# Between a ROC and a Hard Place: Using prevalence plots to understand the likely real world performance of biomarkers in the clinic.

Lendrem BC [1,3,4], Lendrem DW [1,2], Pratt AG [1,2,3], Naamane N [1,2], McMeekin P [5,6], Ng WF [1,2,3], Allen J [1,4], Power M [1,3,4 *], Isaacs JD [1,2,3 *]

[1] Institute of Cellular Medicine, Newcastle University, Newcastle upon Tyne, UK
[2] NIHR Newcastle Biomedical Research Centre, Newcastle University, Newcastle upon Tyne, UK
[3] Newcastle upon Tyne Hospitals NHS Trust, Newcastle upon Tyne, UK
[4] NIHR Newcastle In Vitro Diagnostics Co-operative, Newcastle University, Newcastle upon Tyne, UK
[5] Institute of Health & Society, Newcastle University, Newcastle upon Tyne, UK
[6] School of Health Sciences, Northumbria University, Newcastle upon Tyne, UK

[*]Corresponding Authors:

John D Isaacs, Institute of Cellular Medicine, Newcastle University, Framlington Place NE2 4HH, UK Email: j.d.isaacs@newcastle.ac.uk
Michael Power, NIHR Newcastle In Vitro Diagnostics Co-operative, Newcastle University, Framlington Place NE2 4HH, UK Email: michael.power@newcastle.ac.uk

The Receiver Operating Characteristic (ROC) curve is a widely used tool to evaluate diagnostic and prognostic biomarker performance[1,2,3]. The ROC curve compares the sensitivity and specificity of a candidate biomarker for a range of potential cut-off values for a biomarker assay – see Figure 1A. One of the perceived advantages of the ROC curve is that it is independent of the prevalence of the disease and captures the two key misclassification errors – false positive errors and false negative errors – as a function of biomarker cut-offs.

However, when teaching, it is important to emphasize that while the ROC curve is *independent* of the prevalence rate, the translational performance of a biomarker test in the clinic is *critically* dependent upon that very same prevalence rate[4,5]. While this is well understood by statisticians, it is not always obvious to scientists and clinicians developing new assays.

For example, the "10-90-50 Rule" states that:

- for a disease with a prevalence of **10%**, and
- an assay with both sensitivity and specificity greater than **90%**,
- **50%** of patients testing positive are false alarms.

And if the prevalence of the disease is less than 10% then *most of our positive diagnostic tests will be false alarms* – see Figure 1B.

In this short note, we present a simple tool permitting practitioners to capture assay performance as a function of prevalence rates. Understanding how an assay performs across a range of values for the prevalence is often critical – both commercially and clinically. There is often uncertainty surrounding the estimate of prevalence in the first place. Then, once the test is moved into the clinic, this is compounded by the fact that the prevalence rates vary depending upon how the patients are selected for testing. And, even following adoption of the test, the test may be used for groups of patients for whom the prevalence is rather less than that in the original test population, making the test virtually worthless. Translational performance is a function of both the 'true' disease prevalence *and the clinical selection process for testing*[4,5].

Rather than ignore prevalence, simple plots of candidate assay performance as a function of prevalence rate give a more realistic understanding of the likely real-world performance in the clinic, and a greater understanding of the likely impact of variation in that prevalence on translational performance in the clinic – see Figure 1B. Plotting the misclassification rates – *False Alarms* and *Missed Diagnoses* – as a function of possible prevalence rates allows us to focus on misclassification costs.

In Figure 2, we give a worked example showing prevalence plots for the promising mast cell activation test for IgE-mediated food allergy[6]. The sensitivity and specificity of this test are an impressive 97% and 92% respectively, with a ROC AUC of 0.99 (95% CI: 0.96, 1.00). While the number of patients with food allergies who are missed by the assay is reassuringly low, the number of patients without the disease testing positive is likely to be high, given an estimated prevalence in the UK of just 6%. This may, or may not, be acceptable. In real life, the relative costs associated with false alarms and missed diagnoses are likely to be very different and must be assessed prior to the test entering the clinic: a false alarm may simply mean a patient is subjected to further testing; a missed diagnosis may mean the patient dies.

