# Study Design and Analysis Issues for Diagnostic Monitoring Devices

Bipasa Biswas
CDRH, FDA, 10903 New Hampshire Avenue, Silver Spring, MD 20993

## Abstract
A variety of diagnostic devices often require a need for detection and localization of multiple events within a subject over a time course used mainly for monitoring of events e.g. the detection of events like seizures in temporal intervals from a scalp EEG reading. It is possible to characterize event detection in the presence of a reference standard and a priori time window overlap or proximity criteria, by sensitivity which is the probability of detecting the event given that the event has actually occurred. However, specificity, which is the probability of a non-event given that the event did not occur, cannot be easily estimated. Instead, counting the false positives and reporting the rate of false positives is an alternative to evaluate device performance. A method to assess that the device is not randomly marking events is presented along with discussions of sample size calculations that address the intra-cluster variability due to multiple events per subject.

**Key Words:** Event monitoring, sensitivity, false positive rate, sample size, Poisson Process.

## 1. Event monitoring devices

A general definition of monitoring translates to test or sample, especially on a regular or ongoing basis. Monitoring involves repeated examination of subject condition to assess potential change in status (e.g. change in treatment response, progression of disease). Monitoring can be quantitative (like blood pressure measurement), semi-quantitative or ordinal; or binary (e.g. disease status, event monitoring).

The focus of this paper is event monitoring devices with particular attention to devices used to detect and localize multiple events within a subject over a time interval. Examples of devices that marks an event of interest for event monitoring devices include software's to detect spikes and seizures on a EEG strip reading; electronic fetal monitors that detect specific patterns that are of specific clinical interest like deceleration, sinusoidal fetal heart rate pattern; and continuous glucose monitors to detect hypoglycemic or hyperglycemic events. Each of these devices involves monitoring over a continuous period of time to detect events of interest.

### 1.1 Study Design
To evaluate event monitoring devices that detects and localizes events of interest, it is important to have a clear and specific definition of the event. Thus, it is important to characterize the actual event by specifying the beginning and the end of the event by the clinical reference standard which is discussed in details in De et al (2013). The study to evaluate the performance involves the basic design requirements that include firstly a

sample of subjects from intent of use population; an independent method to assess truth i.e. characterizing the event by a clinical reference standard; masking the output from the test device to the reference and vice versa and an application of both the device as well as the reference method on same subjects for the same interval of time. An interval selection, a priori, to determine localization and detection by the device around the event determined by the reference method is essential to pre-specify prior to analysis.

Some known sources of bias in such designs would be bias due to imperfect reference method i.e. the method used to characterize the actual event, failure to mask the output from the test device and the reference method, use of the test device output to construct the "true" event, and selection bias by sampling subjects that do not reflect the intended use population of the device. Many other sources of bias are discussed in Begg (1987), Pepe (2003), and Zhou et al (2002).

## 1.2 Performance Measures

Based on the known locations of all events and the locations of all marks, the accuracy of the device can be characterized by the strength of association between the numbers and locations of marks and the numbers and locations of the events. For practical purposes, the primary interest lies in properties of a set of marks to adequately detect and localize the events for purpose of decisions of further management. Thus the information can be classified as beneficial information by adequately localized events; missed or inadequately localized events leading to detrimental information and false marks that do not correspond to any event.

In general, diagnostic devices with a qualitative outcome (with two outcomes) e.g. presence or absence of the condition of interest, is evaluated against a reference method used to establish the true condition. Generally, a qualitative test with a binary output is represented by a 2x2 table:

Study Population

| | | Target Condition "Truth" | |
|---|---|---|---|
| | | D=1 | D=0 |
| Test | Y=1 | TP | FP |
| | Y=0 | FN | TN |

The variable D represents the target condition where D=1 means the condition is present and D=0 means the condition is absent and the test is represented by the variable Y where Y=1 means the test is positive and Y=0 means the test is negative. In the above table, "TP" denotes True Positive; "FP" denotes False Positive; "FN" denotes False Negative; "TN" denotes True Negative.

