

Electronic Fetal Monitoring & Computer-Assisted Diagnosis

An FDA Overview

**This document is background material for
the workshop only and should not be
interpreted as an FDA guidance document.**

**Electronic Fetal Monitoring and Computer
Assisted Diagnosis: An Overview**

Table of Contents

I. What do we mean by CAD?	3
II. CAD device types defined by FDA	4
III. How CAD devices generally work	4
IV. Use of CAD devices in clinical practice	7
V. Hierarchical model of efficacy	7
VI. Non-Clinical Testing.....	8
VII. Clinical Testing of Type 3 CAD devices.....	9
A. Outcomes and Endpoints	9
B. Types of Clinical Studies	10
C. Statistical Considerations.....	12
VIII. Testing of Type 2 CAD devices	15
A. Outcomes and Endpoints	15
B. Clinical Testing	16
C. Nonclinical Testing.....	17
D. Statistical considerations for Type 2 Devices.....	17
E. Types of tests and potential biases.....	18
F. Multiple Reader Multiple Case (MRMC) Reader Study Design.....	18
IX. References.....	19

I. What do we mean by CAD?

Over the last 10-15 years, computerized pattern recognition algorithms have been used in conjunction with conventional monitoring and diagnostic modalities for clinical decision support. These so-called CAD systems, i.e., computer-aided detection and/or computer-aided diagnosis are intended to help clinicians find (i.e., detect) signs and symptoms of disease conditions as well as help the clinicians make the actual diagnosis. The term "computer-assisted" is used interchangeably with "computer-aided." CAD systems can aid with one or more parts of the medical evaluation process for the purpose of reducing errors and therefore have the potential to improve care. Today, CAD systems are used for radiographic image interpretation, intensive care monitoring, and intrapartum fetal monitoring, as well as many other clinical applications. An overview of CAD for medical imaging application can be found in the Briefing Package for the March 2008 FDA Radiological Devices Panel (1).

Devices that primarily influence the detection task are sometimes referred to as CADE (computer aided detection) devices, while those that primarily influence the diagnosis task are sometimes referred to as CADx (computer aided diagnostic) devices. Overall, CAD systems may influence one or more of detection, description, diagnosis or reporting tasks, depending on the device design, testing, output, mode of interaction with the clinician and intended use.

The prototypical CADx device is designed to *characterize* a specific finding that has already been detected and determined to be a potential abnormality (such as the evaluation of a questionable finding on a lung CT scan). In obstetrics a CADx device would characterize the finding (e.g., the likelihood a tracing is abnormal) and/or describe the finding (e.g., characterize various fetal heart rate or uterine contraction patterns). In sum, the prototypical CADx device interprets for the clinician what it is the tracing signifies.

The distinction between CADx and CADE devices in obstetrics is hazy, however. A prototypical CADE device is designed to process information to detect findings that may be an abnormality. The term "abnormality" is here meant to include any finding that is not a normal structure or normal variant. CADE systems performing solely a detection task are commonly applied to the detection of abnormalities on medical images (e.g., mammograms) or finding well-recognized FHR patterns (e.g. acceleration, late deceleration, no variability etc.) during intrapartum monitoring. CADE devices function similarly to CADx devices, however, they generate a likelihood rating, (generally not available to the user), of disease condition for every region of interest (ROI) or temporal epoch of monitoring.

CADE for EFM applications may provide feedback to the clinician by triggering an alarm or warning when a specific condition is identified. In general a CADE system displays overlay marks, highlights, outlines, or alarms of some type to identify the presence of a finding within a region (imaging) or time epoch (monitoring) and brings it to the clinician's attention to reduce perceptual errors. In sum, the prototypical CADE device tells the clinician where (imaging) or when (monitoring) to look at a device's output for more careful evaluation.

II. CAD device types defined by FDA

For the sake of discussion, FDA has defined three types of CAD features that could be used with electronic fetal monitoring:

- Type 1 detects min/max thresholds for simple EFM monitoring (e.g., FHR > 150 bpm; FHR < 120 bpm; FHR variability < 5 bpm, etc.)
- Type 2 detects one or more specific EFM patterns that are generally recognized of clinical interest (e.g., late deceleration, hyperstimulation, sinusoidal FHR pattern, etc.)
- Type 3 risk stratification for future event of clinical concern (e.g., biphasic event/urgent delivery)

All of the three different types include a CADE component with Types 2 and 3 likely including some type of CADx component as well. The potential impact of the devices on the clinician's decision process is also likely to increase going from a Type 1 device (less impact) to a Type 3 device which would be expected to have a large potential impact in the decision process.

