Preventive Medicine 110 (2018) 100–105

Contents lists available at ScienceDirect

Preventive Medicine

journal homepage: www.elsevier.com/locate/ypmed

A novel metric that quantifies risk stratification for evaluating diagnostic tests: The example of evaluating cervical-cancer screening tests across populations

Hormuzd A. Katki*, Mark Schiffman

Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, DHHS, Bethesda, MD, USA

ARTICLE INFO

Keywords:
AUC
Biomarkers
Cervical cancer
Diagnostic testing
HPV
Mean Risk Stratification
Number Needed to Test
Risk prediction
ROC
Screening
Youden's index

ABSTRACT

Our work involves assessing whether new biomarkers might be useful for cervical-cancer screening across populations with different disease prevalences and biomarker distributions. When comparing across populations, we show that standard diagnostic accuracy statistics (predictive values, risk-differences, Youden's index and Area Under the Curve (AUC)) can easily be misinterpreted. We introduce an intuitively simple statistic for a 2×2 table, Mean Risk Stratification (MRS): the average change in risk (pre-test vs. post-test) revealed for tested individuals. High MRS implies better risk separation achieved by testing. MRS has 3 key advantages for comparing test performance across populations with different disease prevalences and biomarker distributions. First, MRS demonstrates that conventional predictive values and the risk-difference do not measure risk-stratification because they do not account for test-positivity rates. Second, Youden's index and AUC measure only multiplicative relative gains in risk-stratification: $AUC = 0.6$ achieves only 20% of maximum risk-stratification ($AUC = 0.9$ achieves 80%). Third, large relative gains in risk-stratification might not imply large absolute gains if disease is rare, demonstrating a "high-bar" to justify population-based screening for rare diseases such as cancer. We illustrate MRS by our experience comparing the performance of cervical-cancer screening tests in China vs. the USA. The test with the worst $AUC = 0.72$ in China (visual inspection with acetic acid) provides twice the risk-stratification (i.e. MRS) of the test with best $AUC = 0.83$ in the USA (human papillomavirus and Pap cotesting) because China has three times more cervical precancer/cancer. MRS could be routinely calculated to better understand the clinical/public-health implications of standard diagnostic accuracy statistics.

1. Introduction

After a new biomarker is convincingly associated with disease, a next question is to assess preliminarily, without enumerating costs/benefits/harms, whether the new biomarker is predictive enough to justify formal cost-effectiveness analyses that determine clinical/public-health use. For example, our work involves assessing whether new biomarkers might be useful for cervical-cancer screening in low/middle/high-resource populations, across which both disease prevalence and biomarker distributions vary substantially (Schiffman et al., 2007).

Unfortunately, standard metrics reported for binary biomarkers provide at best indirect information about predictiveness for clinical/public-health use. The odds-ratio is a well-known poor measure of predictiveness (Pepe et al., 2004). When comparing two tests, it is uncommon for one test to have both higher sensitivity and specificity, or both higher positive predictive value (PPV) and negative predictive value (NPV). Two summary statistics, Youden's Index (Youden, 1950) and the Area Under the Curve (AUC) statistic (Hanley and McNeil, 1982), have been correctly criticized for not taking predictive values (i.e. absolute risks) into account, and for not permitting differential weighting of false-positives versus false-negatives (Greenhouse et al., 1950). AUC is the probability that someone with disease has higher risk than someone without disease, which requires only the risk ranks, not the absolute risks (Hanley and McNeil, 1982). Although AUC is used for continuous biomarkers, because the AUC for a binary test is a function of only Youden's Index (Cantor and Kattan, 2000), criticisms of Youden's index also apply to AUC.

To better understand the implications of standard diagnostic accuracy metrics for clinical/public-health utility, we reinterpret standard metrics in light of a novel framework for quantifying risk-stratification. Although "risk-stratification" is a broadly used term, we define it as the ability of a test to separate those at high risk of disease from those at low risk (Wentzensen and Wacholder, 2013), allowing clinicians to

* Corresponding author at: HAK: Division of Cancer Epidemiology and Genetics, National Cancer Institute, 9609 Medical Center Dr. Room 7E592, Bethesda, MD 20892, USA.
E-mail address: katkih@mail.nih.gov (H.A. Katki).

https://doi.org/10.1016/j.ypmed.2018.02.013
Received 12 December 2017; Received in revised form 6 February 2018; Accepted 11 February 2018
Available online 14 February 2018
0091-7435/ Published by Elsevier Inc.

intervene only for those that testing indicates are more likely to develop disease. We introduce two new broadly applicable, linked statistics that have proven useful in our epidemiologic work. We define mean risk-stratification (MRS) as the average change in risk of disease revealed by a test (post-test minus pre-test). We also define a complementary statistic, the Number Needed to Test (NNtest), which quantifies how many people require testing to identify one more disease case than would be identified by randomly selecting people for testing. All statistical details, including comparison to other statistics such as Decision Curves (Vickers and Elkin, 2006), are in a preprint (Katki, n.d.). We have also used MRS to assess agreement between occupational exposure experts (Dopart et al., n.d.).

MRS/NNtest have 3 advantages for comparing test performance across populations with different disease prevalences and biomarker distributions. First, PPV, NPV, and the risk-difference do not measure risk-stratification because they do not account for test-positivity rates, unlike MRS. That is, biomarkers that are rarely positive cannot provide much risk-stratification, regardless of how large the risk-difference is. Second, Youden's index and AUC measure multiplicative relative gains in risk-stratification. However, large relative gains in risk-stratification might not imply large absolute gains if disease is rare. Third, the maximum MRS is limited by disease prevalence. Thus, little risk-stratification is possible for rare diseases such as cancer, demonstrating a 'high-bar' to justify population-based screening. MRS/NNtest re-interpret the risk-difference, Youden's index, and AUC, in the context of disease prevalence and biomarker distributions, which is crucial when considering test performance across populations. AUC cannot be used to rank tests between populations with different disease prevalence, and we show examples from our experience. Our webtool calculates MRS/NNtest (https://analysistools.nci.nih.gov/biomarkerTools/).

