XXI IMEKO World Congress "Measurement in Research and Industry"
August 30 – September 4, 2015, Prague, Czech Republic

TREATMENT OF OUTLIERS IN IN-PROCESS APPLICATIONS: A ROC-BASED APPROACH

Giampaolo E. D'Errico

Istituto Nazionale di Ricerca Metrologica (INRIM), Torino (I), e-mail: g.derrico@inrim.it

Abstract – A new approach, based on the receiver operating characteristics (ROC) technique, is proposed to tackle the problem of adaptive detection of suspected outliers in dynamic measurements. Modelling and analysis are presented and discussed with application to in-process metrology, focusing on adaptive monitoring and control systems.

Keywords: outliers, dynamic systems, in-process metrology, ROC-based treatment

1. INTRODUCTION

Statistical criteria and rules for outlier treatment are provided in the ASTM standard practice [1]. From a theoretical point of view, the possible presence of outlying observations poses at least two kinds of problems, synthesized below.

First, criteria for outlier testing are of a stipulated nature: this implies that an observation considered an outlier according to a criterion may be taken for an inlier (i.e., a non-outlier) by another criterion. Such discordance may happen even if criteria referred to relay upon the same theoretical foundations (e.g. in classical statistics, Dixon criterion [2] and Grubb criterion [3]).

Second, given that a suspected observation is recognized as an outlier, it is not trivial how to perform a consequent statistical treatment. In fact, if an outlier is rejected from a sample set, not only the underlying assumption of independence of observations is violated but the very randomness of the sampling process is eventually lost.

Problems become more difficult in in-process applications. In real world conditions—such as in industrial process monitoring and control—where measurements of time-variant quantities are involved, the sample set itself is a dynamic entity to be constructed in real time.

Therefore, given a general definition of an outlier as 'an observation that deviates significantly from others in the sample set' (see [4], also for other definitions), the observation process is to be terminated before evaluating whether a single observation is deviant or not.

In the present paper, these problems are tackled from a statistical approach developed on the base of the so-called receiver operating characteristics (ROC) analysis [5], [6]. The ROC-based methodics is finalised in terms of an operative model, namely the ROC curve: this is a tool to comparatively evaluate the costs of probable errors over a locus of trade-off points (the ROC curve) between possible cases of false negatives and false positives.

The present paper is organised as follows. In Section 2, state-of-the-art arguments are mentioned briefly. In Section 3 the problem of outliers is situated in the framework of adaptive, self-tuning systems for real time monitoring and process control, where on-line system identification involves in-process metrology. In Section 4 the proposed approach is developed by fitting the ROC technique into the model of soft classification of possible outliers in dynamic measurement. Section 5 closes the paper with concluding remarks and future perspectives.

2. THE METHODOLOGICAL AND TAXONOMIC BACKGROUND

Outliers entail both practical and epistemic aspects: whereas an outlier might indicate a miscalibration or fault in instrumentation, or even a trivial reporting mistake, an unexpected experimental result might be announcing the discovery of an unforeseen phenomenon.

A side-problem of relevant scientific importance is in terms of so-called false discoveries. In these terms, a provisional rejection of a null hypothesis is called a discovery, and false discoveries are related to type I errors. A type I error occurs if the null hypothesis is wrongly (with a probability correponding to the significance level) rejected, when it is true (e.g., [7], [8]).

Moreover, 'significant'—in the statistical sense—is one of the most disputed definitions in scientific literature [9], and it is even claimed that most published findings are false [10].

In-process metrology is challenged by dynamic applications, where the treatment of possible outliers is a task requiring real-time accomplishment. In this framework, the sample set remains itself undefined until the observation cycle is completed. At each cycle, each single sample set is processed in real-time; at the end, all the stored sample sets can be aggregated for post-process analysis of collected data.

A related issue—known under the name of Simpson's paradox—can be raised if data are aggregated from two separated groupings (partitions): for example, two diverse realizations of the whole process under observation. The probabilistic plausibility of a conclusion drawn from the analysis of each data subset might be contradicted (the so-called reversal phenomenon [11]) when data are analised as a whole.

A general taxonomy of outlier detection methods can be found in [4]: these methods includes so called hard classifiers and soft classifiers.

