Published in final edited form as:

Scand J Clin Lab Invest Suppl. 2010 July ; 242: 53–58. doi:10.3109/00365513.2010.493387.

Classification versus association models: Should the same methods apply?

Ziding Feng, PhD

Full Member, Biostatistics and Biomathematics Program, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N., M2-B500, Seattle, WA, 98109-1024, USA

Ziding Feng: zfeng@fhcrc.org

Abstract

Classification and association models differ fundamentally in objectives, measurements, and clinical context specificity. Association studies aim to identify biomarker association with disease in a study population and provide etiologic insights. Common association measurements are odds ratio, hazard ratio, and correlation coefficient. Classification studies aim to evaluate biomarker use in aiding specific clinical decisions for individual patients. Common classification measurements are sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). Good association is usually a necessary, but not a sufficient, condition for good classification. Methods for developing classification models have mainly used that criteria for association models and therefore are not optimal for classification purposes. We suggest that developing classification models by focusing on the region of receiver operating characteristic (ROC) curve relevant to the intended clinical application optimizes the model for the intended application setting.

Keywords

Association; biomarkers; classification; likelihood; logistic regression; odds ratio; ROC curve; sensitivity; specificity

Introduction

The use of biomarkers measured from body fluid or tissue to predict patient outcome is a common practice in medicine. Examples include α -Fetoprotein (AFP) for liver cancer diagnosis and prostate specific antigen (PSA) for prostate cancer diagnosis or prognosis. The vast majority of the traditional biomarker tests do not require modeling. A measure above a pre-selected threshold will prompt a work up test or interventions.

However, a single biomarker is often inadequate for making a clear clinical recommendation due to its false positive and false negative rates. Advances in genomics and proteomics promise better diagnosis or prognosis because new candidate biomarkers are emerging, and there are many of them. Since a new candidate alone will likely have inadequate performance, a logical question is how might it be combined with new or existing biomarkers to improve clinical prediction? Indeed, this is highlighted by the new FDA-approved MammaPrint, a commercial test that evolved from a 2002 report [1] describing an algorithm that combines 70 gene expressions measured from fresh breast tumor tissue to predict breast cancer prognosis.

The most common outcomes to be predicted are binary ($D=1$ for diseased, $D=0$ for no disease) and time-to-event (time to death since diagnosis, time to clinical diagnosis of cancer

since the identification of a pre-cancerous lesion). The most commonly used modeling methods are logistic regression for binary outcome and Cox regression for event time outcome, respectively. Both methods were developed from epidemiologic studies and clinical trials addressing the question of association between risk factors or treatments and outcomes. Classification was not the focus of the settings for which these methods were developed. A natural question is then whether the modeling methods developed from association studies are appropriate for classifications, and if not, what are the appropriate methods?

Association and Classification

Objectives

In association studies, we want to confirm a hypothesis that the risk factor (e.g. biomarker) is associated with a disease. Once confirmed it provides biologic insights to the etiology of the disease and may point to potential interventions for preventing or treating the disease. However, all these statements are made at population level, not at an individual clinical decision-making level, at least not for intervention decisions that might be costly and/or potentially harmful. For example, tobacco is associated with lung cancer. To reduce lung cancer burden in the population we should promote smoking cessation and prevention. However, we do not use smoking status to recommend a lung biopsy to a smoker.

On the other hand, classification is often used for making individual clinical decisions, such as whether to have a more costly and potentially harmful procedure based on the prediction of the model. For example, if a man has elevated PSA, he is often recommended for a prostate biopsy to rule out prostate cancer.

Performance measures

This difference in objectives leads to the differences in measurements of the performance for association and classification models. For association model, odds ratio (OR), the exponential of the regression parameter in logistic regression, is often used for measuring the strength of the association for binary outcome. For event time outcome, hazard ratio, the exponential of the regression parameter in Cox regression, is often used for measuring the strength of the association. Correlation coefficient is usually used for measuring the strength of the association between a continuous outcome and a continuous risk factor. None of these three measures is directly related to decision making. For example, OR is $(p_1/[1-p_1]) / (p_2/[1-p_2])$, where p_1 is the risk of disease for a subject with the risk factor of interest and p_2 is the risk of disease for a subject without the risk factor.