III. How CAD devices generally work

CAD systems typically use software algorithms to find patterns in patient data that indicate a potential abnormality. Just as a clinician is trained to identify abnormalities using case studies, CAD algorithms are first trained to identify patterns using data from a finite sample of patients with and without abnormalities. This sample is called the "training set". Once trained or optimized, the CAD system can then be used on new patients to detect similar abnormal patterns and discriminate them from normal patterns. The CAD device's classification of patterns is intended to be sufficiently sensitive and specific to aid clinicians in their identification and/or assessment of abnormalities. In general, sensitivity and specificity and other measures of performance are estimated by applying the CAD system to a new data set, called a "test set", that is independent of the training set.

Note:

- CAD algorithms will tend to identify patterns similar to the patterns observed in the training set. If different training sets are used, then different CAD algorithms will be produced that will reflect differences in the patterns and patients observed between the training sets.
- Different CAD algorithms will emphasize different patterns and patients even if they are trained using the same set of patients.
- The patterns detected by a CAD are computed from complex transformations of the digital data that may not be directly known or understood by the users of a CAD device.
- CAD algorithms are typically not adaptive; instead they are fixed, and only change with new software revisions.

Many CAD devices internally perform the following steps when applied to medical data, though there is substantial variation of how they are implemented.

1. Signal/Image Processing

The input signal or image is often enhanced, or processed, to facilitate subsequent analysis by the particular CAD. This may include signal processing such as smoothing, sharpening, histogram equalization, etc.

2. Segmentation

Temporal boundaries on the fetal monitor tracing can be mapped to important time intervals (epochs) or to key clinical notations, e.g., rupture of membranes, use of oxygen, labor augmentation (oxytocin), etc. This could also correspond to segmenting tracing data into manageable sections for further processing.

3. Feature Calculation

The features of each epoch/region of the data are calculated. Features may be viewed as patterns that are computer or human estimated quantities characterizing information within the data. These features may or may not be calculable or understandable by the users. Different CAD devices will calculate and use different features, even for the same task.

4. Classification/Discrimination

The features obtained in step 3 are fed to statistical learning algorithms, or classifiers. The classifiers will generally output a single value. This value will indicate an estimated likelihood of an abnormality for the data from which the features were calculated. This likelihood will depend upon how the classifier was trained or optimized. A number of methods can be used to produce a classifier. Examples of such methods include logistic regression, nearest neighbor, neural network, decision tree, kernel machine, and linear or quadratic discriminate analysis. Different CAD devices may use different methods. Some CAD devices may combine multiple classifiers based on different methods and different sets of features.

5. Thresholding and Output

Generally a CADE device will produce a warning (monitoring alarm) or place a mark (imaging) on each data section when the classifier output for that region exceeds a certain threshold. A CADx device, when prompted on a particular section of data will most often display the classifier-based likelihood of an abnormality for that section, either directly or in some scaled or quantized form.

Each of the above steps can be modified by including or eliminating methods or by changing numerical software parameters in each step. This optimization or training can be performed using mathematical regression, performing an extensive search using a computer, having humans tweak parameters in the algorithm, or some combination thereof. When CAD developers make such modifications they are developing, or "training" the CAD device, and the data set being

used is therefore a training set. In general, these changes are made such that the performance of the CAD device increases on the data set. If CAD development is not carefully controlled, the CAD device can become overly specific to the training data set. When this happens, the CAD device may perform nearly perfectly when used on the training data, but performs poorly when applied to new data from the population. This condition is often referred to as “over-training” the CAD device. There are automated training methods to reduce the potential of such problems (e.g., cross-validation re-sampling). These methods require fully automated computer optimization of the CAD device.

Once CAD performance on the training data set is optimized, the CAD device can be applied to new data to estimate its actual performance. If the clinical condition (or disease status) of a new set of patients are known, and these data were not used to optimize the CAD device, and the data were randomly selected from the patient population, then the performance should be an unbiased estimate of how well the device will perform on the general population of patients. This generalization of performance is highly dependent on the quality of the data used to test the device.

CAD devices are computer software that performs extensive processing and analysis on medical data. Different CAD devices contain different software and utilize different processing, algorithms and features to identify abnormalities. The process of training algorithms or selecting features varies across CAD devices and this influences the performance of these devices. Most processing methods, algorithms, training and selection techniques are known and well described in published literature. There are, however, almost unlimited ways in which these methods can be combined and optimized to make a CAD device. To fully understand how a particular CAD device works the following would have to be known about that device:

- General Information
 - What is it targeted to detect?
 - For whom is the device intended?
 - On what kind of system(s) will the CAD be installed?
- Intended Device Usage
 - Will operations and settings be manual (by the physician), semi-automatic, or completely automatic (nothing is under physician control)?
 - What kind of output is generated by the CAD device? Is there user feedback?
 - What is the reading mode (see Section IV)?
- Processing
 - What processing is performed on the data (e.g., filtering, noise reduction, normalization)?
- Features
 - Which features were computed or evaluated during algorithm development?
 - How were features selected?
 - What is the mathematical formulation of the feature?
 - How does the feature relate to medicine/biology?
- Classifiers and Models
 - Which classification methods are used in the CAD algorithm (e.g., simple threshold, decision tree, linear classifier, neural network)
 - How were they optimized?

