Published in final edited form as:

Acad Radiol. 2007 August ; 14(8): 985–991.

Reliable Evaluation of Performance Level for Computer-aided Diagnostic Scheme

Qiang Li

Department of Radiology, University of Chicago 5841 S. Maryland Avenue, Chicago, Illinois 60637
Phone: (773) 834–5096, Fax: (773) 702–0371 Email: qiangli@uchicago.edu

Abstract

Computer-aided diagnostic (CAD) schemes have been developed for assisting radiologists in the detection of various lesions in medical images. The reliable evaluation of CAD schemes is an important task in the field of CAD research. In the past, many evaluation approaches, such as the resubstitution, leave-one-out, cross-validation, hold-out, and bootstrap methods have been proposed for evaluating the performance of various CAD schemes. However, some important issues in the evaluation of CAD schemes have not been systematically analyzed, either theoretically or experimentally. The first important issue is the analysis and comparison of various evaluation methods in terms of some characteristics, in particular, the bias and the generalization performance of the trained CAD schemes. The second includes the analysis of pitfalls in the incorrect use of various evaluation methods and the effective approaches to the reduction of the bias and variance caused by these pitfalls. We attempt to address the first important issue in details in this article, and to discuss the second issue in the Discussion section. We believe that this article would be useful to researchers in the field of CAD research for selecting appropriate evaluation methods and for improving the reliability of the estimated performance of their CAD schemes.

Keywords

computer-aided diagnosis; CAD; resubstitution; leave-one-out; hold-out; cross validation; bias; variance; generalization performance

I. Introduction

Computer-aided diagnostic (CAD) schemes have been developed for detecting various lesions in many medical imaging modalities, including conventional radiography, computed tomography, magnetic resonance imaging, and ultrasound imaging. An important issue for CAD schemes is the reliable evaluation of their performance levels. In early publications in CAD research, resubstitution (RS) method was commonly used for the evaluation of CAD schemes [1,2]. Because the performance level estimated by use of the RS method is optimistically biased, investigators in recent years have begun to employ more reliable evaluation methods such as the leave-one-out (LOO), cross-validation (CV), hold-out (HO), and bootstrap (BS) methods. Some investigators have investigated the effect of the sample size on the bias and/or variance of the estimated performance for classifier [3,4] or for CAD schemes [5-8]. However, to our knowledge, no investigator has systematically analyzed and compared these common evaluation methods in terms of multiple important characteristics, in particular,

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

the bias and generalization performance. As a result, the selection of an evaluation method is generally empirical, and sometimes even arbitrary, in many studies for CAD development. We aim to provide such a comprehensive analysis and comparison in order to help researchers select appropriate evaluation methods for improving the quality and reliability of performance evaluation of their CAD schemes.

An important characteristic for various evaluation methods is the bias in the estimated performance level of a CAD scheme. Some evaluation methods such as LOO, CV, and HO are almost unbiased if they are used appropriately. Incorrect use of these methods, however, can lead to significant biases in the estimated performance levels. Therefore, we identified a number of typical pitfalls in the incorrect evaluation methods for CAD schemes, and conducted experiments to demonstrate quantitatively the extent of bias and/or variance caused by each of the pitfalls [9]. In addition, for promoting and utilizing a high standard for reliable evaluation of CAD schemes, we attempted to make recommendations, whenever possible, for overcoming these pitfalls. Because this part has been published before, we will summarize in the Discussion section some important points regarding the reduction of bias and variance.

II. Generalization performance, estimated performance, and bias

For a CAD scheme trained with a finite sample, there are two performance levels. The first is the generalization performance which measures how well the CAD scheme would achieve for a population of patients (all future new cases). Once a CAD scheme is trained, its generalization performance is a fixed value. Unfortunately, it can not be determined directly in practice because investigators are unable to obtain the population of patients when the CAD scheme is designed. Therefore, one often has to determine an estimated performance by applying the trained CAD scheme to a small test sample. The estimated performance is a random value dependent on the small test sample used, and therefore, can be higher or lower than the generalization performance. If on average, the generalization and estimated performance levels are the same, the estimated performance is unbiased; otherwise, it is biased. Different evaluation methods (RS, LOO, CV, HO, and BS) have different ways to select the test sample, and consequently, can be either biased or unbiased. The bias and generalization performance are two important concepts for evaluation of CAD schemes.