Hard classifiers partition a dataset into two separate subsets according to a fixed threshold to decide sharp assignment of a data-point either to the data-subset of outliers or to the data-subset of non-outliers (inliers).

Soft classifiers refer to adaptive classification criteria to rank each data-point according to its degree of so-called outlierness (or outlyingness) over a scale. For instance, an outlierness degree is used for a classification based on fuzzy sets in [12].

As pointed-out in [13], most application-specific criteria refer to a measure of outlierness to score a data-point over a scale spanning from normal data to uninteresting noise to interesting anomalies.

A Bayesian approach to testing for outliers is developed in [14]; as an alternative to test statistical significance of outliers, a fuzzy treatment of them is presented in [12].

Furthermore, there are diverse outlier models [13]: in the present work, the focus is on a probabilistic and statistical model where the soft classifier can be tuned in order to balance the cost of missing an outlier (false negative) against the cost of misclassifying an inlier as an outlier (false positive).

In this model, tuning the cost balance translates into adapting the trade-off between false negative/false positive probabilities over a curve, namely a ROC curve.

The ROC-based approach was originated in the framework of signal detection theory: it was aimed at modelling the performance of a radar detector (the receiver) to detect approaching aircrafts. After that, its mainstream development is advanced in clinical treatments, medical diagnostics, and imaging applications (e.g., [15], [16]): a tutorial on related ROC methodology can be found in [17].

3. OUTLIERS IN DYNAMIC SYSTEMS

An application field that bears relevance to in-process metrology (e.g., [18], [19] and [20]) is industrial process monitoring and control (e.g., [21]). Especially in the area of design and implementation of adaptive controllers [22], [23], [24], on-line system identification is integrated with real-time adaptation of control laws.

Fig. 1 shows the layout of a self-tuning regulator (STR): the underlying philosophy is that the process parameters may be slowly (respect to the input/output feedback loop) time-varying, so that a real-time estimation is needed.

In a STR, the involved parameters are estimated according to estimation procedures of metrological interest: measurement uncertainty and presence of possible bias are taken into account. A peculiar difficulty of this monitoring and control task is to distinguish signal variations due to the process dynamics from possible intrusion of outliers in measurements.

Another principle scheme of adaptive control is underpinned by the idea that the process parameters are constant but unknown. The layout of this scheme, known as model reference adaptive control system (MRAS), is shown in Fig. 2.

In an MRAS the difficulties are related to the fact that deviations of the identified system from its model may be overestimated in the presence of undetected outliers, giving rise to possible instability of the closed loop system due to overcompensations commanded by the controller.

Procedures for estimation of system parameters and measurement of system signals, which take into account involved uncertainties, are established in this area long since ([23] goes back more than half a century): for example, Kalman filtering is a paradigm in process monitoring (e.g., [18]). However, treatments of possible occurrence of outlying observations in control systems are still a research topic.

As overviewed in Section 2, criteria for outliers detection can be formalised from different methodological standpoints and in view of specific application contexts.

More specifically, focusing on outliers in process identification, some diverse strategies can be found, for example in [25] (based on Bayesian inference and quadratic loss function), [26] (with application to time series models inclusive of aberrant observation or aberrant innovations), and [27] (using a Hampel filter for predictive control robustness).

To manage outliers in dynamic systems, outlying/ inlying observations can be ranked on the base of soft classification criteria, according to a given model.

In the next Section, a ROC-based model is introduced and fitted to purpose.

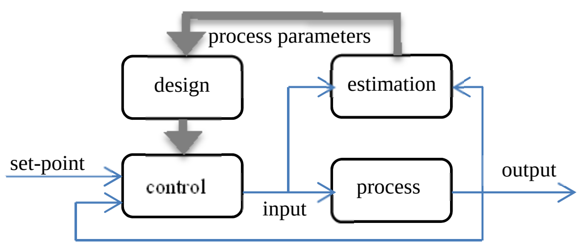

Fig. 1. Schematic block diagram of a self-tuning regulator (input: actuator signal, output: controlled signal); real-time estimation of process parameters (solid connectors and arrows) is a dynamic measurement task.