- Training and Test Databases
 - What types of abnormal and normal patterns and what types of patients are represented in the training and test sets?
 - Are the training and test sets representative of the intended use population?
 - What are the sample sizes of the training and test sets?
- Algorithm Stability
 - Is the CAD algorithm performance robust to minor changes to the algorithm or perturbations to the training or test data? Note, the stability of an algorithm increases as the number of training cases increases, the number or dimensionality of initial features decreases, or the complexity of the CAD decreases.

IV. Use of CAD devices in clinical practice

Reading paradigms (or modes) specify how a CAD device should be used by the clinician when interpreting cases clinically. FDA has identified three general paradigms for CAD implementations:

- First reader mode: The clinician reviews only findings identified by the CAD device. Unmarked findings may not be evaluated by the clinician. A device of this type has been approved for identifying cervical cytology slides that need no further review (AutoPap, P950009).
- Sequential (or Second) reading mode: The clinician first conducts a complete interpretation without the CAD device (unaided read) and then re-conducts an interpretation with the CAD device (aided read). This would probably not be typical for an intrapartum setting.

FDA has approved some CAD devices for this kind of use, e.g., mammography, lung imaging, etc. Some devices intended to be used in a sequential reading mode have been labeled with the "always-never rule," i.e. the clinician should always evaluate the data before turning the CAD on, and the clinician should never ignore findings detected before turning the CAD on in response to the CAD device not marking the finding. This is probably not useful in an intrapartum setting.

- Concurrent reading mode: CAD prompts are available at any time and the clinician performs a complete interpretation in the presence of CAD prompts or warnings. This is more typical of how a clinician is likely to use CAD information in an intrapartum setting.

V. Hierarchical model of efficacy

The discussion of CAD testing is a subset of a larger set of issues in medical diagnosis. In 1990, Fryback and Thornbury (2) described this for radiological imaging, framed in terms of a six-tiered hierarchical model of diagnostic efficacy. This hierarchy was further formalized by a report committee. An abbreviated version of this model is given in Table 1.

Table 1: The Six-Tiered or Hierarchical Model of Efficacy.

Level 1	Technical efficacy	Physical performance measurements, preclinical stand-alone and bench tests
Level 2	Diagnostic accuracy	Sensitivity, specificity, ROC curves, and their summary measures
Level 3	Diagnostic thinking	Effect of imaging test on clinicians’ subjective estimates of diagnostic probabilities, pre-test to post-test
Level 4	Therapeutic efficacy	Effect of diagnostic imaging or test on therapeutic management of patients
Level 5	Patient outcome	Expected value of test information in terms of gains in quality-adjusted life years (QALYs); also, cost per QALY gain.
Level 6	Societal efficacy	Cost-effectiveness and/or cost-benefit analysis from the societal viewpoint

The first tier, i.e., technical efficacy, includes measurements of the physical performance of imaging systems, as well as other bench tests and stand-alone measurements on CADs. The second tier, i.e., diagnostic accuracy, includes measurement of sensitivity, specificity, the receiver operating characteristic (ROC) curve and its summary measures (discussed in the following sections). For premarket evaluation of imaging-based CAD technologies, FDA commonly focuses on the first two or three tiers. Monitoring CAD systems may require an even higher level of efficacy. However, depending on the intended use of the device, testing at other levels may be appropriate. Higher level claims regarding efficacy of patient outcome or societal efficacy are typically in the domain of other agencies. Controlled studies at those higher levels are much broader in scope — and thus typically more difficult and expensive — than those in the first 2-3 efficacy tiers.

A general principle of the hierarchical framework is that, for a CAD to be efficacious at a higher level in the hierarchy, it must be efficacious at the lower levels. This is a necessary but not sufficient condition.

VI. Non-Clinical Testing

CAD devices that are discussed here are used to aid or assist the physician. They are not for use without physician oversight. Nevertheless, nonclinical testing of a CAD device, apart from a physician, can be of interest. Nonclinical testing of CAD devices could include

- *Stand-alone performance testing*: performance of the device by itself, i.e., Does the device identify the abnormality in the absence of physician interaction?
- *Reproducibility*. For example, if two EFM devices are used on a patient simultaneously, their tracings could vary. Do the CAD results vary depending on the tracing used?

Stand-alone CAD performance can be a useful barometer of CAD performance in a clinical use study. Stand alone performance can also be useful for comparing a modification to a CAD device to its original version. If a CAD device is to be used in first reader mode (as defined in Section IV), then stand alone performance may take on additional importance.

