

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

Lendrem BC<sup>1,3,4</sup>, Lendrem DW<sup>1,2</sup>, Pratt AG<sup>1,2,3</sup>, Naamane N<sup>1,2</sup>, McMeekin P<sup>5,6</sup>, Ng WF<sup>1,2,3</sup>, Allen J<sup>1,4</sup>, Power M<sup>1,3,4\*</sup>, Isaacs JD<sup>1,2,3\*</sup>

<sup>1</sup> Institute of Cellular Medicine, Newcastle University, Newcastle upon Tyne, UK

<sup>2</sup> NIHR Newcastle Biomedical Research Centre, Newcastle University, Newcastle upon Tyne, UK

<sup>3</sup> Newcastle upon Tyne Hospitals NHS Trust, Newcastle upon Tyne, UK

<sup>4</sup> NIHR Newcastle In Vitro Diagnostics Co-operative, Newcastle University, Newcastle upon Tyne, UK

<sup>5</sup> Institute of Health & Society, Newcastle University, Newcastle upon Tyne, UK

<sup>6</sup> 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](mailto: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](mailto:michael.power@newcastle.ac.uk)

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

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

11 For example, the “10-90-50 Rule” states that:

- 12 • for a disease with a prevalence of **10%**, and
- 13 • an assay with both sensitivity and specificity greater than **90%**,
- 14 • **50%** of patients testing positive are false alarms.

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

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

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

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

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

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

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

49

50

#### 51 DATA AVAILABILITY STATEMENT

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

53 arXiv:1810.10794 [stat.AP] <https://arxiv.org/abs/1810.10794>

54

#### 55 ACKNOWLEDGEMENTS

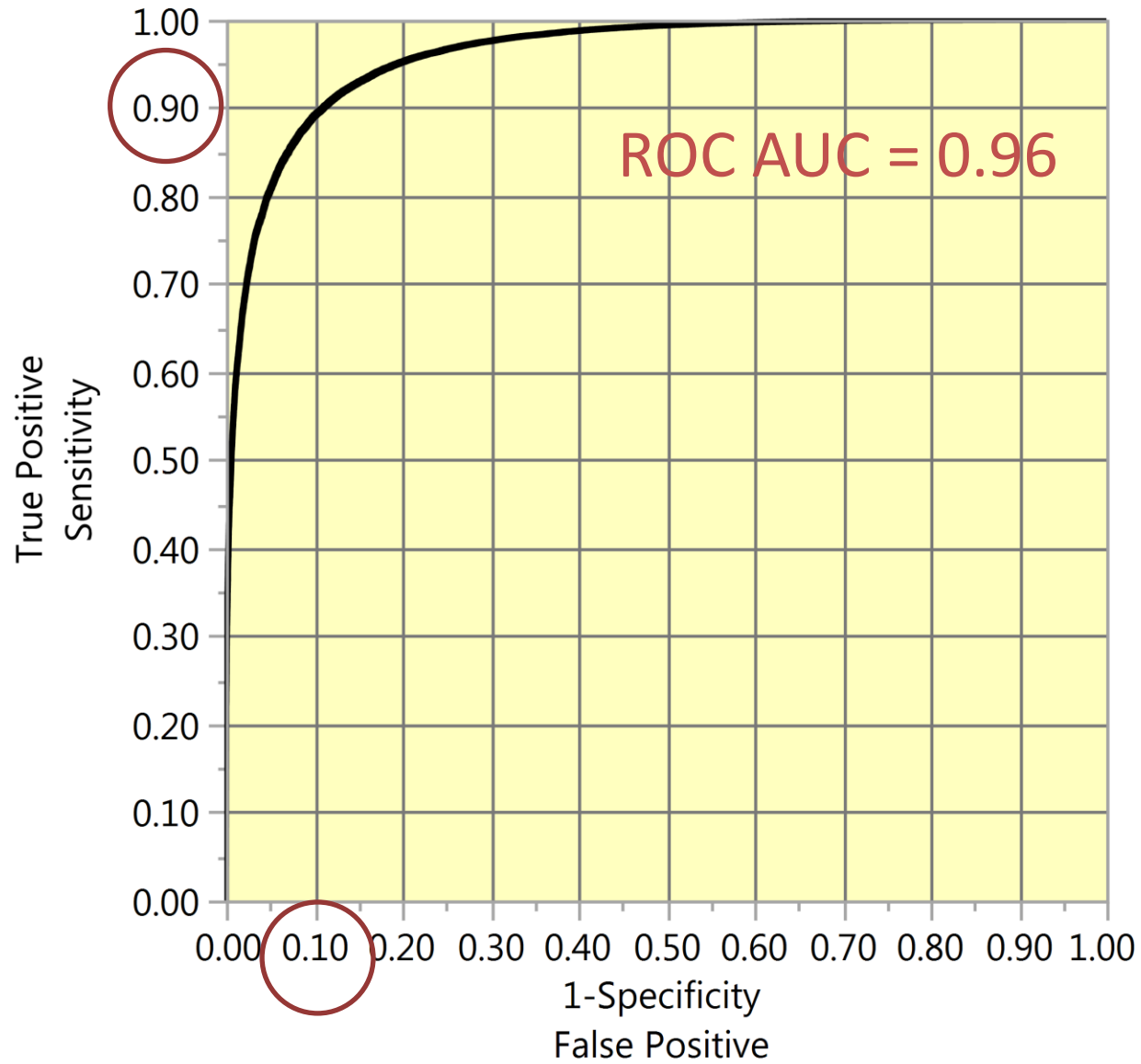
56 The authors were supported at least in part by the NIHR Newcastle Biomedical Research Centre and  
57 the NIHR In Vitro Diagnostics Cooperative at Newcastle University and the Newcastle upon Tyne  
58 Hospitals NHS Trust during the research and/or preparation of the article. The sponsors played no  
59 part in the study design; collection, analysis and interpretation of data; the writing of the report; or  
60 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**.