Prevalence plots focus reviewers on misclassification rates, misclassification costs, and how the assay will translate to the clinic. Without thoughtful consideration of prevalence rates and the relative costs of misclassification errors, it is easy to 1) overstate the potential value of a candidate

biomarker, 2) generate unrealistic expectations of that candidate, 3) incur unnecessary trial costs in evaluating that candidate, 4) incur opportunity costs in denying patients access to better diagnostic tests.

We provide an Excel workbook as a teaching tool. This permits readers to estimate Missed Cases and False Alarms for an assay with any given sensitivity, specificity for a range of prevalence values.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available at

**arXiv:1810.10794 [stat.AP]** https://arxiv.org/abs/1810.10794


## ACKNOWLEDGEMENTS

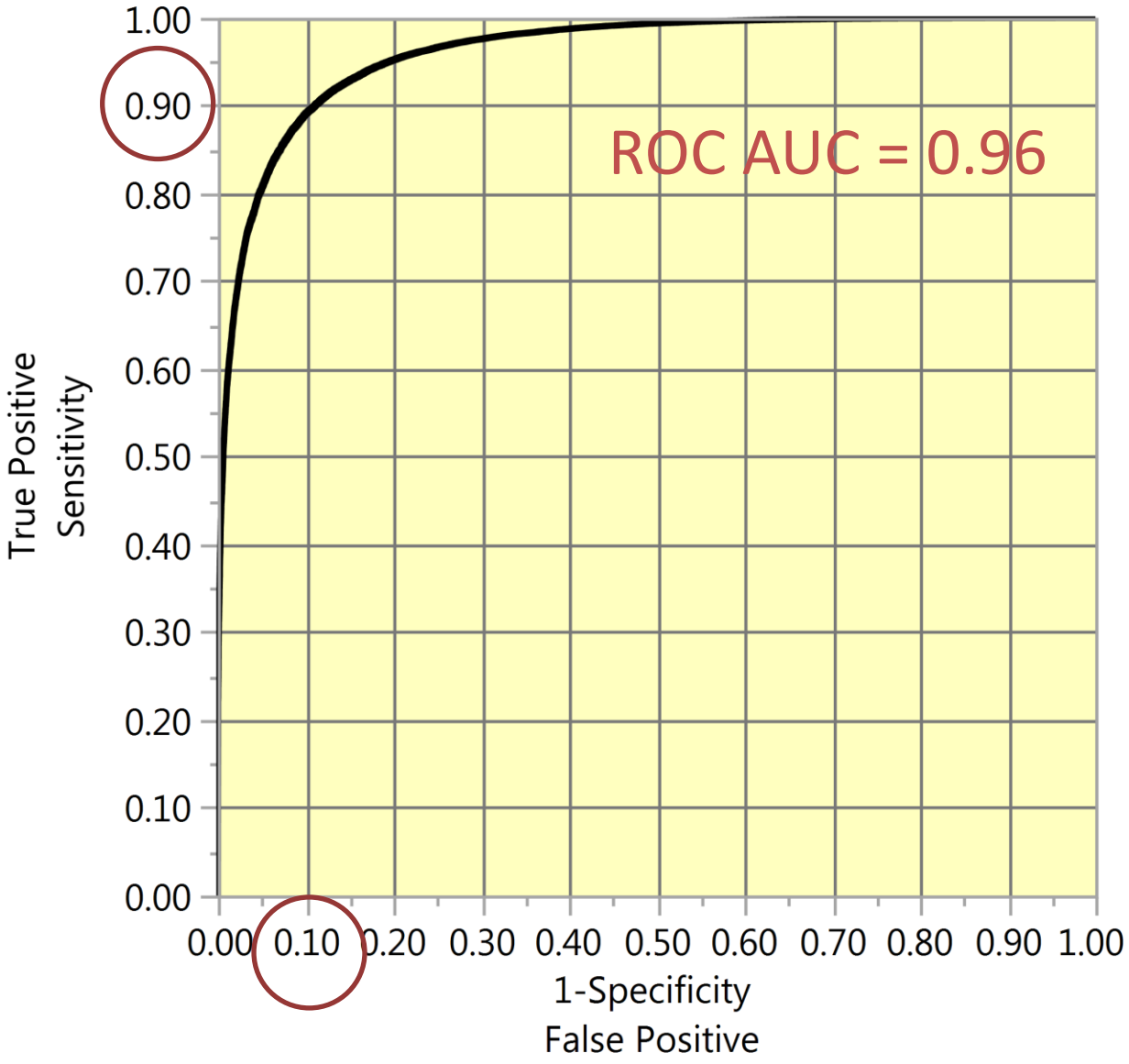
The authors were supported at least in part by the NIHR Newcastle Biomedical Research Centre and the NIHR In Vitro Diagnostics Cooperative at Newcastle University and the Newcastle upon Tyne Hospitals NHS Trust during the research and/or preparation of the article. The sponsors played no part in the study design; collection, analysis and interpretation of data; the writing of the report; or the decision to submit the article for publication.