Then the accuracy of the test is evaluated by either the sensitivity-specificity pair, or the positive predictive value (PPV) and the negative predictive value (NPV) which are defined as follows:.

Sensitivity (TPF) = P(Y=1|D=1) estimated by TP/(TP+FN)
Specificity (1-FPF) =P(Y=0|D=0) estimated by TN/(TN+FP)

Positive predictive value (PPV) = P(D=1|Y=1) estimated by TP/(TP+FP)
Negative predictive value (NPV) = P(D=0|Y=0) estimated by TN/(FN+TN)

The performance measures PPV and NPV depend on the prevalence of the true target condition.

However, unlike a typical qualitative diagnostic device with two outputs, a monitoring device which marks the targets of interest i.e. events in this case, specificity is not estimable as splitting the temporal interval into sub-intervals is complicated. This is illustrated in an example of fetal monitor data readout represented in **Figure 1** where upper part of the readout represents the fetal heart rate and the lower part of the readout represents the mother's uterine contractions. Say a particular pattern for e.g. deceleration of fetal heart rate is the event of interest. Thus to assess this abnormal event location is important and hit determination rule is critical component. Additionally, splitting the readout into intervals is complicated as depending on how these intervals are determined for e.g. **Figure 2** versus **Figure 3** the event detection and localization gets complicated. Thus, in the absence of well-defined interval windows a general determination of true negative is complicated and cannot be done for such situations.

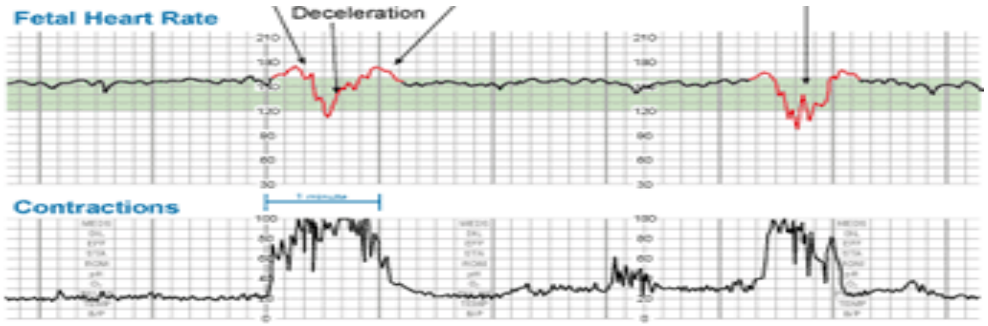


**Figure 1:** Readout from a fetal monitor with upper part representing the fetal heart rate and the lower part represents uterine activity.

