This report summarizes the paper “Limitations of the Odds Ratio in Gauging the Performance of a Diagnostic, Prognostic, or Screening Marker” by Pepe et al, 2003. The report also includes my perspective based on literature review and the course material for “Risk prediction and Precision Medicine”. A glossary for special terminologies* can be found at the end of the report for reference.

A biomarker is defined as a measurable indicator of a biological outcome. A valid biomarker should be quantifiable and be able to discriminate between cases and controls[1]. Biomarkers can be categorized into four categories depending on the stage of the disease, these include diagnostic, prognostic, predictive, and therapeutic. A diagnostic biomarker is used to detect the presence of a disease. A predictive biomarker helps predict an individual's trajectory to a certain therapy. A prognostic marker helps define the clinical course and likely outcome of a disease. A therapeutic marker is a biological target such as a protein receptor or DNA that can be leveraged to treat a disease[2,3].

In general, there are three developmental phases of a biomarker[4,5]:

1. Discovery phase: This phase quantifies the association between the marker and the biological outcome in terms of odds ratios and relative risks and other measures of association.
2. Performance evaluation: This phase evaluates the performance of a biomarker for classifying cases and controls by assessing true positive fraction (TPF) or sensitivity, and false-positive fraction [FPF] or 1-specificity and resulting receiver operating characteristic (ROC) curves.
3. Incremental Value: This phase quantifies the improvement in the performance of an existing well calibrated model after addition of a newly discovered marker.

As evident, each phase has its own methodological considerations including the study design and statistical measures. The paper by Pepe et al specifically emphasizes the point that traditional measures of association such as OR or RR cannot capture the discriminatory performance of a marker even when OR or RR show a very strong association [6-8]. Additionally, it is important to determine if adding a new marker to an existing marker/set of markers will improve its performance (incremental value). Determining the incremental value also relies on a separate statistical measure and traditional epidemiologic methods can lead to misleading results.

For clarity the rest of the paper discusses the following points.

1. Why measures of association cannot be used to discriminate between cases and controls? and why markers from case control studies cannot predict future risk of a disease?
2. How can we evaluate performance of a marker in discriminating between cases and controls and how to assess effect modification?
3. Incremental value of a marker.

Why measures of association cannot be used to discriminate between cases and controls? and why markers from case control studies cannot predict future risk of a disease?

Unfortunately, OR or RR are used increasingly commonly in predictive marker studies. For example, a case control study by Zhang et al showed a significant association between an inflammatory marker* MPO (Myeloperoxidase) and coronary artery disease *(CAD), showing that patients with high MPO levels have 11.9 higher odds of developing CAD than people with lower MPO (OR of 11.9 (95% CI, 5.5-25.5). These findings have biological plausibility since MPO through a molecular pathway causes atherosclerosis*. Although the findings are commendable the authors have proposed a potential role for MPO as an inflammatory marker in CAD and risk assessment[9]. However, an OR, cannot gauge the discriminatory power of the marker to classify cases and controls. In order to summarize the classifying accuracy of a marker one must consider misclassification errors and report TPF i.e. P[marker +ive | outcome +ive] and FPF i.e. P[marker +ive |outcome -ive]. A perfect marker has TPF of 1 and FPF of 0. Moreover, the utility of a marker also depends on the context. Let's say for a disease which entails an aggressive treatment and is emotionally and financially distressing such as cancer, can we really afford to mislabel 2% of the population as false positive? on the other hand for a disease such as antiphospholipid syndrome* which has a seemingly innocuous and inexpensive treatment like a baby aspirin for clot prevention, a higher false positive rate can be acceptable

For a binary marker OR = {TPF/(1 – TPF)}x{(1 – FPF)/FPF}[6] [10]and the following figure shows receiver operator characteristic curves(ROCs) at various OR values.