Classification model performance measures are directly related to decision making: sensitivity is $p(Y > c \mid D=1)$, the probability of a test value is above the threshold c for a diseased patient, i.e. probability of making a correct detection decision. Specificity is $p(Y < c \mid D=0)$, the probability of a test value below c for a non-diseased subject, i.e. one minus the probability of making a false alarm decision. Similarly positive predictive value (PPV) is $p(D=1 \mid Y > c)$ and negative predictive value (NPV) is $p(D=0 \mid Y < c)$. The threshold c emphasizes the need to make a decision. To understand a test performance for all possible thresholds, receiver operating characteristic (ROC) curve is commonly used; $ROC(t)$ is defined as sensitivity at a threshold corresponding to false positive rate t (i.e. $1 - \text{specificity} = t$). ROC curve is a plot of $ROC(t)$ against t . For ROC(t) methodology we refer to the book by Pepe [2].

It is important to note that a strong association between a biomarker and disease is usually a necessary condition, but not a sufficient, condition for its use as a classification test. In epidemiologic studies, odds ratios of a magnitude of 2–3 are often considered a strong

association. In genome-wide association studies (GWAS), odds ratios for SNPs associated with disease are frequently in the magnitude of less than 1.5 [3]. For a test to have classification value, it requires its sensitivity and specificity pair only be obtainable when odds ratio is in the range of 25 to above 100 [4], a strength of association rarely observed in association studies.

Clinical context specificity

Association studies are usually not clinical context specific. Neither odds ratio nor hazard ratio tells us whether the strength is in sensitivity or specificity, or how this association could be used to make a specific clinical decision. Either just provides an overall assessment of the strength of the association. On the other hand, classification performance is usually clinical-context specific. It needs to target a high sensitivity or a high specificity in the context of the intended clinical application, and we often know from the consequence of an incorrect decision the required sensitivity and specificity for the test to be useful as a basis for a clinical decision. For example, in ovarian or pancreatic cancer screening in the general population, often the only relevant region of the ROC curve for a test is from specificity 98–100%. A test with specificity lower than this range, could still be very high, say 95%, and will lead to an unacceptable number of false positive test results and invasive screenings because the vast majority of the general population will not have these rare diseases. In high risk populations or diagnostic triage settings where an invasive or costly diagnostic procedure is the default procedure, a new test needs to have very high sensitivity to rule out patients from the default procedure. Therefore, the relevant performance region often is sensitivity 98–100%. Two biomarkers with the same odds ratios, even with the same area under ROC curve (AUC), may perform very differently in the relevant performance region for a specific clinical context.

The clinical context specificity also holds if we use PPV or NPV as performance measures. Generally PPV depends more on specificity while NPV depends more on sensitivity.

An example

Here the clinical context is early detection of liver cancer among cirrhotic patients. In this population the risk for liver cancer is high: the annual incidence is about 2%. The default surveillance modality depends on geography. In Japan and some institutions in the United States, annual CT/MRI is the surveillance modality. In most regions in the United States, AFP and liver ultrasound (US) is used and CT/MRI is only triggered by abnormal findings in AFP or US. In developing countries, AFP is often the main surveillance modality. Since CT/MRI is related to high cost and radiation, in Japan and the regions where CT/MRI is regularly used for surveillance, the objective of a new biomarker test is to use it in combination with AFP/US to reduce unnecessary CT/MRI. To rule out cirrhotic patients for CT/MRI, a new test needs very high sensitivity (at least 95%) similar to the sensitivity of CT/MRI, with some specificity, say 50%. A test with that performance will spare 50% of patients from repeated CT/MRI, a significant clinical utility. The relevant region in ROC curve is $\text{ROC}(t) = 95\text{--}100\%$. For areas where AFP/US is the default surveillance modality, the reasonable target performance for a new blood test would be to increase sensitivity with high specificity, at least compatible to that of AFP/US, a different region in the ROC curve.

In summary, association and classification models differ fundamentally in their objective, performance measurement, strength of association required, and clinical context specificity. Since methods must meet the needs, in the rest of this paper we will first show that the methods for association are not specifically tailored for classification and therefore not optimal for classification. We then suggest methods that we believe are more suitable for classifications.