If cases are not selected carefully, then stand-alone testing can be subject to statistical bias, i.e., the tendency of a performance estimate to be misaligned with the true performance in the intended use population. Bias can be introduced in the selection of cases. Spectrum effect refers to selecting cases that tend to be easier to detect as normal or abnormal than cases in the intended use population. Spectrum effect tends to enhance stand-alone performance. At the same time, spectrum effect tends to shrink differences in stand alone performance between CAD devices because differences tend to be revealed among cases that are more difficult to classify.

Stand-alone performance can be based on diagnostic accuracy measures such as sensitivity, specificity, positive predictive value, and negative predictive value. The measures are defined relative to ground truth determination.

Ground truth determination could come from an independent modality (e.g., pathologic examination of tissue samples), follow-up over an appropriate time interval to demonstrate no change in a finding (e.g., one-year follow-up for mammographic findings), or from an expert panel assessing the available data. The ground truth determination by a panel of experts is susceptible to reader variability. Independent determination by each panel member provides measurement of the variability associated with the ground truth determination and/or allows for a consensus truth definition.

VII. Clinical Testing of Type 3 CAD devices

Clinical testing refers to physicians using the CAD device as intended. Clinical testing of EFM devices can be done during actual labor or retrospectively using tracings collected before to determine if the device is effective in physicians. When used properly, EFM type 3 device should enable the user to to predict a future event. Ground truth determination may not be available for all patients due to intervention (operational delivery) with the objective of preventing the event (e.g., metabolic acidosis). For more discussion, see the next section of the clinical event to be diagnosed.

A. Outcomes and Endpoints

For the Type 3 intended use, risk stratification for a future event, possible outcomes of interest include:

- Adverse or abnormal event (e.g., metabolic acidosis)
- Intervention (operative delivery other than for lack of progress)
- Non-intervention (spontaneous vaginal delivery)
- Time of intervention
- Diagnostic test results

- True positive (intervene before an adverse event occurs)
- False positive (intervene even though adverse event will not occur)
- True negative (do not intervene when adverse event will not occur)
- False negative (do not intervene before adverse event occurs)

Possible performance endpoints are

- Adverse (or abnormal) event rate
- Intervention rate
- Diagnostic endpoints
 - Sensitivity (true positive fraction among cases with adverse events)
 - Specificity (true negative fraction among cases without adverse events)
 - Positive predictive value (true positive fraction among cases that are test positive)
 - Negative predictive value (true negative fraction among cases that are test negative)
- Agreement between experts using the device and experts not using the device

In Type 3 EFM CAD testing, timing of intervention is critical for preventing an adverse outcome. Timing of intervention can play a role in determining the diagnostic test result. For EFM type 3 devices that predict future events, ground truth determination may not be available for all patients due to intervention (operational delivery) with the objective of preventing the event (e.g., metabolic acidosis). For more discussion, see the next section of the clinical event to be diagnosed.

B. Types of Clinical Studies

Several types of clinical studies are possible. Three are mentioned here:

(1) Prospective Randomized Controlled Trial (RCT)

In a prospective RCT, patients are randomized to two or more modalities, e.g., readings with or without the assistance of CAD. Randomization is typically stratified by “reader” (i.e., physician) to control for reader effects.

A prospective RCT is generally regarded as the best trial design for measuring the effect of the CAD on important clinical endpoints such as adverse event rate and intervention rate. However, prospective RCTs have drawbacks:

- The trial can be of significant risk to patients because patients are managed on the basis of the unapproved CAD device.
- The trial may have to be very large to have sufficient power to detect a statistically significant difference between the two modalities in the rate of a rare adverse event (e.g., CP). Alternatively, a more common event (e.g., metabolic acidosis, defined by cord artery pH < 7.10 and BDecf > 12 mmol/L) could be used as a surrogate for the rare event.
- The possibility exists that CAD effects could be explained by physicians reading a tracing differently when they use the CAD versus when they do not, irrespective of the CAD output (i.e., a differential in the Hawthorne effect between arms).

Unfortunately, many common endpoints for measuring diagnostic accuracy of the device cannot be estimated from the trial. The reason is that the ground truth determination of whether or not an adverse event occurred at spontaneous vaginal delivery can only be observed if spontaneous vaginal delivery was permitted, that is, if the physician did not intervene with an operational delivery. An intervention is a test positive result, but a determination of whether it is a true or false positive cannot be made because we do not know if an adverse event would have occurred had a normal delivery been permitted. Consequently, among the common diagnostic accuracy endpoints of sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV), only NPV can be estimated from the study because it does not depend on true or false positive results. However, under certain assumptions, accuracy when using CAD relative to when not using it can be estimated from the study. Endpoints of relative accuracy are discussed in the Statistical Considerations Section.