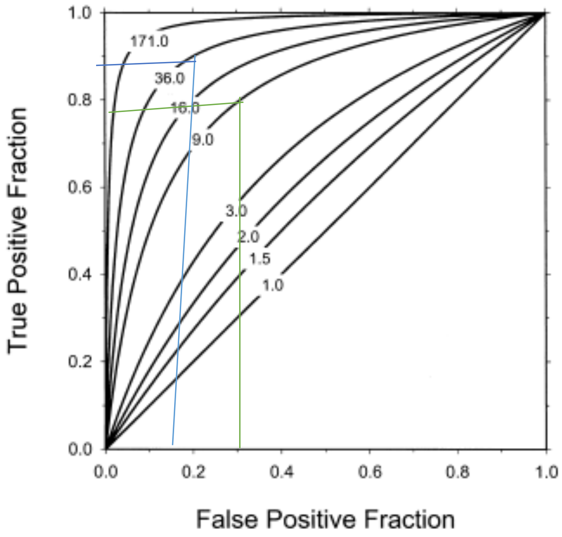


Fig 1. Limitations of the Odds Ratio in Gauging the Performance of a Diagnostic, Prognostic, or Screening Marker by Pepe et al, 2003

As evident an OR as high as 9 correctly labels 75% of cases as having positive test, while mislabels 30 % controls as false positive (green). For a valid marker the ROC should lie high above the diagonal line. Considering the ROC function for OR=9 it can be seen that the marker performance is somewhat weak at this very high OR. Furthermore, at a higher accuracy when FPF = 0.15 and TPF =0.85 (blue), the OR = 36, a magnitude of association seldom seen in epidemiological studies. Even if such a large OR is reported it can be seen that it spans across various combinations of TPF and FPF, for example for a large TPF of 97.3 the FPF is well over 50%

In case of a continuous marker. The size of the odds ratio depends on the units in which a marker is measured. In the following figure, the marker is scaled so that a unit increase represents two standard deviations = one unit. The extent of separation between the two curves depicts how well a marker performs in classifying cases and controls. It can be seen that even a high OR such as 9 has a considerable overlap and a larger separation would require an OR of over 300.

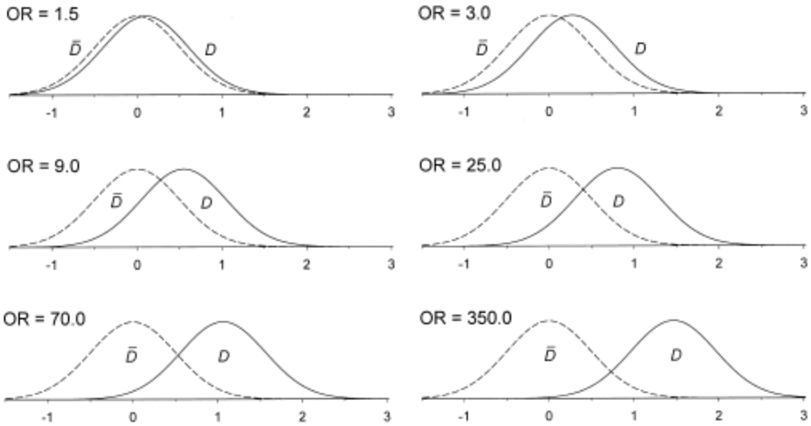


Fig 2. Limitations of the Odds Ratio in Gauging the Performance of a Diagnostic,Prognostic, or Screening Marker by Pepe et al, 2003

This also holds true even when marker performance is evaluated at different cutoff points for each quartile in the distribution for controls.

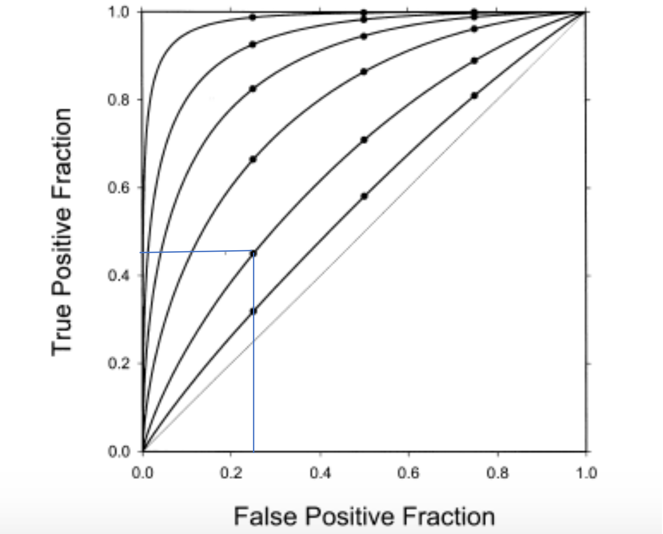


Fig 3. Limitations of the Odds Ratio in Gauging the Performance of a Diagnostic, Prognostic, or Screening Marker by Pepe et al, 2003

Note that when the OR for the upper quartile vs. the lowest quartile is 4.1 for the ROC function at OR=3, 45% of cases are correctly identified and 25% percent of controls are misclassified as having the disease.

