition of NPV. The initial PPA analysis included technical disagreements between US clinicians and STAN experts on the decision to intervene that were per protocol considered to be agreements. FDA asked that these cases be excluded or considered to disagreements.

FDA made additional analyses of PPA and NPA. Some of the more conservative additional analyses FDA performed on the PPA calculation yield worse results than even the sponsor's original analysis. FDA's PPA analysis that, based on the above clinical perspective, exchanges the 17 SVD technical disagreements that were included as positive agreements for PPA and the 82 technical disagreements that were considered as negative agreements for NPA has a more positive outcome. Specifically when these cases are swapped in calculation PPA and NPA, both the hypotheses PPA > 75% and NPA > 75% are met. The study would then succeed in that all three primary hypotheses NPV > 75%, PPA > 75%, and NPA > 75% are now met.

The PPA result submitted in the PMA and the additional analyses demonstrate that there are multiple ways to evaluate this study endpoint that lead to different conclusions. This and the fact that the CUS succeeded in meeting the NPA and NPV endpoints should be taken into account in deciding whether or not the CUS was successful in demonstrating that with adequate training STAN can be adopted into clinical practice in the US.

The study does not permit calculation of the diagnostic accuracy endpoints of sensitivity, specificity, and PPV. These endpoints cannot be computed because the true status of a case cannot be known when it is intervened on. That is, the cord artery pH level that would have been obtained had the case not been intervened on is not known. Therefore, the study only permits calculation of NPV, the proportion of non-interventional cases with normal outcomes. The study protocol did not assign relative weight to the three different primary endpoints. The clinical importance of each of the three primary endpoints (i.e. PPA, NPA and NPV) is still under consideration by FDA.