Modeling methods for association and classification

Though not directly relevant for modeling, it is important to point out that study design for evaluating biomarker performance for classification has unique characteristics. First, the clinical context should be clearly stated and from that follows the definition of study population, definition of diseased and controls, the time and the setting under which specimens and data to be collected, the minimum performance criteria to set up the study hypothesis, and study size. The key is that the clinical context drives everything else. See Pepe et al. [5] for detailed descriptions of study design issues for a pivotal biomarker classification validation study. Failure to use the principles, termed PROBE study design standards, could lead to biased study conclusions not replicable when the test is used in the real clinical context, with potentially devastating consequences if it is used for important clinical decisions. The following discussions assume an appropriate study design has been done and we now have data on multiple biomarkers or a biomarker to be combined with existing predictors to develop a classification model.

The most commonly used modeling method to combine multiple predictors for binary disease outcome is logistic regression [6]. A logistic model has the form $\log(p/(1-p)) = X\beta$ where X is a vector of predictors $x_0, x_1, x_2, x_3, \dots$, e.g. x_0 is 1 and β_0 is related to disease prevalence, x_1 is biomarker value, $x_2, x_3, \dots$ are other clinical predictors. $\exp(\beta_i)$ is OR associated with i -th predictor. One can use $X\beta$ as a combined “new test” to draw the ROC curve, use a specific threshold for classification, and calculate sensitivity, specificity, PPV, and NPV associated with this threshold.

Though logistic regression was developed in epidemiologic association study, it has a nice property for classification. If the underlying logistic model is correct or at least $X\beta$ is a monotone function of the true model, then the classifier defined above is the optimal classifier in the sense that it has the maximum sensitivity given any specificity and vice versa, i.e. we can not improve further its ROC(t) curve for any point t [7]. The reason is that by the famous Neyman-Pearson Lemma [8], likelihood ratio-based decision rule, $p(y|D=1)/p(y|D=0) > c$, is the optimal decision rule. Any monotone increasing function of the likelihood ratio, such as risk score $p(D=1|y)$, and logit transformation of risk score $\log(p(D=1|y)/(1-p(D=1|y)))$ are also optimal [9, 10]. The last one is $X\beta$ under logistic regression model.

However, a proposed model is almost never to be the truth. With that, the logistic regression model is in general not optimal for optimizing the ROC curve, i.e. not optimal for classification decision.

Researchers have used the likelihood ratio principle to construct a multiple predictor classifier even when they do not know the truth model. The idea is that if we can maximize the ROC curve or likelihood ratio directly and non-parametrically, the resulting classifier must be optimal. Since maximizing the whole ROC curve is computationally difficult, and since the optimal ROC curve must have the maximal AUC, Pepe et al. [11] used AUC as an objective function to search biomarker combinations that maximize AUC. In their simulations they found when a logistic regression model is incorrect, maximizing AUC directly could lead to a much better classifier than that from the logistic regression model alone, while when the logistic model is correct, the two methods gave similar performance.

If sample size is sufficiently large, maximizing AUC, if computationally feasible, can indeed lead to the optimal classification model. For finite sample size, it could lead to a classifier that optimizes AUC for a given data but not for the population, an over fit. Since we don't know in which region of the ROC curve the incorrect modeling has occurred, another approach is to focus on maximization of the relevant region of the ROC curve. Baker [9]

used likelihood ratio to maximize the ROC region between sensitivity 98–100% for prostate cancer population screening context.

Baker divided each marker into a number of intervals and formed a grid for two markers (for d markers it will be a d -dimensional grid). Starting from the extreme, i.e. the cell corresponding to the highest level for each marker, representing highest specificity, he selected a new cell to be added, one by one, using likelihood ratio, i.e. sensitivity/specificity ratio for the cell, as the selection criteria. That is to find a region by combining cells to maximize sensitivity/specificity ratio while satisfying specificity at a high value between 98–100%.