III. Analysis and comparison of various evaluation methods

III-1. Methods

This study was conducted based on datasets obtained from 4 Monte Carlo simulation experiments with 4 different sample sizes of 220 (20 abnormal + 200 normal), 250 (50 + 200), 300 (100 + 200), and 400 (200 + 200). The prevalences of abnormal objects were thus 9.1%, 20.0%, 33.3%, and 50.0%. Each Monte Carlo experiment consisted of 100 trials. In each trial, we randomly generated a sample of synthetic objects with one of the above sample sizes. Each object had 6 features. For normal objects, the six features each obeyed a Gaussian distribution with a mean of 0 and respective standard deviations of {0.7, 0.8, 0.9, 1.0, 1.1, 1.2}. For abnormal objects, the six features each obeyed a Gaussian distribution with respective means of {1.0, 1.0, 0.9, 0.9, 0.8, 0.8} and respective standard deviations of {1.2, 1.1, 1.0, 0.9, 0.8, 0.7}. The first and second features, the third and fourth, and the fifth and sixth were correlated with a correlation coefficient of 0.7, 0.5, and 0.3, respectively. There was no correlation between other pairs of features.

The sample of simulation data in each trial is partitioned into a training set and a test set for training and testing a CAD scheme, respectively. In the RS method, the entire sample is used for both training and testing of a CAD scheme. In a k-fold CV method, the entire sample is first randomly partitioned into k disjoint subsets of nearly equal size, and then each of the k

subsets is used as a test set for evaluation of a CAD scheme trained on the other (k-1) subsets. When the size of the subset is equal to 1, the CV is equivalent to the LOO method. In the HO method, the entire sample is partitioned into two subsets (not necessarily, but often, of equal size), one of which is used only for training of the CAD scheme, and the other only for testing of the trained CAD to obtain the estimated performance. In the BS method [10], a training dataset is generated by sampling with replacement N times from the N available cases (N=220, 250, 300, 400) in the entire dataset. The entire dataset is also employed as a test dataset for evaluating the performance level of a CAD scheme. In each trial, the BS sampling was repeated 100 times, and the average performance level of the 100 iterations was reported.

The CAD algorithm employed consisted of two steps. The first step was the selection of features based on Wilks' lambda and the corresponding F-value [11,12]. The second step was the application of a classifier based on Bayesian artificial neural network (ANN) [13,14]. The output of the ANN was employed for the determination of performance level, i.e., the area under observer operating characteristic curve (AUC). We trained the CAD algorithm and calculated its estimated performance levels by use of the RS, LOO, HO, and BS methods. We also created a massive test dataset with 10,000 normal and 10,000 abnormal objects for the determination of the generalization performance levels of the CAD schemes trained with RS, LOO, HO, and BS evaluation methods. We assumed that the sample of 20,000 objects was a good approximation to the population. The mean difference between the generalization performance and the estimated performance was determined as the bias of the estimated performance.

III-2. Results

Figure 1 shows the generalization performance levels of the CAD schemes trained with the RS, LOO, HO, and BS methods for the 100 trials of the Monte Carlo experiment. It is apparent from Fig. 1 that the curves of the generalization performance for the RS and LOO methods almost overlapped, which indicated that the CAD algorithms trained with the RS and LOO methods had almost the same generalization performance levels. Another important observation was that the CAD algorithms trained with the RS and LOO methods achieved a higher generalization performance than did the CAD algorithms trained with the HO and BS methods. Therefore, with a given sample of cases, CAD schemes trained with the RS and LOO will provide better results for a large number of future cases than that trained with the HO and BS.