(2) Prospective Masked Study with Retrospective Analysis

A non-randomized masked study is a study where patients are monitored by readings without the assistance of CAD. To obtain information about CAD performance, the CAD output is recorded during the study, but masked to readers. After the study is over, the CAD output can be analyzed retrospectively to estimate CAD performance. A retrospective analysis of a prospective study is sometimes termed a prospective-retrospective analysis.

Retrospective analysis of the CAD output could reveal interventions that could have taken place on the basis of CAD to prevent adverse events. Thus, in theory, the relative reduction in the false negative rate ($1 - \text{sensitivity}$) could be estimated. For more discussion of endpoints that can be estimated from such masked studies, see the Statistical Considerations Section.

(3) Retrospective Multi-Reader Multi-Case (MRMC) Study

In a prospective RCT, patients are randomized to two or more reading modalities (e.g., readings with or without the assistance of CAD). The interpretation mode for prospective studies generally involves a single reader for each case in the course of routine clinical practice. Such studies may necessitate a prohibitive numbers of readers and cases due to reader and case variability as well as low prevalence of the event being detected.

Alternatively, in one type of reader study, which is retrospective but can be applied to prospectively collected data, a set of readers interpret a common set of patient data, in each of two competing reading conditions (e.g., readers unaided versus readers aided by CAD). Such studies are referred to as multiple reader multiple case (MRMC) design. The MRMC design can be “fully-crossed” whereby all of the readers independently read all of the cases. This design offers the most statistical power for a given number of truth-verified cases. Common statistical endpoints include sensitivity/specificity, PPV/NPV, and receiver operating characteristic (ROC) metrics.

The reading behavior in a retrospective study might not be the same as in a clinical setting for reasons that include the following:

- Retrospective reading is not the same as reading during an actual labor when making a clinical decision at the right time matters.

- When monitoring for rare events, retrospective studies are often enriched with patients experiencing the event to increase statistical power to obtain statistically significant results.

The hope is that when comparing reading modalities, relative performance is approximately preserved, even if absolute performance is not, so that CAD effects extrapolate to the clinical setting.

The data for a retrospective MRMC study could come from a prospective masked study. Patients with complete data would be natural to select for the study. However, for EFM Type 3 devices, patients with complete data are those in whom an intervention (operative delivery) did not take place. These patients are a random sample of test negative patients only (test-negative by EFM w/o CAD). Therefore, for the same reasons as mentioned above, an MRMC study on these patients will produce biased estimates of sensitivity, specificity, and PPV, but an unbiased estimate of NPV (assuming other factors do not bias the estimate). Under certain assumptions, some relative performance measures can be estimated without bias (see Statistical Considerations Section).

C. Statistical Considerations

(1) Evaluating a Diagnostic Test

Diagnostic tests are evaluated on their ability to discriminate patients with the condition of interest from patients without it. The evaluation can be characterized as a trade-off between truly and falsely detecting the condition. A diagnostic test is said to be informative if it is better than random classification for the condition. Several criteria imply that a test is informative, including (i) the true positive fraction (sensitivity) is greater than the false positive fraction ($1 - \text{specificity}$), (ii) PPV is greater than $1 - \text{NPV}$, and (iii) or the area under the ROC curve is greater than 0.5. For example, a diagnostic test with sensitivity 0.8 and specificity 0.2 is NOT informative, whereas one with sensitivity 0.7 and specificity 0.6 is informative.

Demonstrating that a test is informative may not be sufficient for it to have clinical utility. Clinical effectiveness depends on the magnitude of the test’s discriminatory ability as well as the seriousness in making false positive and false negative errors.

(2) Possible Outcomes from a Prospective Study

EFM Type 3 devices are designed to predict a future condition of interest, e.g., an adverse event upon spontaneous vaginal delivery. The possible outcomes of a prospective study can be outlined as follows:

<u>Intervention?</u>	<u>Adverse Event?</u>
0	0
0	1
1, operational	[0]
1, operational	[1]

1, too late 1
 1, lack of progress -

If an operational delivery intervention did not take place (lines 1 and 2), then the status of whether an adverse event occurred at spontaneous vaginal delivery can be observed. If an operational delivery intervention did take place (lines 3 and 4), then the status of the adverse event is missing, denoted by brackets in the table. If an intervention took place but was too late to prevent an adverse event (line 5), one possibility is to group this patient with those of line 2. If an intervention took place due to lack of progress (line 5), one possibility is to group this patient with those of line 1.

Because the status of the adverse event is known only for patients in whom an intervention didn’t take place, a test negative result (no intervention) can be determined to be true negative (no adverse event) or false negative (adverse event), but a true positive cannot be determined to be true positive or false positive. Consequently, among the commonly used diagnostic endpoints of sensitivity, specificity, PPV, and NPV, only NPV can be estimated from the study.