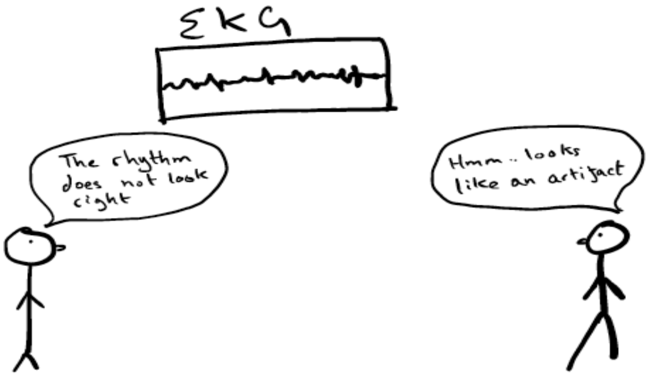
Additionally, for predictive markers, a marker from case controls studies can only provide insight into the risk of a disease compared to someone without the that marker. A marker in itself does not predict the future timeframe for the course of the disease unless an accurate baseline risk is known. This baseline risk, also known as the nuisance parameter (α) from case control studies is an inaccurate parameter of the baseline risk. On the other hand, absolute risk model which takes calibrated baseline risk from cohort studies into consideration can predict the probability of a disease occurring over a particular period. This holds true without comparison to another group.

In summary OR is a scaler value that quantifies an association and cannot be utilized to classify cases and controls. Additionally, extremely large associations are needed for a marker to qualify as a valid marker, and even then, there exist several combinations of TPFs and FPFs at a certain OR value. Moreover, the discriminatory ability of a marker does not inform the future risk of a disease and absolute risk model are needed to predict future risk of a disease. This brings to our next question about how can we assess the performance of a marker?

How can we evaluate the performance of a marker in discriminating between cases and controls? and how to assess effect modification?

As mentioned in the previous section the performance of a marker is a function of two parameters TPF and FPF. For a binary marker the TPF and FPF can provide an insight into the performance of a marker, and OR can be written as OR = {TPF/(1 – TPF)} X {(1 – FPF)/FPF}. For continuous markers ROC function is utilized and can provide a set of (FPF, TPF) depending on the criteria determined for the marker. Moreover, unlike OR, the ROC function does not depend on the measurement units of a marker and is not interpreted as per unit change of a marker e.g., a marker measuring a serum concentration of a biological molecule can be easily compared with another marker measuring a physical measure such as size of a cancerous tumor.

Effect modification: While the marker in itself may perform well in classifying cases and controls, other factors can influence marker performance such as fluid in the lungs that can obscure MRI findings for lung pathologies or clinician dependent EKG interpretation of an arrythmia*.



Clinician 1 Clinician 2

It is important to know how marker performance varies according to these factors so that they may be accounted for during interpretation. This type of effect modification requires other statistical tests rather than the traditional epidemiological one. Traditionally logistic regression utilizes an interaction term between the marker and the effect modifier (Z) to assess whether the magnitude of association (OR, RR) varies with the effect modifier.

$$\mathrm{logit}P[D = 1|X, Z] = \alpha + \beta_1 X + \beta_2 Z + \beta_3 XZ.]$$

But since it is already determined that OR is a scaler value and does not provide information about TPF and FPF, other ways have to be sought to determine effect modification. For a binary marker and when Z is binary, this can be achieved by evaluating the extent to which TPF and FPF vary with in the presence and absence of Z. Pearson chi square statistics can be used to test the hypothesis that H0: FPF(Z = 1) = FPF(Z = 0), or two sided alternative hypothesis where FPF(Z = 1) and FPF(Z = 0). Similarly, when Z is continuous logistic regression can be used to evaluate how TPF in cases vary with Z i.e. TPF(Z) = P(X = 1|D = 1, Z)[11 12].

For a continuous marker, we can determine if the ROC function for a marker significantly differs with Z (If Z is binary) by calculating a P value for the difference between the AUC in the presence and absence of Z.