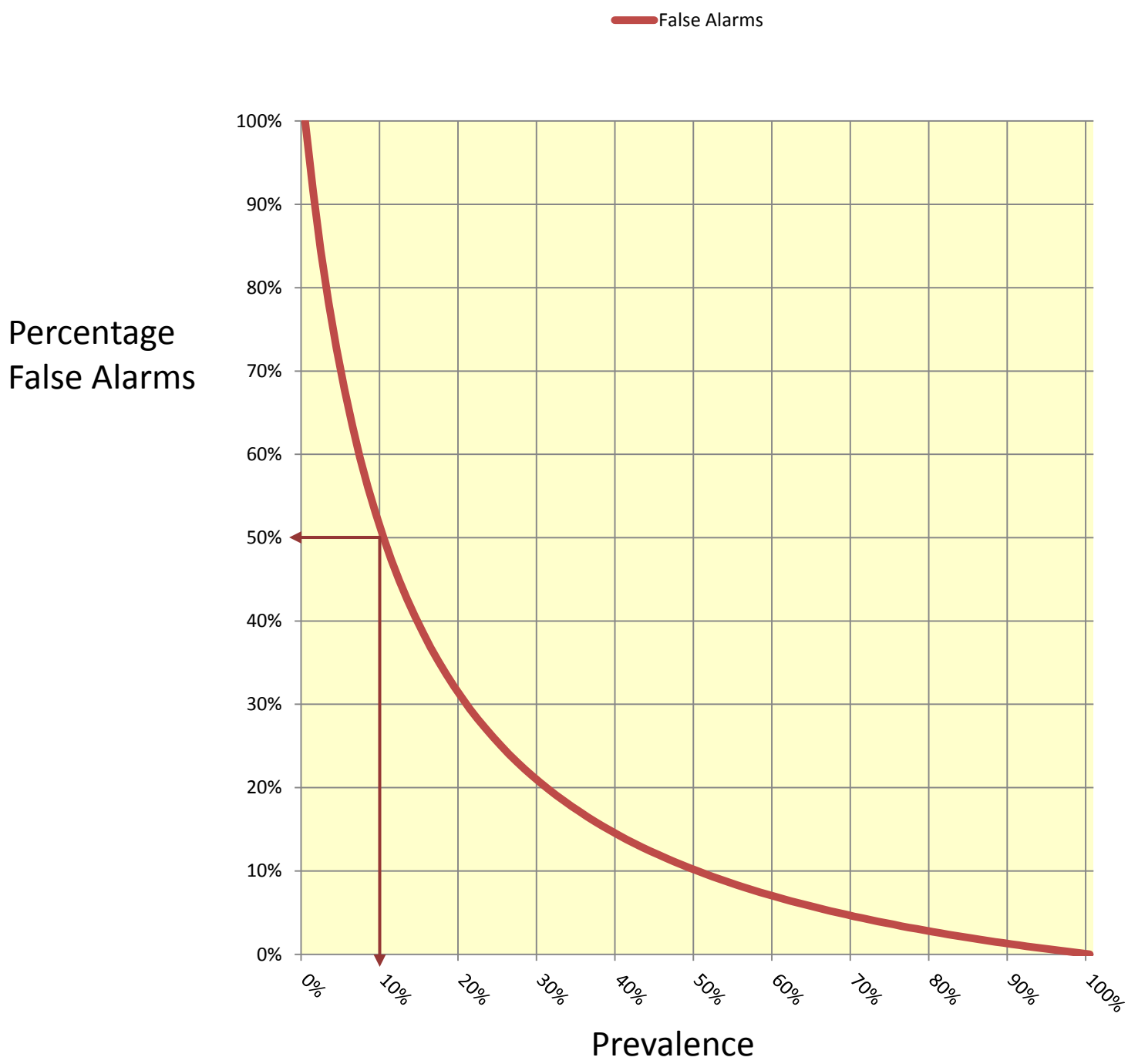
**Figure 1: The 10-90-50 Rule.** While the ROC AUC for our candidate biomarker looks promising (**A**) with a sensitivity and specificity of 90%, the performance of the assay in the clinic depends critically on the prevalence of the disease (**B**). The false positive and false negative rates are both 10%, but if the prevalence of the disease in the test population is 10% then 50% of all positive tests will be false alarms. The false alarm rate depends critically upon the prevalence in the test population. Plotting test performance as a function of prevalence gives a more realistic understanding of likely performance in the clinic. See text for details.