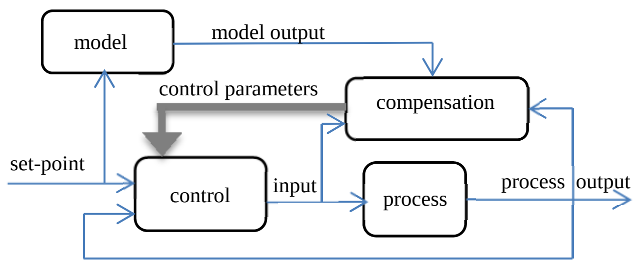

Fig. 2. Schematic block diagram of a model reference adaptive control: given the same set-point, the process output is compared to the model output.

4. ROC-BASED MODELLING AND IMPLEMENTATION TOOLS

Let θ be the threshold value for the outlierness degree δ over a given scale: a data-point with $\delta > \theta$ is assigned to the data-subset of outliers; otherwise, data-points with $\delta \leq \theta$ are assigned to the data-subset of inliers.

The probability of erroneous assignments, either a false positive (an inlier misclassified as an outlier) or a false negative (a missed outlier) is related to the discriminating level of the chosen threshold θ.

The choice of the scale for δ values and of the value θ is a matter of decision-making, taking into account that they affect the distribution of some relevant variables, namely:

TP, true positives (the data-point is an outlier indeed and is correctly assigned to the outlier subset);

TN, true negatives (the data-point is an inlier indeed and is correctly assigned to the inlier subset);

FP, false positives (the data-point is not an outlier, but it is erroneously assigned to the outlier subset);

and *FN*, false negatives (the data-point is an outlier indeed, but it is erroneously assigned to the inlier subset).

The sensitivity (*Se*) and the specificity (*Sp*) of the classification process are defined in terms of false positive probability and false negative probability.

Putting $Tot^{+} = TP + FN$ and $Tot^{-} = TN + FP$, the sensitivity and the specificity can be expressed in terms of the following rates, respectively:

$$Se = TP/Tot^{+} \quad (1)$$

$$Sp = TN/Tot^{-} \quad (2)$$

Let o represent a binary variable such that $o = 1$ if the data-point is an outlier, otherwise $o = 0$; let r represent a bivalent—Yes/No (Y/N)—variable such that r = Y if the classification response is positive (for the data-point to be classified as an outlier), otherwise r = N. Let

$$\Pi(o = 1) \quad (3)$$

denote the pre-process (prior) probability for the data-point to be classified as an outlier, and let

$$\Pi(o = 0) = 1 - \Pi(o = 1) \quad (4)$$

denote the prior probability for an inlier.

Using this notation, *Se* and *Sp* can be reformulated in terms of conditional probability:

$$Se = \Pi(r = \mathrm{Y}|o = 1) \quad (5)$$

$$Sp = \Pi(r = \mathrm{N}|o = 0) \quad (6)$$

Moreover, to predict the performance in application that depends also on prior probabilities the positive predictive value, *PPV*, and the negative predictive value, *NPV*, are introduced in the following terms.

The positive predictive value is the inverse probability of *Se*:

$$PPV = Se\Pi(o = 1)/[Se\Pi(o = 1) + (1 - Sp)\Pi(o = 0)] \quad (7)$$

The negative predictive value is the inverse probability of *Sp*:

$$NPV = Sp\Pi(o = 0)/[Sp\Pi(o = 0) + (1 - Se)\Pi(o = 1)] \quad (8)$$

Other useful definitions are the true positives rate *TPR* and the false positives rate *FPR*; they are estimated by:

$$TPR = Se \quad (9)$$

$$FPR = 1 - Sp \quad (10)$$

In order to outline the proposed implementation rational, let the "3-sigma" criterion be considered in the following example.

Let the dataset $X = \{x_i: 0 < i < n\}$ be a sample of n data-points from an unknown univariate distribution with both the mean μ and the standard deviation σ to be estimated.

Let the outlierness degree of a data-point x_i be evaluated in terms of the distance $\delta_i = |x_i - \mu|$ and let the threshold be $\theta = \mu + 3\sigma$.

In this example, let the set X be composed of nine sampled points: $X = \{0, 1, 3, 3, 6, 8, 8, 300, 1000\}$. Even if some outliers are suspected to affect the sample dataset X, at least the maximum value $x_9 = 1000$, the situation is counterintuitive and can be analyzed as follows.

In fact, if the distribution parameters are estimated by the sample parameters, they are: $\mu = 147.6667$ and $\sigma = 334.279$. Then: $\delta_9 = 852.333$ and $\theta = 1150.501$.