2. Background: cervical-cancer screening tests

Human papillomavirus (HPV) causes almost all cervical cancer (Schiffman et al., 2007). HPV-based screening is replacing cervical cytology ("Pap smears"), but many new tests are available or being developed (Castle et al., 2011; Wentzensen et al., 2015). For low/middle-income countries, a low-cost, but unreliable and inaccurate, test is visual inspection with acetic acid (VIA) (Gravitt et al., 2010; Shastri et al., 2014). Screen-positive women by any test are referred for definitive disease ascertainment by colposcopy. To expedite screening guidelines development (Massad et al., 2013; Schiffman et al., 2017), we propose using MRS/NNtest to better identify potentially useful biomarkers.

To illustrate MRS/NNtest, we present data from 2 collaborations. Colleagues in China evaluated 3 tests (HPV testing, cervical cytology, and VIA) in an unscreened population of 30,371 women, to select a test as the basis for future nationwide screening (Zhao et al., 2016). Second, in support of US screening guidelines, we previously analyzed data on 331,818 women screened at Kaiser Permanente Northern California (KPNC) with cervical cytology, HPV testing, or both concurrently ("cotesting") (Katki et al., 2011).

Table 1b
Summary characteristics of cervical screening tests in two populations: an unscreened population in China (1.6% precancer/cancer prevalence)[17] and a previously heavily screened population in the USA at Kaiser Permanente Northern California (0.55% precancer/cancer prevalence)[18]. Acronyms: HPV (human papillomavirus), Pap (Papanicolaou Test), KPNC (Kaiser Permanente Northern California), VIA (visual inspection with acetic acid), LR+ (Likelihood Ratio positive), LR− (Likelihood Ratio negative), AUC (Area under the receiver operating characteristic curve), MRS (Mean Risk Stratification), NNtest (Number Needed to Test). Note that the odds ratio in Table 1a equals LR+/LR−.

Population	Cervical screening testing modality	LR+	LR−	Youden's index	AUC	MRS	NNtest
Unscreened population in China	HPV	6	0.03	82%	91%	2.60%	77
	Pap	6	0.06	80%	90%	2.60%	77
	VIA	5	0.50	45%	72%	1.43%	140
Screened population KPNC (USA)	HPV/Pap cotesting	10	0.28	67%	83%	0.73%	275
	HPV	15	0.31	65%	83%	0.71%	280
	Pap	9	0.69	30%	65%	0.32%	633

Table 1a–b shows standard metrics for each test in China and KPNC. When comparing two tests in the same population, there is usually a tradeoff of sensitivity vs. specificity, or PPV vs. cNPV. This dilemma makes it hard to draw firm conclusions on the basis of any single statistic.

3. Methods: Mean Risk Stratification (MRS) and Number Needed to Test (NNtest)

In the absence of test results or other pre-test information, each individual can only be assigned as a best guess the same population-average risk $\pi = P(D+)$. After taking a test, people learn how their predicted individual risks differ from population-average. Two outcomes are possible:

1. With probability $p = P(M+)$, the test is positive. The person's risk increases from $\pi = P(D+)$ to Positive Predictive Value ($PPV = P(D+|M+)$), an increase of $PPV - \pi$.
2. With probability $P(M-) = 1 - p$, the test is negative. The person's risk decreases from $\pi = P(D+)$ to complement of Negative Predictive Value: $cNPV = 1 - NPV = P(D+|M-)$. The person's risk decreases by $\pi - cNPV$.

Mean Risk Stratification (MRS) is a weighted average of the increase in risk among those who test positive and the decrease in risk among those who test negative:

$$MRS = \{PPV - \pi\}p + \{\pi - cNPV\}(1-p). \quad (1)$$

MRS is the average difference between predicted post-test individual risk and population-average (pre-test) risk. Simply, MRS is the

Table 1a
Standard characteristics of cervical screening tests in two populations: an unscreened population in China (1.6% precancer/cancer prevalence)[17] and a previously heavily screened population in the USA at Kaiser Permanente Northern California (0.55% precancer/cancer prevalence)[18]. Acronyms: HPV (human papillomavirus), Pap (Papanicolaou Test), KPNC (Kaiser Permanente Northern California), VIA (visual inspection with acetic acid), PPV (positive predictive value), cNPV (complement of negative predictive value), Sens (Sensitivity), Spec (Specificity), LR+ (Likelihood Ratio positive), LR−(Likelihood Ratio negative), AUC (Area under the receiver operating characteristic curve), MRS (Mean Risk Stratification), NNtest (Number Needed to Test).

Population	Cervical screening testing modality	Odds ratio	Test Positivity	PPV	cNPV	Sens	Spec
Unscreened population in China	HPV	206	17.3%	9.1%	0.05%	98%	84.0%
	Pap	108	16.7%	9.4%	0.10%	95%	84.6%
	VIA	11	10.7%	8.3%	0.83%	55%	90.1%
Screened population KPNC (USA)	HPV/Pap cotesting	37	7.5%	5.4%	0.16%	74%	92.9%
	HPV	47	5.1%	7.6%	0.17%	70%	95.3%
	Pap	13	3.8%	4.7%	0.36%	34%	96.3%