(3) Prospective Randomized Controlled Trial

In a prospective randomized controlled trial, diagnostic performance when using the CAD can be estimated relative to when not using the CAD. Patients are randomized to these two reading conditions. The data can be summarized as follows:

		Arm1: EFM			Arm2: EFM+CAD		
		AE	noAE		AE	noAE	
Test	-	n10	n00	n.0	m10	m00	m.0
	+	[n11]	[n01]	[n.1]	[m11]	[m01]	[m.1]
		[n1.]	[n0.]	n..	[m1.]	[m0.]	m..

For example, n10 is the number of patients who had an adverse event (AE) but were test negative by EFM. Brackets indicate missing data. For example, the number of patients m11 who had an adverse event but were test positive by EFM+CAD is missing (spontaneous delivery was not allowed). Because of randomization, the prevalence of an AE is the same in both arms. For this reason, the ratio of the specificities for EFM+CAD and EFM can be estimated as

$$rSp = (m00/m..) / (n00 / n..)$$

Likewise the ratio of false negative fractions (1 – sensitivity) can be estimated as

$$rFN = (m10/m..) / (n10 / n..)$$

Of course, the intervention rate and the rate of adverse events actually observed (as opposed to those that could have occurred had the interventions not taken place) can be estimated in each arm. These quantities are important in evaluating the clinical utility of the CAD, although they do not directly measure its diagnostic performance.

(4) Prospective Masked Study with Retrospective Analysis

For a non-randomized masked study in which CAD readings are evaluated retrospectively, the data can be summarized as follows:

		No AE			AE		
		EFM+CAD			EFM+CAD		
		-	+		-	+	
EFM	-	n000	n001	n00.	n100	n101	n10.
	+	[n010]	[n011]	[n01.]	[n110]	[n111]	[n11.]
		[n0.0]	[n0.1]	[n0..]	[n1.0]	[n1.1]	[n1..]

The NPV of EFM can be estimated as $n00. / (n00. + n10.)$. However, the diagnostic accuracy of EFM+CAD is limited to evaluation among EFM – patients, that is, patients allowed to have spontaneous delivery because only in those patients is the status of the adverse event at the unimpeded delivery time available.

Depending on the intended use, for some EFM CAD type 3 devices one could make the assumptions that (i) an intervention would certainly have taken place had the CAD alerted the reader and (ii) CAD would not have overruled a decision to intervene by EFM reading alone. Essentially, intervention takes place if either the EFM reading or the CAD itself renders a test positive result. This combination of EFM+CAD can be called the “believe the positive” rule. Under this assumption, one can show that the ratio of specificities of EFM+CAD to EFM alone can be estimated by

$$rSp = n000/n00.$$

Note that the estimate cannot be greater than 1, that is, when the “believe the positive” rule is used, specificity of EFM+CAD cannot be better than specificity of EFM alone.

Likewise, under the same assumption, the proportional reduction in the false negative rate (1 – sensitivity) due to EFM+CAD reading compared with EFM alone can be estimated as

$$FNreduct = n101/n10.$$

The proportional reduction is the difference between the false negative rate for EFM alone and the false negative rate for EFM+CAD divided by the false negative rate for EFM alone. With the “believe the positive” rule, the false negative rate cannot increase when adding CAD to EFM, thus the estimate can be no greater than one.

(5) Retrospective MRMC Study

In a retrospective MRMC study, the data can be summarized as follows:

		AE			no AE		
		EFM+CAD			EFM+CAD		
		-	+		-	+	
EFM	-	n000	n001	n00.	n100	n101	n10.
	+	n010	n011	n01.	n110	n111	n11.
		n0.0	n0.1	n0..	n1.0	n1.1	n1..

None of the numbers is missing, suggesting that the common diagnostic endpoints of sensitivity, specificity, NPV, and PPV are available for both EFM alone and EFM+CAD. However, the estimates of these quantities are naïve. Data for this study would be collected from a prospective RCT or a masked study. Presumably, only patients with complete data would be considered. Patients with complete data are those in whom an intervention didn’t take place because otherwise adverse event status upon spontaneous delivery is missing. Studying only these patients biases estimates of the four diagnostic endpoints mentioned except for NPV. Relative performance would also be biased. Under the “believe the positive” rule, estimates of some relative performance endpoints may be unbiased when comparing EFM+CAD retrospective readings with EFM readings from the prospective study from which the data were collected, but how to utilize the additional EFM retrospective readings in the MRMC study is unclear.