Figure 2(a) shows the generalization performance levels and the mean estimated performance levels, with error bars of one standard deviation, for the 100 trials by use of the RS method. Because the mean estimated performance levels were much higher than the generalization performance levels, in particular when the available samples were small, the estimated performance levels were optimistically biased. Figures 2(b) and 2(c) show the generalization performance levels and the mean estimated performance levels, with error bars of one standard deviation, for the 100 trials by use of, respectively, the LOO and HO methods. Because the mean estimated performance levels by use of the LOO and HO methods were very close to the corresponding generalization performance levels, the LOO and HO methods were almost unbiased. Figure 2(d) shows the generalization performance levels and the mean estimated performance levels, with error bars of one standard deviation, for the 100 trials by use of the BS method. As with the RS method, the estimated performance levels were optimistically biased, in particular when the available samples were small. However, the variance in the estimated performance level was comparable to that of the RS, and was considerably lower than those of the LOO and HO methods.

III-3. Application of various evaluation methods

Table 1 summarizes the goodness of the generalization performance, the bias of the estimated performance, and three additional important characteristics, i.e., the uniqueness of the CAD algorithm, the variance of the estimated performance, and the computational cost, for the RS, LOO, HO, and BS evaluation methods, as well as 2-fold and 10-fold cross-validation methods. The uniqueness of the CAD scheme implies that a single CAD scheme is obtained with a specific evaluation method. It is apparent from the table that each evaluation method has its critical shortcoming. For example, the RS method estimates an optimistically biased performance; the HO method provides a CAD with a degraded generalization performance for future new cases; the LOO method does not provide a unique CAD algorithm for a given sample because each case in the sample corresponds to a different CAD scheme; and the BS. No single evaluation method is universally superior to the others.

In the early stage of CAD development conducted in many research institutions, investigators often want to reliably estimate the performance of their CAD algorithm based on a given finite sample, whereas it is often not important whether a unique CAD algorithm exists. In such situations, the LOO method appears superior to the others because it can provide an almost unbiased performance estimate for a CAD algorithm with “good” generalization performance. Neither the RS method nor the HO method is appropriate for these situations, because the former provides a biased performance estimate, and the latter does not provide a performance level as good as what the LOO method provides.

In some other situations, investigators not only are interested in the performance of their CAD algorithm, but also want to conduct an observer performance study with a unique trained CAD algorithm, in order to examine whether their CAD algorithm helps radiologists improve their diagnostic performance. The essential requirements in such situations are the uniqueness of a CAD algorithm and the unbiased performance estimate for the cases to be used in the observer study. It is apparent that, in such situations, the HO method is more appropriate than the other methods. To employ the HO method, one first reserves a test subset from the available sample for the observer performance study, then trains the CAD algorithm with the remaining cases, and finally applies the trained CAD algorithm to the reserved test set to obtain an unbiased performance estimate.

Further, in more realistic situations, investigators want to design a practical CAD system for future use, such as all commercial CAD systems and some prototype CAD systems designed for clinical evaluation in some research institutions. Such a CAD system should be a unique one with a good generalization performance for new cases; and, if possible, the generalization performance of the CAD system should be reliably estimated at the time when the system is designed. Although none of the evaluation methods satisfies all the above requirements, the combination of the RS and LOO methods should be adequate for such situations. The CAD designers can first employ the RS method to design a unique CAD system with a good generalization performance. They can then employ the LOO method to estimate the generalization performance reliably. Such an estimate of generalization performance is an unbiased one for the LOO method, and is also an unbiased one for the RS method employed in the design of the CAD system, because, as shown in Fig. 1, the generalization performance for the RS and LOO methods are almost the same.

IV. Discussions

As shown in Table 1, the LOO, CV, and HO methods are almost unbiased. However, this is true only if these evaluation methods are used appropriately. Improper use of these methods can lead to significant biases in the estimated performance levels. We reviewed all research articles on the development and evaluation of CAD schemes that were published in *Medical*

Physics [15-21] and *IEEE Transactions on Medical Imaging* [22-24] in 2004. We found that, among the 10 papers we reviewed, nine employed incorrect evaluation methods leading to increased bias and/or variance in the estimated performance levels [9].

A CAD scheme generally consists of three major steps, i.e., feature selection, classifier, and cutoff threshold selection. In the ten CAD-related papers we reviewed, most investigators included only the classifier step in their evaluation methods such as the LOO and CV methods, and left out the other steps such as feature selection and cutoff threshold selection. This pitfall led to an increased bias in the estimated performance levels of CAD schemes [9]. We found that each step can be a source of bias in the estimated performance level. For example, in our experiment, the step of feature selection was the main source of bias and the classifier was a secondary source of bias [9]. Therefore, it is important to include not only classifier but also all other steps in an appropriate evaluation procedure.