Since such a search can lead to a disjoint region, that is both anti-intuitive biologically (a high value is indicative for disease, but when it is higher it is indicative for control, and when it goes even higher it is indicative for disease again) and has more chance to over fit the data. Some restrictions should be applied. Baker used *Jagged ordered* and *Rectangular ordered* but they could be better termed as *or* rule and *and* rule, respectively.

The *or* rule (Jagged ordered rule) predicts a disease for a two marker situation, if marker 1 is above a *or* marker 2 is above b . The *and* rule (rectangular rule) predicts a disease if marker 1 is above a *and* marker 2 is above b . The search path to maximize the likelihood ratio is restricted to this given rule so the decision region is always connected.

We believe that the *or* rule is most often used because of its natural biological appeal. Most cancers are known to be heterogeneous, and there are a number of known or unknown histological or molecular sub-classes for each cancer. If there is a biomarker for each sub-class, then the *or* rule will combine these markers to increase sensitivity without much decrease in specificity.

The *and* rule is more suitable for combining markers that have very high sensitivity but poorer specificity. For example, a tumor and its surrounding tissues always have some inflammation, but other benign diseases could also exhibit inflammation. CA19-9 is elevated in both pancreatic cancer and chronic pancreatitis. PSA is elevated in both prostate cancer and benign prostate disease.

Baker did not discuss the possibility of combining *or* and *and* rules. It is conceivable that this combination could improve the performance. Let's take a hypothetical example. Suppose there are five biomarkers, each elevated in 20% of non-overlapping cases and 10% non-overlapping controls. Using *or* rule to combine these five markers leads to a test with 100% sensitivity and 50% specificity. Let's call this combined test marker A.

Another biomarker B has 100% sensitivity and 50% specificity. If the distribution of marker A and B are independent in controls, using *and* rule to combine marker A and B, i.e. predict disease if marker A is positive *and* marker B is positive, the final test rule will have 100% sensitivity and 75% specificity.

Algorithms

These methods are conceptually simple but difficult to implement when the number of candidates, d , is larger than two. One needs to do a grid search in d -dimensional space. There is also a question of how to form a grid, i.e. how many intervals should we form for each marker. Too-fine grids lead to insufficient numbers in each cell and the likelihood ratio statistics for each cell become noisy, while too-coarse divisions might miss an optimal cutoff point. For an example of four markers and 228 controls and 137 cases, Baker used 5

quartiles, leading to 625 cells in 4-dimensional space. More research should be done in developing efficient and robust algorithms for larger d , say 4 to 10.

Avoid over fitting

Even after we add restriction to the model space, overfitting of data is still a serious threat when the number of candidates is big and the sample size is small. The requirement for sample size increases exponentially with the number of candidates. Therefore it is important to limit the number of candidates to a minimum. To do so, one should first understand the performance characteristics for each biomarker and its biology so there is a strong rationale for inclusion. One can use cross-validation to estimate the performance of a combination rule. To do it right, the steps of the cross-validation should be specified in advance. Going back and forth and using many cross-validations to select one combination rule with the best cross-validation performance defeats the purpose of cross-validation.

Classification models for prognostic tests—Cox regression model [12] is the most popular method for combining predictors to model event time outcome. It has a similar drawback to the logistic regression model for binary outcome in that if the underlying model is incorrect, it is totally unknown whether the resulting classifier has any optimality for classification purpose. It also has the same issue as the logistic regression model in that the resulting classifier does not focus on the relevant performance region of the ROC curve for a specific clinical application. The ROC curve for a prognostic test has a time dimension, e.g. ROC curve for predicting 10-year prostate cancer mortality among prostate cancer patients. The proportional hazard assumption for Cox regression is unlikely to hold for prognostic markers as one could imagine a biomarker performs better when it is measured near the event than it does when it is measured far away from the event. For this reason, Zheng et al. [13] used a weighted logistic regression model to model event time and found that when the proportional hazard assumption is violated, it performed better than the Cox model for classification purposes. When the proportional hazard assumption holds, both methods have similar classification performance in terms of ROC curves.

Prognostic marker study design has an unique issue of selection of cases and controls over time. Some investigators tend to define controls as those who never had an event during follow up and cases as those who had an event (died or disease recurred) before a certain time t . This sampling scheme violates the sampling principles for event time studies in which nested case-control or case-cohort samplings are preferred methods [14, 15]. The optimal sampling method, in terms of efficiency and bias in estimating the classification performance, e.g. 10-year ROC(0.02), for prognostic test evaluation, has not been fully developed.

