EDUCATIONAL TRACK FOR RESIDENTS－REVIEW ARTICLE

Statistics in Nuclear Cardiology: Evaluating Diagnostic Accuracy

Kathryn A. Williams, MS[1], David Harrild, MD, PhD[2] and David N. Williams, PhD[3]

Received: July 16, 2016/Revised manuscript received: July 22, 2016/Accepted: July 25, 2016

Abstract

Developing novel nuclear cardiology approaches requires evaluating their diagnostic ability. Statistical measures such as sensitivity, specificity, receiver operating characteristic (ROC) curve, and area under the curve (AUC) are useful to evaluate the diagnostic value of novel imaging parameters. This paper reviews key statistical methods used in the evaluation of diagnostic tests and highlights their use in clinical research settings.
Keywords: Diagnostic testing, Nuclear cardiology, ROC, Statistics
Ann Nucl Cardiol 2016；2（1）：174-177

New imaging parameters are required to evaluate diagnostic outcomes. In nuclear cardiology imaging, sensitivity (Se), specificity (Sp) and accuracy are frequently used to evaluate the diagnostic capability of these new parameters. Receiver operating characteristic (ROC) curves play an important role in demonstrating the advantage of new imaging parameters compared to conventional approaches. In the first issue of Annals of Nuclear Cardiology, Tanaka et al. present a good example of the use of these statistical techniques in their article on the diagnostic value of left ventricular dyssychrony parameters to detect multivessel coronary artery disease (CAD)(1). This paper reviews these key statistical approaches and highlights their use in the article by Tanaka et al.

Diagnostic testing with binary imaging parameters

A variety of cardiac imaging techniques may be used to aid in the diagnosis of suspected cardiac disease in high-risk patients. To be effective, parameters derived from imaging techniques must be able to distinguish between patients with cardiac disease or conditions and those without. For example, to use left ventricular (LV) mechanical dyssynchrony to identify the presence of multivessel CAD, Tanaka et al. used phase analysis with gated stress single-photon emission computed tomography (SPECT).

For all diagnostic tests within a symptomatic population, the goal is to provide an accurate diagnosis (without false diagnoses) at a justifiable cost. To assess the performance of a test, it is usually compared to a "gold" standard (e.g. histology, blood test, better imaging modality). The gold standard (GS) is not always perfect, but must be accepted as the best available measure based on standards and validation of the test (i.e. the best results in many studies). The GS itself may not be used routinely for a variety of reasons, such as cost, time constraints, invasiveness, or expertise required. The GS is a best available surrogate for the true presence or absence of disease. The presence or absence of disease or condition references throughout this article relate to whether the GS showed presence or absence of disease or condition.

In order to assess the ability of an imaging test, key results need to be considered. These results are best displayed in the following Table 1 where the test results (positive or negative) are contrasted against the GS results by classifying each of n patients into one of four categories:

- True Positive (TP)=both test and GS positive
- False Positive (FP)=test positive but GS negative
- False Negative (FN)=test negative but GS positive
- True Negative (TN)=both test and GS negative

Diagnostic test development provides the metrics for the

doi：10.17996/ANC.02.01.174

1) Kathryn A. Williams
Senior Biostatistician, Design & Analysis Core, Clinical Research Center, Boston Children's Hospital, Harvard Medical School Teaching Hospital, 21 Autumn Street, Boston, MA 02115, USA
E-mail: Kathryn.Williams@childrens.harvard.edu
2) David Harrild
Department of Cardiology, Boston Children's Hospital, and Department of Pediatrics, Harvard Medical School, Boston, MA, USA
3) David N. Williams
Clinical Research Center, Boston Children's Hospital, and Harvard Medical School, Boston, MA, USA

Table 1 Assessing a Test against a Gold Standard with n patients

		Gold Standard (GS)		
		GS Positive	GS Negative	
Test	Test Positive	TP a	FP b	PPV=a/(a+b)
	Test Negative	FN c	TN d	NPV=d/(c+d)
		Se=a/(a+c)	Sp=d/(b+d)	n=a+b+c+d

With n patients being tested, the counts for each combination of positive and negative between the Test and the GS are the cells of the table labeled a, b, c, d.