In conclusion, this x_9 value is—according to the 3-sigma criterion—an inlier indeed, because the condition $\delta_9 < \theta$ is verified.

Such a situation might be used to illustrate that the 3-sigma criterion is defective in detecting outliers (see, e.g. [28]): in particular, this may happen when the dataset is not sampled from a normal distribution—as is the case for X in the example).

However this example is also useful to illustrate, from the point of view of the present approach, that the threshold—given a criterion—can be adapted in order to track signal variations that are a consequence of the time-varying system dynamics, instead of the presence of any outliers.

In a ROC-based approach, adaptation of thresholding translates into a false positive/false negative tradeoff allowed to vary in the ROC space, being constrained to lay on the locus of points described by a ROC curve. To exemplify, a ROC graph is shown in Fig. 3 [16] in the ROC space, namely *FPR* in abscissa and *TPR* in ordinate.

The ROC curve rises form the point (0, 0) (no false positive cases are only obtained at the cost of no true positives cases) and reaches the point (1, 1) (100% of true positive rate is obtained only at the cost of 100% false positive rate) with the varying slope $Se/(1 - Sp)$.

Therefore, the thresholding adaptability—namely, in the framework of applications such as those described in Section 3, the ability to track signal variations while monitoring for outliers—has a cost.

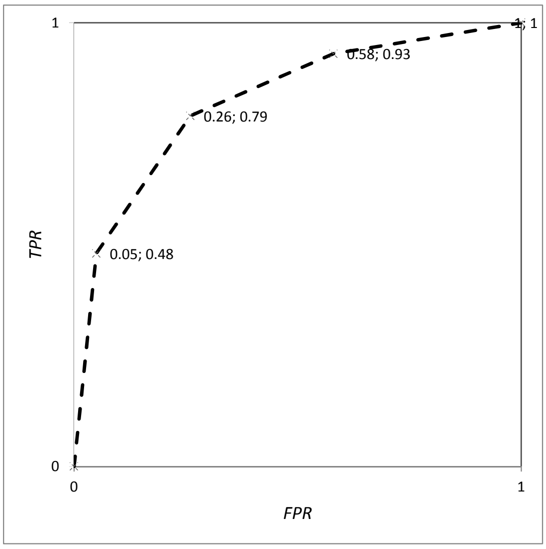

Fig. 3. ROC plot of a 4-level thresholding example (by use of calculator for ROC curves [31]).

The cost is that, moving along the ROC curve, the balance between sensitivity and specificity may imply further penalization; according to Eq. (7) and Eq. (8), the predictive values *PPV* and *NPV* are affected too.

More aspects of interest related to rating the comparative performance of soft classification processes are in terms of the area under the ROC curve [29], [30]. This area under the curve (AUC) is a global descriptor of the performance described by the ROC curve: the larger the AUC, the better the performance, with its limiting value AUC=1 representing an ideally errorless performance.

Table 1. A balanced sample set, with variable threshold levels.

Point no.	FP/TP class	True case	Threshold θ
1	FP	(–)	1
2	TP	(+)	0.95
3	TP	(+)	0.9
4	TP	(+)	0.85
5	TP	(+)	0.8
6	FP	(–)	0.75
7	TP	(+)	0.7
8	FP	(–)	0.65
9	FP	(–)	0.6
10	TP	(+)	0.55
11	FP	(–)	0.5
12	TP	(+)	0.45
13	FP	(–)	0.4
14	FP	(–)	0.35
15	FP	(–)	0.3
16	TP	(+)	0.25
17	TP	(+)	0.2
18	FP	(–)	0.15
19	TP	(+)	0.1
20	FP	(–)	0.005

As an example, let a sample set be composed of 20 cases as follows: 10 true positive cases (+) and 10 true negative cases (–). The set is tested and the resulting scores are represented by ordered points listed in Tab. 1, where each point is matched against the related threshold θ and its score is found to be greater than θ. A case is classified positive if its score is greater than the current threshold level, θ: if this classification corresponds to a (+), a TP is obtained; however a misclassification (FP) occurs if the case is instead a true negative (–) one.

For instance, the case represented by point no. 1 is (–) and is labelled FP in Tab. 1 because its score was found to be greater than the current threshold value, θ = 1; on the other hand, point n. 2 represents a (+) case and—given the current threshold value, that is now θ = 0.95—is labelled TP; and so on.