Summary and additional comments for classification studies

In this paper we argue that association and classification differ fundamentally in:

1. objectives
 - a. association: population relationship
 - b. classification: individual decision
2. performance measurements
 - a. association: odds ratio, hazard ratio, correlation coefficient
 - b. classification: sensitivity, specificity, ROC(t), PPV, NPV. Classification requires much stronger association between biomarkers and disease.

3. clinical context specificity

- a. association: without specific clinical context
- b. classification: specific clinical context demands specific performance characteristics as relevant region in ROC curve

The classification modeling methods mostly are based on association modeling methods. They are therefore not optimal for classification purposes. We recommend use of the likelihood ratio principle and focus on the relevant region of ROC curves for the intended clinical application. The combination of *or* and *and* rules are suggested due to its biological appeal.

Special attention should be paid to appropriate study design for evaluation of classification biomarkers. The gains in efficiency by using optimal methods or more sophisticated methods are usually marginal compared to popular association modeling methods such as logistic and Cox regression. However, a weak study design could bring serious bias in estimating the performance of the test. We recommend PROBE design standards.

There are many other methods for classification modeling, including but not limited to linear discriminant analysis, additive models, tree-based methods, boosting, neural networks, support vector machine, nearest-neighbors, etc. Details covering these topics can be found in the book by Hastie, Tibshirani, and Friedman [16]. Many of them are used in machine learning and data mining areas. Their potentials for clinical classification remain to be seen. Since the simpler methods focused on in this paper have biological appeal to clinicians and only recently have been formally studied for classification in statistical literature, we think they hold more hope for immediate use in clinical classification once their properties are better understood and efficient algorithms become available.

Acknowledgments

This work is supported in part by the National Institutes of Health (U01 CA086368).

References

1. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*. 2002; 415(6871):530–6. [PubMed: 11823860]
2. Pepe, MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press; New York: 2003.
3. Epstein CJ. Medical genetics in the genomic medicine of the 21st century. *Am J Hum Genet*. 2006; 79:434–38. [PubMed: 16909381]
4. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol*. 2004; 159:882–90.
5. Pepe MS, Feng Z, Janes H, Bossuyt PM, Potter JD. Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. *J Natl Cancer Inst*. 2008; 100:1432–38.
6. Weedon MN, McCarthy MI, Hitman G, Walker M, Groves CJ, Zeggini E, Rayner NW, Shields B, Owen KR, Hattersley AT, Frayling TM. Combining information from common type 2 diabetes risk polymorphisms improves disease prediction. *PLoS Medicine*. 2006; 3:1877–82.
7. Green, DM.; Sweets, JA. *Signal Detection Theory and Psychophysics*. New York: Wiley; 1966.
8. Neyman J, Pearson ES. On the problem of the most efficient tests of statistical hypothesis. *Philosophical Transactions of the Royal Society of London*. 1933; 231:289–337. Series A.

9. Baker SG. Identifying combinations of cancer markers for further study as triggers of early intervention. *Biometrics*. 2000; 56:1082–87. [PubMed: 11129464]
10. McIntosh MS, Pepe MS. Combining several screening tests: optimality of the risk score. *Biometrics*. 2002; 58:657–64. [PubMed: 12230001]
11. Pepe MS, Cai T, Longton G. Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics*. 2006; 62:221–29. [PubMed: 16542249]
12. Cox DR. Regression models and life tables (with discussion). *J Roy Stat Soc Ser B*. 1972; 34:187–220.
13. Zheng Y, Cai T, Feng Z. Application of the time-dependent ROC curves for prognostic accuracy with multiple biomarkers. *Biometrics*. 2006; 62:279–87. [PubMed: 16542256]
14. Liddell FDK, McDonald JC, Thomas DC. Methods of cohort analysis: appraisal by application to asbestos miners, *J Roy Stat. Soc Ser A*. 1977; 140:469–91.
15. Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*. 1986; 73:1–11.
16. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*. Springer; New York: 2001.