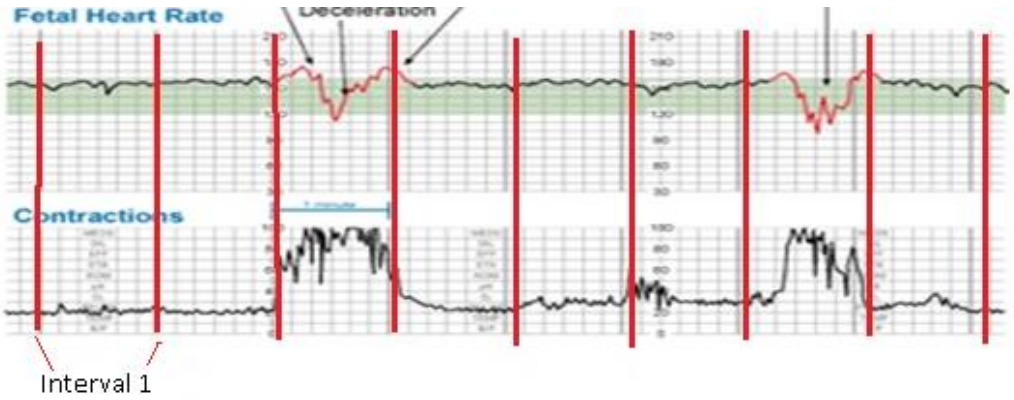


**Figure 2:** Splitting the readout into intervals.

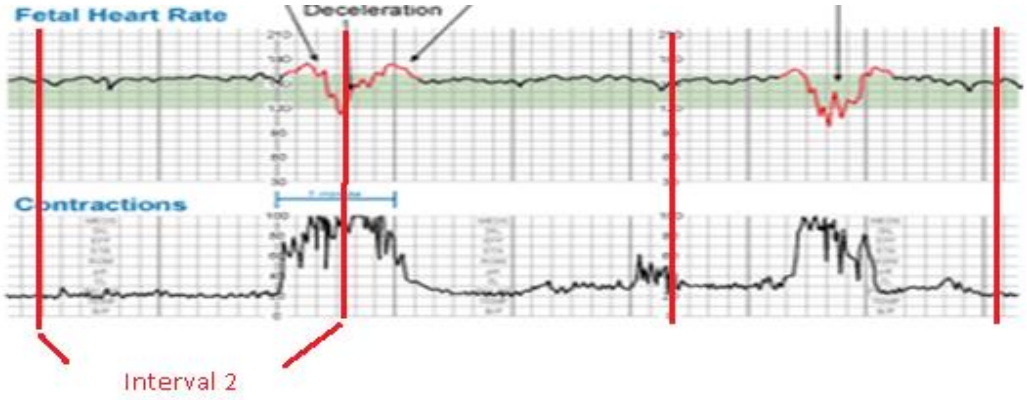


**Figure 3:** Splitting the readout into intervals.

Instead the focus is on detection and localization of the events. Thus, it is essential to specify the beginning and the end of an event by the reference standard, and a well-specified interval around the event to determine adequate localization of the mark. To determine whether device mark corresponds with an actual event, the mark should lie inside the interval of the event also known as proximity criteria or acceptance region.

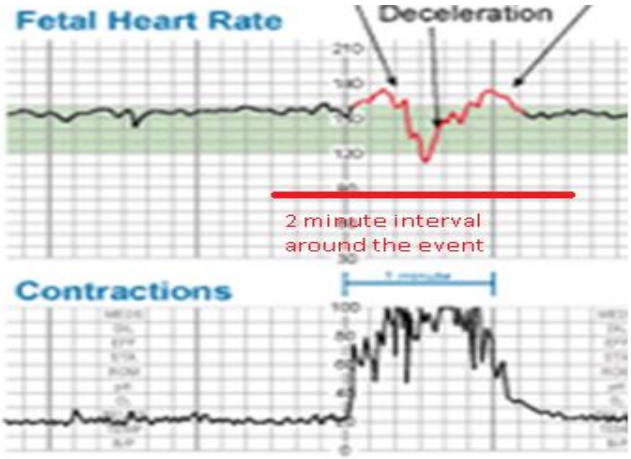


**Figure 4:** Hit determination rule for detection and localization of event (deceleration in fetal heart rate).

For most practical purposes a certain limit is acceptable within which a localization error can be ignored. Thus in **Figure 4** as long as the device marks in the 2 minute interval around the event the event will be considered detected and localized. Such a mark is called true positive and the localized event is considered detected. Marks that do not have a corresponding event within the regions determined by the proximity criteria are called false positives and finally events that have no marks closer than determined by the proximity criteria is a false negative. The definitions of true positive, false positive and false negative depend on the adopted proximity criteria. Classification based on such rules can be complicated by establishing correspondence between the marks and the events and handling multiple marks that are adequately close to the same event. The

statistical framework in this paper assumes for a considered proximity criteria, a single event corresponds to a single mark.