TP: true positive, FP: false positive, FN: false negative, TN: true negative, PPV: positive predictive value, NPV: negative predictive value, Se: sensitivity, Sp: Specificity

clinical use of the test. It gives information about "How good is this imaging test at diagnosing the cardiac disease or condition?" While the primary medical goal is to know whether a patient has a positive imaging test diagnosis, it is equally important to have an indication of how correct the test is. There are 7 key measures for evaluating how well the test measures the true presence or absence of disease:

- Sensitivity (Se): the proportion (or percent) of individuals **with** the disease or condition who are correctly classified. This true positive rate reflects the probability of a positive test if you have the disease. It suggests how good the test is when you have the disease or condition (Table 1).
- Specificity (Sp): the proportion of individuals **without** the disease or condition who are correctly classified. This true negative rate reflects the probability of a negative test if you do not have the disease. It suggests how good the test is when you do not have the disease or condition (Table 1). The measure "1-Specificity" gives the false positive rate and it is also a key measure.
- Positive Predictive Value (PPV): the proportion of individuals with a positive test who actually have the disease or condition. It is the probability of disease if you have a positive test. It suggests how good a positive test is at correctly identifying whether you do have the disease or condition (Table 1).
- Negative Predictive Value (NPV): the proportion of individuals with a negative test who lack the disease or condition. It is probability of NO disease if you have a negative test. It suggests how good a negative test is at correctly identifying whether you do NOT have the disease or condition (Table 1).
- Accuracy (ACC): the proportion of individuals with the disease or condition that test positive and individuals without the disease or condition that test negative. Calculated as the true positive plus true negative rate, this measure reflects the proportion of tests with the correct results; (a+d)/n (Table 1).
- Positive Likelihood Ratio (LR+): the likelihood of a positive test given disease relative to the likelihood of a positive test given no disease; Se/(1-Sp).
- Negative Likelihood Ratio (LR-): the likelihood of negative test given disease relative to the likelihood of a negative test given no disease; (1-Se)/Sp.

These 7 measures help to quantify the diagnostic ability of an imaging test. Of the 7 measures, four (Se, Sp, LR+, LR-) are properties of the test, regardless of the population to which it is applied. In contrast, PPV, NPV and ACC depend on both the test and the prevalence of the disease. Therefore, the design of the study that provides these measures (and in particular the study prevalence of disease) must be considered as well as the prevalence in the population that will be tested in the future. A test with excellent sensitivity could have a very poor PPV because, if the condition is rare, most of the positives will be false positives. Conversely a test with excellent specificity could have a poor NPV, because if the condition is common, most of the negatives will be false negatives. For further clarification, suppose the study is done by sampling 100 cases and 100 controls, the prevalence in the study data will be 50%, but in practice the disease may be extremely rare and the PPV and NPV estimated from the study data will not be useful. On the other hand, if the study design was cross-sectional or prospective (randomly sampled from the population) then there is at least a better chance that prevalence in the study will reflect the population prevalence, and in this case PPV and NPV from the study can be generalized. In clinical practice, in which only the imaging test result is known, the 7 measures and the design of the study from which the measures were derived can be used to inform the ordering physician about the utility of the test at identifying presence of the disease.

Ideally all of the first five measures should be close to a proportion of 1.0 or 100%. In reality, this is rarely the case, and clear tradeoffs are present. The "SnOut & SpIn" guideline (2) is useful in understanding the use of Sensitivity and Specificity.

- SnOut (or SnNOut) -highly **Se**n**s**itive and **N**egative results are good for ruling **Out** disease
- SpIn (or SpPIn) -highly **Sp**ecific and **P**ositive results are good for ruling **In** disease.