VIII. Testing of Type 2 CAD devices

A. Outcomes and Endpoints

Possible outcomes for Type 2 devices include:

- Baseline Heart Rate
- Variability
- Acceleration
- Decelerations
 - Early
 - Variable
 - Late
- Tachycardia
- Bradycardia

Possible performance endpoints are:

- Diagnostic (detection) endpoints
 - Sensitivity (true positive fraction among cases with the abnormality)
 - Specificity (true negative fraction among cases without the abnormality)
 - Positive predictive value (true positive fraction among cases that are test positive)
 - Negative predictive value (true negative fraction among cases that are test negative)

B. Clinical Testing

The type II CAD EFMs are generally used to detect features like baseline fetal heart rate, variability, acceleration, deceleration, and alerts. Similar to glucose monitors, which have alerts for identifying incidences of hypo- and hyper- glycemia, the EFM CADs with outcomes as described could be evaluated. These features could be evaluated against a reference standard. Substantial equivalence could be evaluated against a suitable predicate by performing agreement studies. Study samples should be representative samples from the intended use population and should have enough cases to have a precise estimate of the measures of agreement.

Depending on the features to be identified, we could construct 2x2 tables and examine the percentage of correct detection of features and correct detection of absence of features, (i.e. late deceleration).

		Predicate	
		Late Deceleration	Not Late Deceleration
Test	Late Deceleration	a	b
	Not Late Deceleration	c	d
Total		n1	n2

Percent correct detection of late deceleration (sensitivity) = $100 \% * (a/n1)$

Percent correct detection of absence of late deceleration (specificity) = $100 \% * (d/n1)$

Note that to capture the false negative information, the whole monitoring time needs to be divided into clinically meaningful segments where the feature is checked at each segment for its presence and/or absence. If only the features detected by the test device are verified, then one can only have an unbiased estimate of the positive predictive value (PPV) of the test. Diagnostic accuracy measures such as sensitivity, specificity and negative predicted value cannot be evaluated in an unbiased manner.

If the design permits segmentation of the entire monitoring episode to capture sensitivity and specificity, then the estimates should be reported with 95% confidence intervals, taking correlation into account. Since multiple readings will be evaluated from the same patient, which could lead to highly correlated data, the variance should be properly estimated to reflect the correlation structure.

If these devices are used as an adjunct to clinical decision making, then designs specific to such issues should be constructed. A multi-reader, multi-case design would be appropriate. A generalized estimating equation model can be constructed to estimate average percent correct detections (separately for presence as well as absence).

C. Nonclinical Testing

(1) *Check for reproducibility of the device.*

If possible, two or more devices could be attached to the mother to estimate the closeness between readings between the devices. The outcomes of continuous fetal heart rate could be modeled as repeated measures data to determine any significant deviation between the devices.

(2) *Phantom studies to check for accuracy of the feature detection.*

D. Statistical considerations for Type 2 Devices

If the outcome to be assessed is continuous (i.e. heart beats per minute or baseline heart rate) then an appropriate method comparison studies should be designed. Method comparisons between the test modality and a comparator can be assessed by Bland-Altman plots or standard regression (evaluating the regression equation against a straight line with slope=1 and intercept=0).

For dichotomous or binary outcomes, a 2x2 table as reported above can be constructed to estimate percent correct detection of abnormality and percent correct detection of absence of abnormality. To address this issue appropriately a tracing could be divided into mutually exclusive clinically relevant segments (i.e. segments with 30 minutes of tracing) and record features detected by the modality. Establish the presence or absence of the abnormality by a well established comparator or a clinical consensus to estimate sensitivity and specificity. Issues that need to be carefully addressed are that multiple observations are recorded per fetus and this may or may not lead to a highly correlated data. Emir et al (5) discuss ways to estimate such correlated observations in details. These estimates need to be reported with 95% confidence intervals which should take the correlated nature of the data into account.

Sample size may be an issue if each observation (in this case segments) is treated as independent units then a small number of cases may seem appropriate. However, multiple observations coming from the same patient may be highly correlated and hence the estimates may have little precision. The sample of patients should sufficiently cover the spectrum of patients (i.e. should have adequate number of patients with normal tracings as well as abnormal tracings)..

E. Types of tests and potential biases

CAD devices discussed here are used to aid or assist the clinician. They are not for use without clinician oversight. There are two primary categories of tests for CAD devices:

- *Stand-alone performance testing*: performance of the device by itself (i.e., Does the device identify the abnormality in the absence of clinician interaction?)
- *Reader performance testing*: performance of clinician using the device (i.e. Does the clinician recognize additional disease when using the CAD device?)

These two types of tests provide different types of information. Stand-alone test results provide a measure of intrinsic functionality of the device. Reader studies measure the impact on clinician performance. If, for example, the stand-alone sensitivity of a device was good, but the device only identified abnormalities that clinicians found on their own, then the device would be unlikely to improve performance. On the other hand, clinician performance may be improved by a device with poor stand-alone sensitivity, if the device identifies abnormalities that clinician tend to miss, and the clinicians recognize these as important when they are identified by the CAD device.