The simplest type of data collected is on a single subject. Based on the locations of e known events, and applying the proximity criteria (overlapping window length) for matching the device marking to an event, one can define which known events are detected and adequately localized (true positive marks), and which marks do not correspond to any events (False Positive marks). Then True Positive (TP) and False Positive (FP) are realization of random variables X and Y where

#TP=X ($X \leq e$)
#FP=Y
#FN=e-X

Frequencies of TP and FP are characterized by proportion of detected events (TPF) and the rate of false responses per subject (FPR) given as:

$$\mathrm{TPF} = \frac{\mathrm{E(X)}}{\mathrm{e}};$$
$$\mathrm{FPR} = \mathrm{E(Y)}$$

If the random variable H represents detection and localization of event from a random population of events, then the TPF is given by

$$\mathrm{TPF} = \frac{\mathrm{E(X)}}{\mathrm{e}} = \frac{\mathrm{E}(\sum_{i=1}^{e} \mathrm{I}(\mathrm{H_i} = 1))}{\mathrm{e}} = \frac{\mathrm{e} \times \mathrm{P(H = 1)}}{\mathrm{e}} = \mathrm{P(H = 1)}$$

Thus TPF, the probability of detection and localization of an event, is interpreted as a probability and FPR is interpreted as a rate. The evaluation of performance is based on both TPF and FPR simultaneously, since one can increase one measure of performance at the expense of other. Thus, the frequency of correct decisions should always be assessed in conjunction and relative to the number of errors. For a perfect device the device marks all events with no error i.e. X=e and Y=0 implying TPF=1 and FPR=0. The poorest performing device will yield only erroneous marks i.e. X=0 implying TPF=0.

Generalizing to S subjects the data can be represented as:

$$\{e_s, x_s, y_s\} \; s = 1,2, \ldots, S$$

Where s indicates one of S subjects and $e_s$ is the number of events in subject s; $x_s$ is the number of TP in subject s and $y_s$ is the number of FP in subject s.
Then the estimates of event specific performance measures are given by

$$\widehat{\mathrm{FPR}} = \frac{\sum_{s=1}^{S} y_s}{S}$$
$$\widehat{\mathrm{TPF}} = \frac{\sum_{s=1}^{S} x_s}{\sum_{s=1}^{S} e_s} = \sum_{s=1}^{S} \frac{e_s}{\sum_{s=1}^{S} e_s} \left(\frac{x_s}{e_s}\right) = \sum_{s=1}^{S} w_s \widehat{\mathrm{TPF}_s} \qquad (1)$$

And the estimates of subject specific performance measures are given by

$$\widehat{\mathrm{FPR}} = \frac{\sum_{s=1}^{S} y_s}{S}$$
$$\widehat{\mathrm{TPF}} = \sum_{s=1}^{S} \frac{\left(\frac{x_s}{e_s}\right)}{S} = \sum_{s=1}^{S} w\widehat{\mathrm{TPF}_s} \qquad (2)$$

The difference between event-specific and subject-specific performance lies in the weights associated in calculating the TPF. These weights will be same if each subject

contributes equal number of events and then subject-specific TPF is the same as the event-specific TPF. But for most practical purposes the subjects may not contribute the same number of events and thus it is important to distinguish the measures.

## 1.2 A random test in case of event monitoring

A simple approach to characterize the random placement of marks is to consider a homogeneous Poisson Process with distribution of random number $n$ of marks at random locations along the temporal interval of a reading with length say $L$. Say $\lambda$ is the average number of marks of the Poisson Process and $L$ is the total length of interval over which events are evaluated. For this paper, multiple marks for an event are counted as one mark and say $l$ is the length of the interval around the event and for simplicity is located entirely inside the reading of length $L$, is considered as the proximity criteria. We can define the following indicator function:

$$\delta = \begin{cases} 1 & \text{if the event is marked within interval of length } l \\ 0 & \text{if the event is missed} \end{cases}$$

Then the probability of marking the event by chance is given by $p = \frac{l}{L}$