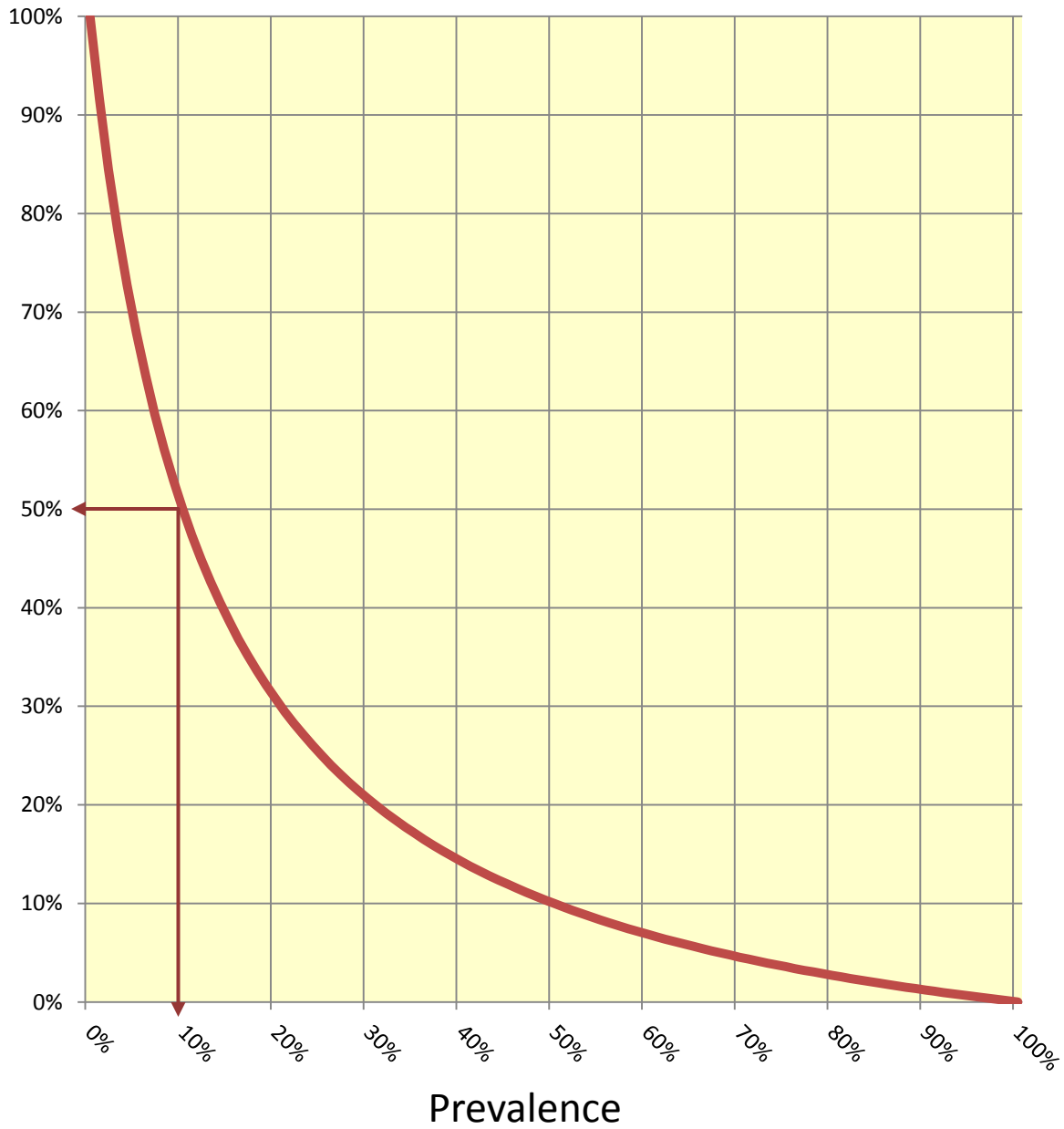
## REFERENCES

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*.



False Alarms

Percentage  
False Alarms



False Alarms Missed Cases

Percentage Cases