Stand-alone or reader performance testing is subject to many sources of statistical bias that can often be minimized through good study design. Statistical bias is a tendency for a performance estimate in a study to be misaligned with the true performance in the intended use population.

Stand-alone and reader performance testing both require ground truth determination. Ground truth determination includes whether or not disease is present within a patient as well as the precise location of disease. Ground truth determination could come from an independent modality (e.g., pathologic examination of tissue samples.), follow-up over an appropriate time interval to demonstrate no change in a finding (e.g., one-year follow-up for mammographic findings, or from an expert panel assessing the available data. Ground truth could also come from a panel of experts reading the same modality as the reviewing clinicians. This may be appropriate for intrapartum monitoring where clinical intervention may complicate the determination of the underlying true pathology. Ground truth determination by a panel of experts is susceptible to reader variability. Independent determination by each panel member provides measurement of the variability associated with the ground truth determination and/or allows for a consensus truth definition.

F. Multiple Reader Multiple Case (MRMC) Reader Study Design

In a prospective study design, patients are randomized into two or more reading modalities (e.g., readings with or without the assistance of CAD). The interpretation mode for prospective studies generally involves a single clinician evaluating each patient in the course of routine clinical practice. Such studies may necessitate a prohibitive number of readers and cases to properly account for reader and case variability as well as the low disease prevalence.

In an FDA regulatory setting, large-scale prospective studies are occasionally a condition for premarket clearance/approval of a device because they can help answer questions requiring long-term follow-up or capturing rare event rates.

An alternative to prospective studies may be the use of retrospective data in controlled reader studies. Although reading behavior in a retrospective study might not be the same as in a clinical setting (3), the hope is that when comparing reading modalities, relative performance is approximately preserved, even if absolute performance is not, so that CAD effects extrapolate to the clinical setting.

In one type of reader study, which is most commonly applied to retrospective data but can be applied to prospectively collected data, a set of readers interpret a common set of patient data, in each of two competing reading conditions (e.g., readers unaided versus readers aided by CAD). Such studies are referred to as multiple reader multiple case (MRMC) design. The MRMC design can be "fully-crossed" whereby all of the readers independently read all of the cases. This design offers the most statistical power for a given number of truth-verified cases. Common statistical endpoints include sensitivity/specificity, PPV/NPV, as well as receiver operating characteristic metrics. Wagner et al. provide a tutorial of the evaluation of imaging systems and computer aid devices (4).

The basic endpoints discussed above may not all be appropriate in intrapartum monitoring studies. For example, specificity may not be well defined for a monitoring task. A more appropriate alternative could be false positives/epoch or some other measure of performance. Each different measure brings statistical complexity that must be addressed in any reader study to appropriately assess and compare these devices.

IX. References

1. FDA Radiological Devices Panel Meeting, March, 2008, Briefing Package. <http://www.fda.gov/ohrms/dockets/ac/08/briefing/2008-4349b1-01%20FDA%20Radiological%20Devices%20Panel%20Meeting%20Intro.pdf>. Effective March 2008. Accessed October 2, 2008.
2. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. *Med-Decis-Making* 1991; 11:88-94.
3. Gur D, Bandos AI, Cohen CS, et al. The "laboratory" effect: comparing radiologists' performance and variability during prospective clinical and laboratory mammography interpretations. *Radiology* 2008; 249:47-53.
4. Wagner RF, Metz CE, Campbell G. Assessment of Medical Imaging Systems and Computer Aids: A Tutorial Review. *Acad-Radiol* 2007; 14:723-748.
5. Emir B, Wieand S, Su JQ, Cha S. Analysis of Repeated Markers Used to Predict Progression of Cancer. *Statist. Med.* 1998; 17:2563-2578.

Additional References

Devoe L, Golde S, Kilman Y, Morton D, Shea K, Waller J, A comparison of visual analyses of intrapartum fetal heart rate tracings according to the new national institute of child health and human development guidelines with computer analyses by an automated fetal heart rate monitoring system, *Am J Obstet Gynecol.* 2001 Jun;184(7):1587-8.

Ayres-de-Campos D, Bernardes J. Comparison of fetal heart rate baseline estimation by SisPorto 2.01 and a consensus of clinicians. *Eur J Obstet Gynecol Reprod Biol* 2004; 117(2):174-8.

Ayres-de-Campos D, Costa-Santos C, Bernardes J. Prediction of neonatal state by computer analysis of fetal heart rate tracings: the antpartum arm of the SisPorto multicentre validation study. *Eur J Obstet Gynecol Reprod Biol* 2005; 118(1):52-60.

Jezewski M, Wrobel J, Labaj P, et al, Some Practical Remarks on Neural Networks Approach to Fetal Cardiotocograms Classification, Proceedings of the 29th Annual International Conference of the IEEE EMBS