The plot in Fig. 4 is obtained step by step, taking into account the misclassification (FP) or correct classification (TP) of each single point, according to the descending order of threshold levels, as listed in Tab. 1. Thus, the plot in Fig. 4 is the graph of a step function.

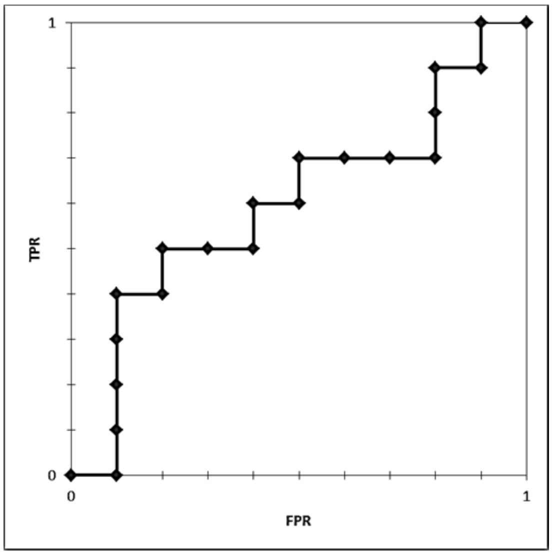

Fig. 4. Step ROC function plotted by use of data in Tab. 1.

A set with balanced distribution (i.e., with 10 true positive cases and 10 true negative cases) is used in this example to simplify the illustration of the construction of the ROC graph, shown in Fig. 4 (see also [6]). In order to cope with outliers, large changes in distributions are expected to occur—by definition of an outlier. However, it is worth noting that a ROC curve is insensitive to variations in distribution, being based upon rates (namely FPR and TPR), rather than upon absolute counts of cases. Furthermore, distribution skews of more than 10^2 are not unrealistic in real world applications and, for example—as pointed out by [6]—in industrial processes, variations in operating conditions may cause measurable variations in production quality.

The work is in progress. As to the implementation tools, diverse SW packages—also freely available on the internet—can be fitted to the purpose. For example, the graph in Fig. 3 was obtained (with application to another problem dealt with by the author of the present paper) by use of the web-based calculator for ROC curves, namely the JROCFIT calculator [31].

5. CONCLUSION

In the present paper, problems related to outlier detection and analysis in the area of dynamic systems is dealt with from a metrological point of view.

An approach is proposed on the base of ROC methodology that, after its origin in the framework of signal detection theory, is being applied in various fields, mainly for diagnostics, and imaging thresholding.

This methodology is focused in the present paper on adaptive thresholding aimed at classifying a suspected outlier in terms of an outlierness degree. Such a soft classification is used to adaptively balance false negative/false positive probabilities for in-process metrology applications.

An outcome of this study is fitting the ROC model for a new application purpose, namely processing outliers in dynamic measurements. The strategy is outlined in its preliminary architecture.

Perspective developments include simulations for performance analysis in terms of accuracy, robustness and promptness in process monitoring applications.

REFERENCES

[1] ASTM Standard E 178-08, 1961, 2002 (2008) "Standard practice for dealing with outlying observations", ASTM International, West Conshohocken, PA, USA, 2008.

[2] W. J. Dixon, "Analysis of extreme values", *Ann. Math. Statist.*, vol. 21, no. 4, pp. 488–506, 1950.

[3] F. Grubbs, "Procedures for detecting outlying observations in samples", *Technometrics*, vol. 11, no. 1, pp. 1–21, 1969.

[4] I. Ben-Gal, "Outlier detection". In: Maimon O. and Rockach L. (Eds.), *Data Mining and Knowledge Discovery Handbook*, Springer, 2010.

[5] C. E. Metz, "Receiver operating characteristic analysis: a tool for the quantitative evaluation of observer performance and imaging systems", *J Am Coll Radiol*, vol. 3, pp. 413–422, 2006.

[6] T. Fawcett, "An introduction to ROC analysis", *Pattern Recognition Letters*, vol. 27, pp. 861-874, 2006.

[7] G. E. D'Errico, "The problem of false discoveries from a metrological point of view", *J. Phys.: Conf. Ser.*, vol. 459**,** 012041 (6 pages), 2013.