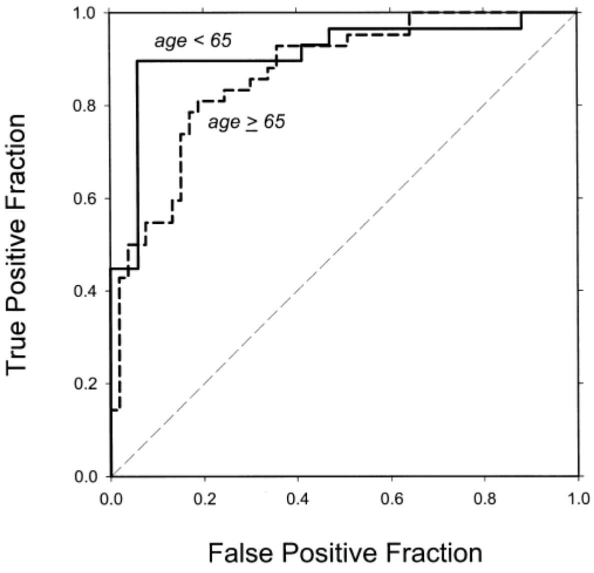


Fig 6. Limitations of the Odds Ratio in Gauging the Performance of a Diagnostic, Prognostic, or Screening Marker by Pepe et al, 2003

The above plot shows that the AUC does not significantly differ (p=0.44) for prostate specific antigen* (PSA) between patients older and younger than 65, hence age is not an effect modifier of PSA. We can employ a similar technique to compare the performance of two different marker who serve the same purpose in the same population. When a marker is continuous, and Z is also continuous a regression technique for ROC curves must be employed.

Incremental Value:

One important step in the development of a marker is how an addition of marker will contribute to the performance of an existing well calibrated model consisting of clinical risk factors or previous markers. Using traditional epidemiological model one can assess the additional value of a marker by using a multivariable model to see how OR changes for an association between the new marker (X2) and the outcome D=1 after adjusting for an existing marker (X1).

$$\operatorname{logit} P(D = 1|X_1, X_2) = \alpha + \beta_1 X_1 + \beta_2 X_2,$$

However, since a measure of association cannot quantify the predictive performance of a marker, it cannot serve as a reasonable approach to assess the incremental value of a new marker. In this case its performance can be evaluated by how the ROC function improves after adding the new marker to an existing marker in use. To solidify this concept, lets evaluate the improvement in predictive performance before and after addition of a newer pancreatic cancer marker (CA-125) to an existing pancreatic cancer marker (CA-19-9).

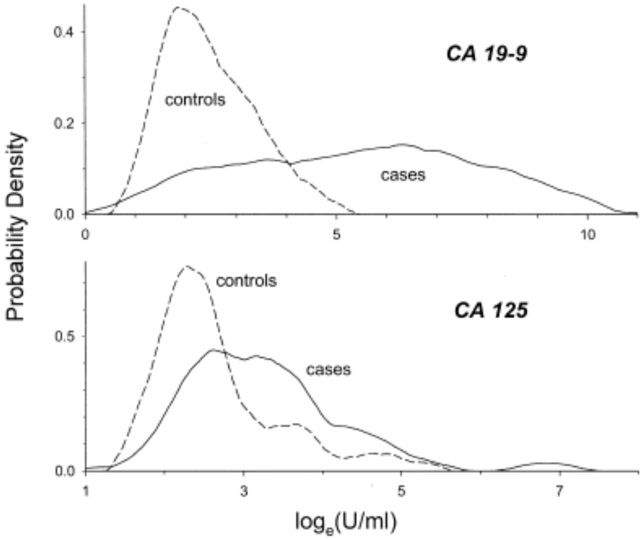


FIGURE 4. Frequency distributions of two markers for pancreatic cancer. Refer to Wieand et al. (20) for a description of source data.

Fig 4. Limitations of the Odds Ratio in Gauging the Performance of a Diagnostic,Prognostic, or Screening Marker by Pepe et al, 2003

Using the logistic regression model and the above figure, there is a statistically significant association (OR= 2.54 (p = 0.002), between CA 125 and pancreatic cancer after adjusting for CA-19-9. However, this does not hold true when improvement in ROC function is assessed after looking at the combined effect of CA-125 and CA-19-9, as negligible improvement in the AUC is seen after the addition of CA-125.