It is important to keep in mind that with rare diseases SpIn is not appropriate since ruling in disease based on a patient's positive test is not reasonable (PPV is too low). It is also important to remember false positives and false negatives may not be equally undesirable; in a given diagnostic situation one may have higher costs, either financial or ethical, than the other. In situations where initial tests are used to identify patients that need additional more invasive tests, more false positives may be acceptable in the first round of testing

Ann Nucl Cardiol 2016；2（1）：174-177

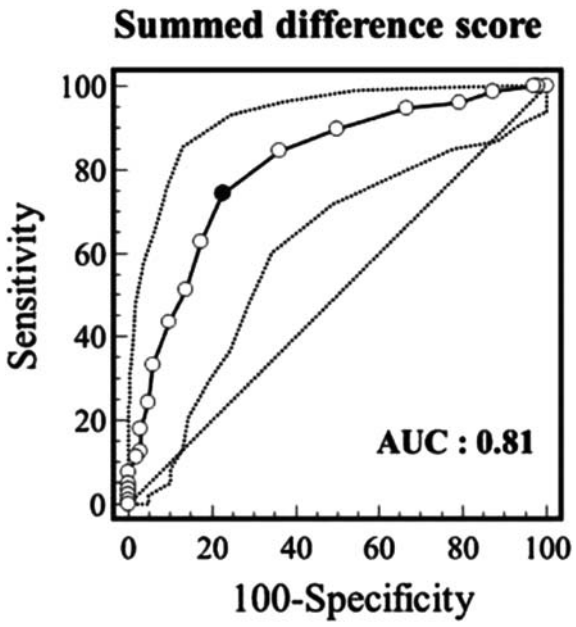

Fig. 1 Receiver Operating Characteristic (ROC) curve illustration from Tanaka et al.

The ROC curve and its 95% confidence intervals and the AUC at 0.5 are represented on the figure. The cutpoint for multivessel CAD was defined as a summed difference score of ＞5. With this dichotomizing of SDS the sensitivity is 74% and specificity is 78% (the black dot on the ROC curve maximizes the sum of the sensitivity and specificity).

because these false positives are expected to be caught in the next round(s). Finally, the last two of the 7 measures, LR+ and LR-, can vary between zero and infinity. For LR＋ higher values are better, while for LR-, the opposite is true.

Diagnostic testing of continuous imaging parameters

With nuclear cardiology imaging, typical diagnostic measures do not provide clear binary responses (Yes or No) to diseases or conditions but may instead offer one of a range of values. Sensitivity and specificity measures can be calculated across the range of values. This is accomplished for each possible value (cutpoint) by dichotomizing the data at that cutpoint and then calculating the Se and (1-Sp) for each cutpoint. A receiver operating characteristic (ROC) curve plots the sensitivity measures against the false positive rate (1-specificity) for the range of cutpoints to help visualize performance of the test (Fig. 1). The area under the ROC curve (AUC -- also known as the c statistic) provides a measure of the overall discrimination ability of the test, e.g. how well the test distinguishes between those who have the disease versus those who do not. The AUC ranges between 0 and 1.0 (or 0 and 100%) and can be assessed as follows:

- c=0.5 suggests no discrimination (i.e. we might as well flip a coin)
- 0.7≤c<0.8 acceptable discrimination
- 0.8≤c<0.9 excellent discrimination
- c≥0.9 outstanding discrimination.

The more the curve follows the y axis and then the x axis the better the discrimination. The diagonal line shows no discrimination (c=0.5; Fig. 1). Another useful interpretation of AUC is that it provides the probability that if you take a random person with the disease and a random person without the disease, the person with the disease will score “higher” on the test (assuming a high score indicates higher risk of disease) than the non-diseased person.

In addition, a measurement value that maximizes sensitivity and specificity can be derived from the curve. It is impossible to maximize both sensitivity and specificity simultaneously. One has to be traded for the other. Therefore the most common method is to choose the cutpoint that maximizes the sum of Se and Sp. This provides a cutpoint to convert the continuous imaging parameters into a binary measure that can be used in the diagnostic statistics calculations.

ROC curves can be used to compare two or more imaging measures designed to diagnose the same cardiac disease or condition. The ROC curve that has the larger area below it (larger AUC) will provide more discrimination, but if the curves cross, the curve with a smaller area might provide more discrimination in an important range of the measure. This range should be considered. Delong et al. developed a statistical method for comparing the areas under two or more correlated ROC curves (3).

Additional considerations

Several excellent articles have been published on the challenges of diagnostic testing and the benefits of using the measures summarized in this article (4-9). There is more to the evaluation of diagnostic tests than could reasonably be included in this article. The Net Reclassification Index (NRI), Integrated Discrimination Improvement (IDI), validating results in independent data sets, combining tests into a single diagnostic measure, calculating PPV and NPV for different prevalences and comparing correlated ROC curves are all examples of additional topics in the wonderful world of statistics for diagnostic tests.