Thus, for a given rate $\lambda$, we can express

$$\begin{aligned} \text{TPF} = 1 - \text{P}\{\textstyle\sum_{i=1}^{n}(\delta_i = 0)\} &= 1 - \text{E}[(1-\text{p})^{\text{n}}] \\ &= 1 - \sum_{\text{k}=0}^{\infty}(1-\text{p})^{\text{k}}\frac{\lambda^{\text{k}}}{\text{k!}}\text{e}^{-\lambda} \\ &= 1 - \text{e}^{-\lambda}\sum_{\text{k}=0}^{\infty}\frac{[\lambda(1-\text{p})]^{\text{k}}}{\text{k!}} \\ &= 1 - \text{e}^{-\lambda}\,\text{e}^{\lambda(1-\text{p})} \\ &= 1 - \text{e}^{-\lambda\text{p}} \end{aligned}$$

$$\text{FPR} = \text{E}[\text{n} - \textstyle\sum_{\text{i}=1}^{\text{n}}\delta_{\text{i}}] = \text{E}(\text{n}) - \text{E}(\sum_{\text{i}=1}^{\text{n}}\delta_{\text{i}}) = \lambda - \lambda\text{p} = \lambda(1-\text{p})$$

Thus if $\varphi = \frac{\text{p}}{(1-\text{p})}$ then

$$\begin{aligned} \text{TPF} = 1 - \text{e}^{-\lambda\text{p}} &= 1 - \text{e}^{-\lambda\text{p}\left(\frac{(1-\text{p})}{(1-\text{p})}\right)} \\ &= 1 - \text{e}^{-\lambda(1-\text{p})\left(\frac{\text{p}}{1-\text{p}}\right)} \\ &= 1 - \text{e}^{-\text{FPR}\times\varphi} \end{aligned} \qquad (3)$$

As an example if one would like to determine if a software that detects deceleration of fetal heart rate in electronic fetal heart rate machine readout in a one hour reading, is better than random, then one can use the equation above to determine for a given false positive rate and the adopted proximity criteria how much more the TPF has to be. Thus for EFM measurements with 1 hour reading per subject we have $L = 60\,minutes$ and say for matching the device marking to an actual events the proximity criteria is an interval of length $l = 2$ minutes to mark on the EFM strip for event determination.

Then $\text{p} = \frac{2}{60}$=0.0333; $\varphi = \frac{0.0333}{1-0.0333}$=0.0345

Thus if a device marks an event and the matching interval is of length 2 minutes has FPR=2.0 then any $\text{TPF} > 1 - \exp(-2 \times 0.0345) = 0.067$ will be better than random markings.
Now say the FPR=20.0 instead of 2.0 as above then any $\text{TPF} > 1 - \exp(-20 \times 0.0345) = 0.502$ will be better than random markings.
In the above examples, it was implicitly assumed that the TPF-FPR pairs for the tests are known without errors. A formal comparison can be made by computing confidence intervals of TPF and FPR.

## 2.Sample Size and Reporting

If the goal is to evaluate a diagnostic event monitoring device explained in this paper against a known performance goal then a sample size can be based on evaluating the performance against the performance goal with adequate power and alpha (the probability of type 1 error). Say $(\text{TPF}_0, \text{FPR}_0)$ are the minimally required performance goals for TPF and FPR then the null hypothesis is
$\text{H}_0 = \{\text{TPF} \leq \text{TPF}_0, \text{ or } \text{FPR} \geq \text{FPR}_0\}$
The sample size for number of events and number of false positive marks is determined for size= $\alpha/2$ and power $(1-\beta)$ at an alternative hypothesis $\text{H}_1: \text{TPF} = \text{TPF}_1$ and $\text{FPR} = \text{FPR}_1$.

$$\text{n}_{\text{TPF}} = \frac{(\sigma_{\text{TPF}})^2 \left(\text{z}_{\alpha/2} - \text{z}_{(1-\beta)}\right)^2}{(\text{TPF}_1 - \text{TPF}_0)^2}$$

Where $(\sigma_{\text{TPF}})^2 = \text{d}_\text{u}\text{TPF}_1(1 - \text{TPF}_1)$
$\text{d}_\text{u} = \bar{\text{m}}^{-1}\left\{1 + \left(\frac{\bar{\bar{\text{m}}}}{\bar{\text{m}}} - 1\right)\rho\right\}$ $\bar{\text{m}} = \text{E}(\text{m})$ and $\bar{\bar{\text{m}}} = \text{E}(\text{m}^2)$. Where m is the number of events per subject and $\bar{\text{m}}$ is average number of events and $\bar{\bar{\text{m}}}$ is the average of square of the events and ρ is the intra-cluster correlation. The details of the sample size calculation can be found in Jung et al (2001).