[8] G. E. D'Errico, "Multiple hypothesis testing for metrology applications", Accred Qual Assur, vol. 19, pp. 1–10, 2014.

[9] G. Brumfiel, "Significant", *Nature*, vol. 455, pp. 1027–1028, 2008.

[10] J. P. A. Ioannidis, "Why most published research findings are false", *PLoS Medicine*, vol. 2, no. 8-e124, pp. 696–701, 2005.

[11] G. E. D'Errico, "Aggregation of comparisons data and reversal phenomena of metrological interest", *Measurement*, vol. 50, pp. 319–323, April 2014.

[12] G. E. D'Errico and N. Murru, "Fuzzy treatment of candidate outliers in measurements", *Advances in Fuzzy Systems*, vol. 2012, Article ID 783843, 6 pages, 2012.

[13] C. C. Aggarwal, *Outlier Analysis*, chap. 1, Springer, N.Y., USA, 2013.

[14] G. E. D'Errico, "Testing for outliers based on Bayes rule", *Proc. IMEKO XIX World Congress*, pp. 2368–2370, Lisbon, Portugal, Sept. 6–11, 2009.

[15] G. E. D'Errico, "Receiver operating characteristic: a tool for cell confluence estimation", *Proc. IEEE International Symposium on Medical Measurements and Applications (MEMEA 2015)*, pp. 576–579, Torino, Italy, May 7–9, 2015.

[16] G. E. D'Errico, "Adaptive cell detection by application of roc methodology", *XXI IMEKO World Congress*, Prague, Czech Republic, Aug. 30–Sept. 4, 2015.

[17] C. E. Metz, "ROC analysis in medical imaging: a tutorial review of the literature", *Radiol Phys Technol*, vol. 1, pp. 2–12, 2008.

[18] G. E. D'Errico, "À la Kalman filtering for metrology tool with application to coordinate measuring machines", *IEEE Trans. Ind. Electron*., vol. 59, no. 11, pp. 4377–4382, Nov 2012.

[19] G. E. D'Errico and N. Murru, "An algorithm for concurrent estimation of time-varying quantities", *Meas. Sci. Technol.*, vol. 23, no. 4, 045008 (9 pages), April 2012.

[20] G. E. D'Errico and N. Murru, "In-process estimation of time-variant contingently correlated measurands", *Int. J. Metrol. Qual. Eng.* Vol. 3, pp. 137–143, 2012.

[21] G. E. D'Errico, "A systems theory approach to modelling of cutting temperature with experimental identification", *Int. J. Mach. Tools Manufact.*, vol. 37, no. 2, pp. 149–158, 1997.

[22] K. J. Åström and B. Wittenmark, *Adaptive Control* (2nd edition), Addison–Wesley, Reading, MA, USA, 1995.

[23] R. Bellman, *Adaptive Control Processes. A Guided Tour*, chaps. IX, XV, Princeton University Press, Princeton, NJ, USA, 1961.

[24] G. E. D'Errico, "Adaptive systems for machining process monitoring and control", *J. Materials Processing Technology*, vol. 64, pp. 75-92, 1997.

[25] J. M. Lemos, "Adaptive control in the presence of outliers", *Proc. 14th Mediterranean Conference on Control and Automation*, Ancona (I), 28–30 June, 2006.

[26] B. Abraham and A. Chuang, "outlier detection and time series modeling", *Technometrics*, vol. 31, no. 2, pp. 241–248, 1998.

[27] R. K. Pearson, "outliers in process modeling and identification", *IEEE Trans. Contr. Syst. Technol*., vol. 10, no. 1, pp. 55–63, 2002.

[28] C. Leys, C. Ley, O. Klein, P. Bernard and L. Licata, "Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median", *Journal of Experimental Social Psychology*, vol. 49, pp. 764–766, 2013.

[29] D. Bamber, "The area above the ordinal dominance graph and the area below the receiver operating characteristic graph", *Journal of Mathematical Psychology*, vol. 12, pp. 387–415, 1975.

[30] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve", *Radiology*, vol. 1, pp. 29–36, 1982.

[31] J. Eng, "ROC analysis: web-based calculator for ROC curves". Baltimore: Johns Hopkins University (updated 2014 March 19; cited July 2015). Available from: http://www.jrocfit.org.