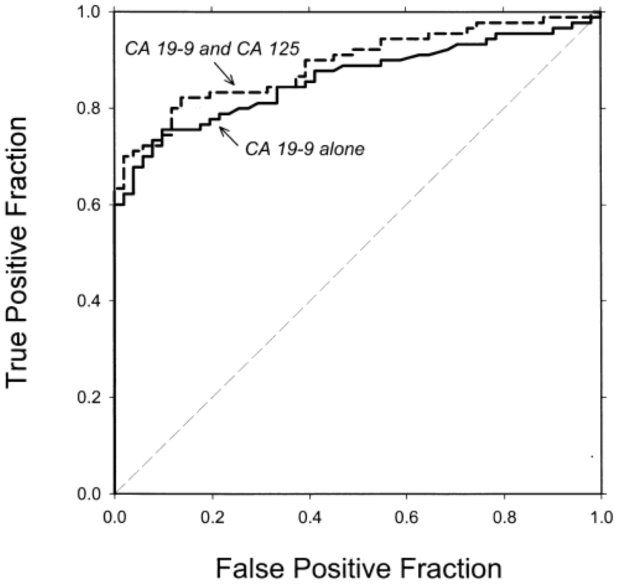


Fig 7. Limitations of the Odds Ratio in Gauging the Performance of a Diagnostic, Prognostic, or Screening Marker by Pepe et al, 2003

Given the context of the disease, if a 5 % FPF is acceptable then adding CA-124 only raises TPF by 3%, raising question about the clinical utility of adding a cost prohibitive marker that offers only a minimum improvement in AUC.

In conclusion, this report summarizes how commonly used association measures do not hold the intrinsic property of assessing marker performance at different stages of development, which require different study designs and statistical testings based on ROC curves, marker and covariate characteristics (binary or continuous). The paper also highlights that the discriminatory ability of a marker from a case controls study does not inform the future risk of a disease and absolute risk models are needed to predict the future risk of a disease.

Glossary of terms:

Inflammatory marker MPO (Myeloperoxidase): Molecule secreted in the body in response to an injury

Coronary artery disease: Blockage of the vessels of the heart

Atherosclerosis: Fat deposition in the vessels that can cause blockage

Antiphospholipid syndrome: A clotting disorder, first line treatment is aspirin

Arrythmia: Irregular rhythm of the heartbeat.

Prostate specific antigen: Marker for Prostate cancer.

References:

1. Huang Y, Pepe MS. Biomarker evaluation and comparison using the controls as a reference population. Biostatistics 2009;10:228-44.
2. Durães C, Almeida GM, Seruca R, Oliveira C, Carneiro F. Biomarkers for gastric cancer: prognostic, predictive or targets of therapy? Virchows Arch 2014;464:367-78.
3. Lin L-L, Huang H-C, Juan H-F. Discovery of biomarkers for gastric cancer: A proteomics approach. Journal of Proteomics 2012;75:3081-97.
4. Pepe MS, Etzioni R, Feng Z, et al. Phases of Biomarker Development for Early Detection of Cancer. JNCI: Journal of the National Cancer Institute 2001;93:1054-61.
5. Parikh CR, Thiessen-Philbrook H. Key concepts and limitations of statistical methods for evaluating biomarkers of kidney disease. J Am Soc Nephrol 2014;25:1621-9.
6. Boyko EJ, Alderman BW. The use of risk factors in medical diagnosis: Opportunities and cautions. Journal of Clinical Epidemiology 1990;43:851-8.
7. Kattan MW. Judging New Markers by Their Ability to Improve Predictive Accuracy. JNCI: Journal of the National Cancer Institute 2003;95:634-5.
8. Emir B, Wieand S, Su JQ, Cha S. Analysis of repeated markers used to predict progression of cancer. Stat Med 1998;17:2563-78.
9. Zhang R, Brennan ML, Fu X, et al. Association between myeloperoxidase levels and risk of coronary artery disease. Jama 2001;286:2136-42.
10. A. Lachenbruch P. The odds ratio. Controlled clinical trials 1997;18:381-2.
11. Leisenring W, Pepe MS. Regression Modelling of Diagnostic Likelihood Ratios for the Evaluation of Medical Diagnostic Tests. Biometrics 1998;54:444-52.
12. Smith PJ, Hadgu A. Sensitivity and specificity for correlated observations. Stat Med 1992;11:1503-9.