$$\text{n}_{\text{FPR}} = \frac{(\text{FPR}_1 \frac{\overline{1}}{\text{m}} + \sigma_\text{C}^2)\left(\text{z}_{\alpha/2} - \text{z}_{(1-\beta)}\right)^2}{(\text{FPR}_1 - \text{FPR}_0)^2}$$

where $\sigma_\text{C}^2$ =between cluster variance. The details of the sample size calculation can be found in Hayes and Bennett (1999).

The sample size can also be based on comparing to a predicate in the same study and the details are not considered further in this paper. The above sample size derivation is for a single case when the evaluation of performance is set against a performance goal.

While reporting performance of such event monitoring devices, with focus on localizing and detecting events of interest, both TPF and FPR are reported simultaneously along with 95% confidence intervals. The nonparametric estimation of these measures can be given by (1) above and the 95% confidence intervals should take the correlation into account due to multiple events and multiple marks from the same subject. Thus in particular the variance of the point estimates of TPF and FPR should include the correlation due to clusters. Variance estimation of these point estimates can be assessed using the variance estimators for clustered proportions (Rao and Scott 1992; Donner and Klar 2000).

## 3. Conclusion

For practical purposes the primary interest in evaluating diagnostic monitoring devices which mainly detects and localizes events of interest in temporal intervals, the primary interest often lies in properties of the marks to adequately detect and localize the events for purposes of making decisions for further management. Simplest way to characterize the performance can be based on the association of the number and the location of marks by the device and the number and location of the actual events which have been independently determined by the reference standard. Unlike typical diagnostic devices with binary outcomes (presence or absence), the true negative determination for such event detection and localization devices is generally not definable and thus the usefulness of the device marking is characterized by true event detection and localization (true positives), false markings (false positives) and missed events (false negatives).

Thus to evaluate performance, there is a need to specify a clinical reference standard, a priori window of practical interval length as proximity criteria for detecting and localizing events by a mark, and report both true positive fraction (TPF) and false positive rate (FPR) along with 95% confidence intervals of the point estimate. Usefulness of the device lies if the true positive markings are significantly greater than the false positive marks but unlike a typical diagnostic device with binary outcome it is not a straight forward assessment whether the markings are random or informative.

Sample size is based on clustered data as same subjects provide multiple targets and the correlation needs to be accounted due to clustered data while reporting results.

## Acknowledgements

The author acknowledges the help and support of Arkendra De PhD, Lakshmi Vishnuvajjala PhD and Division of Biostatistics.

## References

(1) Bandos A, Rockette HE, Song T, Gur D. Area under the Free-Response ROC Curve (FROC) and a Related Summary Index, Biometrics. 2009 March ; 65(1): 247–256

(2) Hayes RJ, Bennett S. Sample size calculation for cluster-randomized trials. International Journal of Epidemiology 1999; 28:319–326.

(3) Donner A, Klar N. Design and Analysis of Cluster Randomization Trial in Health Research. Wiley, 2000.

(4) Jung SH, Kang SH, Chul AW. Sample size calculations for clustered binary data, Statist. Med. 2001; 20:1971–1982.

(5) Rao JNK and Scott AJ. A simple method for the analysis of clustered binary data. Biometrics 1992; 48:577-585.

(6) De A, Meier K, Tang R, Li M, Gwise T, Gomatam S, Pennello G. Evaluation of Heart Failure Biomarker Tests: A Survey of Statistical Considerations. Journal of Cardiovascular Translational Research. 2013; 6:449-457 doi:10.1007/s12265-013-9470-3

(7) Pepe MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press, 2003.

(8) Zhou XH, Obuchowski, NA, McClish DK. Statistical Methods in diagnostic Medicine – second Edition. Wiley, New York 2011.

(9) Begg CB. Biases in the assessment of diagnostic tests. Statistics in Medicine. 1987, 6:411-423