Supplement for "A risk stratification approach for improved interpretation of diagnostic accuracy statistics"

eAppendix. Relationship of MRS and NNTest to other statistics: Agreement, Kappa, Population Attributable Risk, Attributable Community Risk, Kraemer's Kappa statistics, Net Reclassification Index, Integrated Discrimination Improvement, and Total Gain.

Quantities from the risk stratification distribution are related to agreement, Population Attributable Risk (PAR), and "weighted kappa" statistics for evaluating diagnostic tests proposed by Helena Chmura Kraemer [1,2]:

$$\kappa(0,0)=\frac{Spec-P(M-)}{P(M+)}=\frac{PPV-P(D+)}{P(D+)},$$
$$\kappa(1,0)=\frac{P(D+)-cNPV}{P(D+)}=\frac{Sens-P(M+)}{P(M-)},$$
$$\frac{1}{\kappa(w,0)}=\frac{w}{\kappa(1,0)}+\frac{1-w}{\kappa(0,0)}.$$

Recall that MRS is a weighted mean of the two components *PPV-P(D+)* and *P(D+)-cNPV*. Thus κ(0,0) and κ(1,0) are those two components standardized by their maximum values to ensure they are between [-1,1]. Although Kraemer's kappa statistics are valuable, the standardization loses epidemiologic interpretation. In particular, even if either κ(0,0) or κ(1,0) is 1, if disease is rare enough then there is still little absolute risk stratification provided by the test. Note κ(1,0)=*PAR*, showing that PAR is *P(D+)-cNPV* standardized by disease prevalence. The numerator of PAR is the Attributable Community Risk [3], first defined in table 35, pg. 230 of [4].

The weighted harmonic mean κ(w,0) is a weighted Kappa agreement statistic [5] (w=0.5 yields the unweighted Cohen's Kappa.). Weighted Kappa never equals the MRS for any weight. However, because $P(D+,M+)-P(D+)P(M-)=P(D-,M-)-P(D-)P(M-)$,

$$MRS=P(D+,M+)-P(D+)P(M-)+P(D-,M-)-P(D-)P(M-)$$

MRS is the percent agreement minus the sum of the product of the margins, which is the numerator of the unweighted Kappa. Kappa uses a denominator to standardize to being between -1 and 1. Although Kappa has clear interpretation when it is -1, 0, or 1, in our experience, other values are hard to interpret epidemologically or clinically. Since the numerator of Kappa is MRS, it is the denominator used for standardization that is the culprit.

All of the above statistics are standardized versions of either MRS, or components of MRS. Although standardization gains simple statistical interpretation at the extremes and middle, it loses epidemiologic and clinical interpretation as absolute risk stratification.

MRS is neither Net Reclassification Index (NRI) nor the Integrated Discrimination Improvement (IDI) [6]. For comparing a binary test to no test, the NRI is twice Youden's index and the IDI is Youden's index. When comparing two binary tests for the same disease, IDI is the difference in their Youden's indicies, while the ratio of the MRSs is the ratio of the Youden's indices.

Total Gain (TG) is a statistic for measuring the explanatory power of a continuous covariate x in a binary regression model $P(Y=1\,|\,x)=p(x)=G(\alpha+\beta x)$ where G is typically the logistic function [7]. Denote overall disease prevalence as $P(Y=1)=p$. By the mean value theorem there exists an x^* such that $P(Y=1\,|\,x=x^*)=p(x^*)=p$. Then

$$TG=2\left|\int_{x^*}^{-\infty}(p(x)-p)\,dF(x)\right|.$$

This simplifies to $TG = 2\left|P(Y=1, x > x^*) - P(Y=1)P(x > x^*)\right|$. Although this expression is close to MRS, TG is always non-negative, and negative MRS is allowable.

Furthermore, the existence of x^* depends on x being continuous (and differentiable). Thus TG cannot apply to discrete-valued tests, like that we consider in this paper. However, we can extend TG to a discrete covariate x if every value of $\{x : x \geq x^*\}$ increases risk, i.e. $P(Y = 1 \mid x) > P(Y = 1)$ if $x \geq x^*$, then we can define $M+ \equiv \{x \geq x^*\}$ to extend TG to a discrete covariate. However, this extension of TG does not simplify to MRS if the cutpoint x^* does not exactly divide x so that $x \geq x^*$, always increases risk and $x<x^*$ always decreases risk. An example of this is Pap testing in China [8]. Pap results are quaternary: negative, ASC-US, LSIL, and HSIL. The customary cutpoint for Pap positivity is non-negative, which is ASC-US or worse. In China, the overall disease prevalence is 1.6%. Testing LSIL and HSIL have PPVs of 5.4% and 35% respectively, but testing ASC-US has a "PPV" of 1.2% that is lower than disease prevalence. Therefore defining x^* as ASC-US includes a value that decreases risk in the set of results considered positive (namely, ASC-US). Thus MRS based on defining positivity as ASC-US or worse is not TG. In this example, only if x^* is LSIL would the discrete TG equal MRS (and only up to a sign).

References

1. Kraemer HC. *Evaluating Medical Tests: Objective and Quantitative Guidelines* Sage Publications Inc, 1992.
2. Kraemer HC. Reconsidering the odds ratio as a measure of 2x2 association in a population. *Stat Med* 2004;**23**(2):257-270.
3. Wacholder S. The impact of a prevention effort on the community. *Epidemiology* 2005;**16**(1):1-3.
4. McMahon B, Pugh TF, Ipsen J. *Epidemiologic Methods*. Boston: Little, Brown, 1960.
5. Fleiss JL, Cohen J. The equivalence of the weighted Kappa and the intraclass correlation coefficient as a measure of reliability:. *Educational and Psychological Measurements* 1973;**33**:613-619.

6. Pencina MJ, D'Agostino RB, D'Agostino RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 2008;**27**(2):157-72; discussion 207-12.
7. Bura E, Gastwirth JL. The Binary Regression Quantile Plot: Assesing the Importance of Predictors in Binary Regression Visually. *Biometrical Journal* 2001;**43**:5-21.
8. Zhao FH, Hu SY, Zhang Q, Zhang X, Pan QJ, Zhang WH, Gage JC, Wentzensen N, Castle PE, Qiao YL, Katki HA, Schiffman M. Risk assessment to guide cervical screening strategies in a large chinese population. *Int J Cancer* 2016.