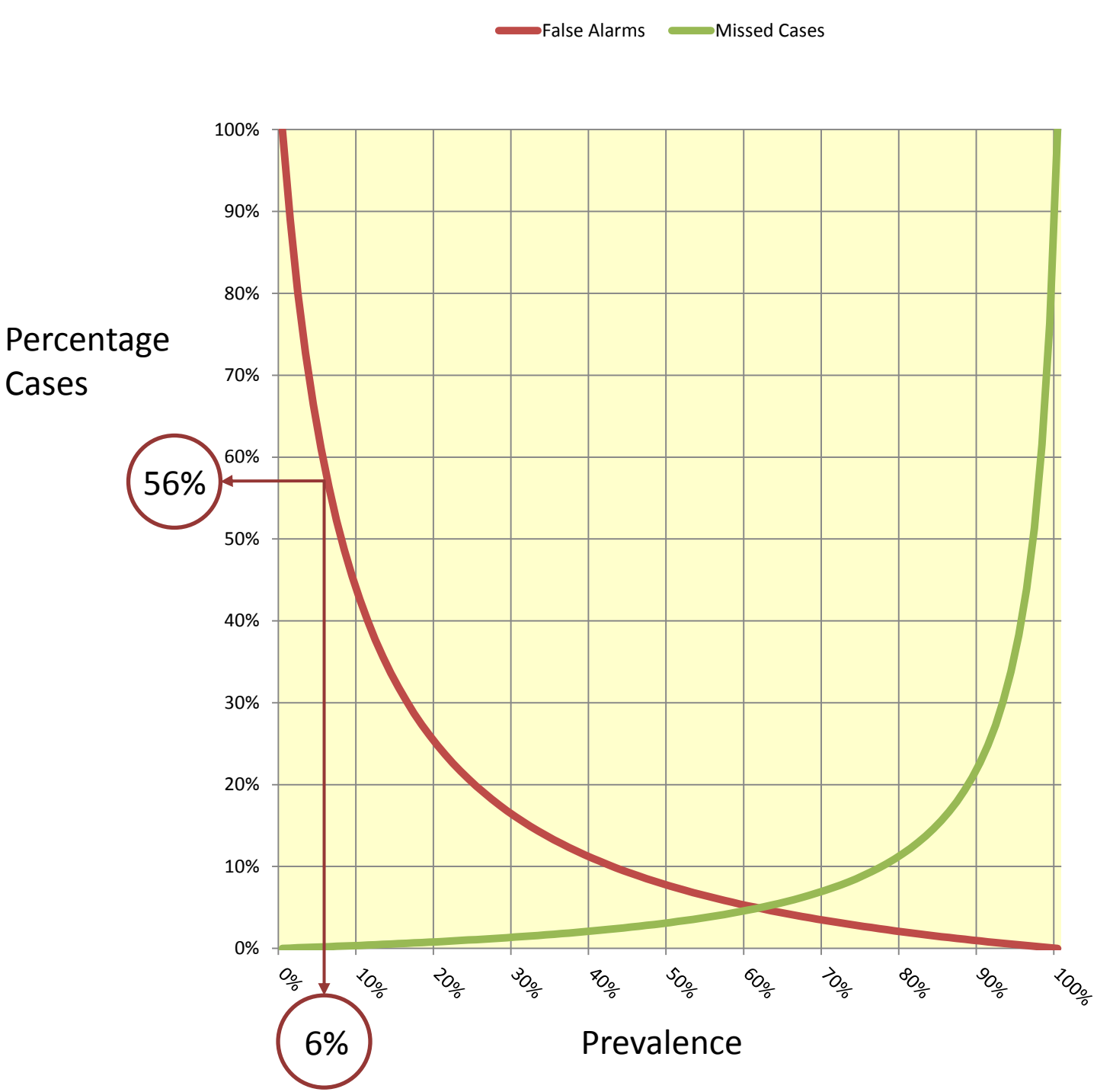
**Figure 2: Prevalence plots for the mast cell activation assay.** While the assay looks promising - with a sensitivity of 97%, specificity of 92% and a ROC AUC of 0.99 - translation to the clinic depends critically upon the prevalence in the test population. As the prevalence increases, the percentage of missed cases increases and the percentage of false alarms decreases. If the prevalence rate is zero then any positive test results are false positives and the false alarm rate is 100%. If the prevalence rate is 100% then any negative tests are false negatives and the missed case rate is 100%. *The vertical line shows the estimated prevalence of IgE-mediated food allergy at 6%.* At this prevalence rate, approximately 56% of all positive tests will be false alarms – see Supplementary Excel Workbook **Bench-2-Bedside.xls**.