A clinical research application

The study by Tanaka et al. considered three accepted imaging measures for detecting multivessel CAD: summed stress score (SSS), summed rest score (SRS) and summed difference score (SDS). In addition, they analyzed two novel measures: LV dyssynchony applying phase analysis to measure phase standard deviation (Phase SD) and histogram bandwidth (HB). Since all are continuous measures, ROC analysis showed that SDS had the highest AUC (0.81) (Fig.1). The optimal cutpoints were determined using the ROC curves. These cutpoints maximize the sum of sensitivity plus

specificity. Tanaka et al. concluded that SDS alone provided the highest accuracy of 76%, with sensitivity of 74% and specificity of 78%. By adding the consideration of Phase SD and HB with SDS to detect multivessel CAD, the sensitivity increased to 82% while specificity was 76%. In multivariate analysis, these three imaging parameters were significantly associated with multivessel CAD ($p<0.05$).

Tanaka et al. effectively used ROC curve analysis to identify which imaging parameters had the best test performance and determine that the addition of two new metrics (phase SD and HB, after stress) enabled superior identification of patients with multivessel CAD. Other articles such as the position paper on the use of noninvasive cardiac imaging in the diagnosis and evaluation of ischemic heart disease by Beanlands et al. (10) make effective use of diagnostic testing statistics to present their case.

Conclusion

It is essential that practitioners within the field of nuclear cardiology continue to advance diagnostic imaging tests and make full use of statistical tests such as those reviewed in this article. The 7 key statistics described should be calculated in order to understand the potential utility of new diagnostic tests. LR＋ and LR- are important statistical measures that should be more frequently used than they are at present. Increasing incorporation of these statistical techniques into the field of nuclear cardiology will allow for continued improvement in the accuracy and utility of its diagnostic testing.

Acknowledgments

We are indebted to Dr. Henry A. Feldman and Dr. Leslie A. Kalish of the Clinical Research Center at Boston Children's Hospital for their insightful review of the manuscript.

Sources of funding

None

Conflicts of interest

The authors have nothing to disclose.

Reprint requests and correspondence:
Kathryn A. Williams, MS
Senior Biostatistician, Design & Analysis Core, Clinical Research Center, Boston Children's Hospital, Harvard Medical School Teaching Hospital, 21 Autumn Street, Boston, MA 02115, USA
E-mail: Kathryn.Williams@childrens.harvard.edu

References

1. Tanaka H, Chikamori T, Hida S, et al. Diagnostic value of vasodilator-induced left ventricular dyssynchrony as assessed by phase analysis to detect multivessel coronary artery disease. Ann Nucl Cardiol 2015; 1(1): 6-17.
2. Pewsner D, Battaglia M, Minder C, et al. Ruling a diagnosis in or out with "SpPIn" and "SnNOut": a note of caution. BMJ 2004; 329: 209-13.
3. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988; 44: 837-45.
4. Altman DG, Bland JM. Diagnostic tests 1: sensitivity and specificity. BMJ 1994; 308: 1552.
5. Altman DG, Bland JM. Diagnostic tests 2: predictive values. BMJ 1994; 309: 102.
6. Altman DG, Bland JM. Diagnostic tests 3: receiver operating characteristic plots. BMJ 1994; 309: 188.
7. Altman DG, Deeks JJ. Diagnostic tests 4: likelihood ratio test. BMJ 2004; 329: 168-9.
8. Greenhalgh T. How to read a paper. Papers that report diagnostic or screening tests. BMJ 1997; 315: 540-3.
9. Guyatt G, Rennie D, Meade MO, Cook DJ, eds. Users' guides to the medical literature 3rd Edition, Chicago: AMA Press, 2014.
10. Beanlands RSB, Chow BJW, Dick A, et al. CCS/CAR/CANM/CNCS/CanSCMR joint position statement on advanced noninvasive cardiac imaging using positron emission tomography, magnetic resonance imaging and multidetector computed tomographic angiography in the diagnosis and evaluation of ischemic heart disease – executive summary. Can J Cardiol 2007; 23: 107-19.