**REFERENCES**


1. Bossuyt, P.M. (2010) 'Clinical validity: defining biomarker performance', *Scand J Clin Lab Invest Suppl*, 242, 46-52.
2. Florkowski, C.M. (2008) 'Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests', *Clin Biochem Rev*, 29 Suppl 1, S83-7.
3. Saito, T. and Rehmsmeier, M. (2015) 'The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets', *PLoS One*, 10(3), e0118432.
4. Power, M., Fell, G. and Wright, M. (2013) 'Principles for high-quality, high-value testing', *Evid Based Med*, 18(1), 5-10.
5. Fanshawe, T.R., Power, M., Graziadio, S., Ordonez-Mena, J.M., Simpson, J. and Allen, J. (2018) 'Interactive visualisation for interpreting diagnostic test accuracy study results', *BMJ Evid Based Med*, 23(1), 13-16.
6. Bahri, R., Custovic, A., Korosec, P., Tsoumani, M., Barron, M., Wu, J., Sayers, R., Weimann, A., Ruiz-Garcia, M., Patel, N., Robb, A., Shamji, M.H., Fontanella, S., Silar, M., Mills, E.N.C., Simpson, A., Turner, P.J. and Bulfone-Paus, S. (2018) 'Mast cell activation test in the diagnosis of allergic disease and anaphylaxis', *J Allergy Clin Immunol*.


ROC AUC = 0.96
True Positive
Sensitivity
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
1-Specificity
False Positive


False Alarms
Percentage
False Alarms
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Prevalence


False Alarms
Missed Cases
Percentage Cases
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
56%
6%
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Prevalence