Even if all three steps are included in an appropriate evaluation method, a bias may still occur. For example, if we use a complex classifier such as a ANN, we have to determine some important parameters of the ANN, such as the number of units in each hidden layer, the number of epochs for training the ANN, etc. Investigators often train their ANNs with a training dataset, and select a “good” combination of parameters by looking at the results of ANNs on the “test” dataset. Because parameter selection is part of the training process, the “test” dataset will become part of the training dataset. Thus, this pitfall leads to a bias in the estimated performance. To overcome this pitfall, the performance level of a CAD scheme based on a complex classifier should be estimated by use of a three-subset evaluation method instead of a conventional two-subset method [9]. If the three-subset evaluation method is not practical, the investigators should be aware of the bias in the estimated performance level obtained by use of the two-subset method.

Other factors may also affect the reliability of the estimated CAD performance level. For example, a way to reduce the variation of the estimated performance levels by use of the half-half CV and HO methods is the stratified partition of the entire dataset, in which the training and test subsets have an approximately equal number of abnormal objects and an equal number of normal objects. Stratification may also be made in regard with other important characteristics of objects such as lesion sizes, patterns, and the institutions from which the cases were collected.

The LOO, CV, and HO evaluation method should be patient-based rather than object-based. In a patient-based evaluation method, all objects from a patient belong to either the training dataset or the test dataset, but not both. In this case, the leave-one-out method becomes the leave-one-patient-out method. The patient-based evaluation method can reduce the bias caused by the correlation between the objects in the same patient.

Our experiments were based on simulation data, and therefore, have limitations. For example, we assumed two Gaussian distributions for the two categories, which may not be correct in practice; we used samples of relatively small sizes with a relatively small number of features; and we employed a relatively small repetition number (100 trials) for calculating various mean performances. The use of a complex classifier - the Bayesian ANN - may make it difficult for the reader to reproduce the reported results and to draw conclusions for his or her own field of application. Although these limitations are inevitable in this study because of our limited capacity, the readers should be aware that some details may vary if other simulation experiments are conducted, or if other classifiers are employed. For example, if the number of hidden units of the ANN increased, then the bias of the RS and BS methods would become larger. On that other hand, if a linear classifier was employed instead of the Bayesian ANN,

then the bias of the RS and BS methods would become smaller because the linear classifier tends to be less overtrained than the ANN.

V. Conclusion

With the definition of a generalization performance and a bias in this article, we compared the characteristics of various evaluation methods, discussed methods to reduce the bias and variance in the estimated performance levels by identifying and addressing typical pitfalls in various evaluation methods. This article is important for researchers in the field of CAD research to reliably evaluate the performance levels of the CAD schemes trained with finite datasets.

Acknowledgments

This work was supported by USPHS grants CA62625, CA64370, and CA113820. Q. Li is a consultant to Riverain Medical Group, Miamisburg, OH. CAD technologies developed at the Kurt Rossmann Laboratories for Radiologic Image Research, the University of Chicago, have been licensed to companies including R2 Technologies, Riverain Medical Group, Deus Technologies, Median Technology, Mitsubishi Space Software Co., General Electric Corporation, and Toshiba Corporation. It is the policy of the University of Chicago that investigators disclose publicly actual or potential significant financial interests that may appear to be affected by research activities.

References

1. Chan HP, Doi K, Vyborny CJ, Schmidt RA, Metz CE, Lam KL, Ogura T, Wu Y, MacMahon H. Improvement in radiologists' detection of clustered microcalcifications on mammograms: The potential of computer-aided diagnosis. *Invest Radiol* 1990;25:1102–1110. [PubMed: 2079409]
2. Kobayashi T, Xu X, MacMahon H, Metz CE, Doi K. Effect of a computer-aided diagnosis scheme on radiologists' performance in detection of lung nodules on radiographs. *Radiology* 1996;199:843–848. [PubMed: 8638015]
3. Fukunaga K, Hayes RR. Effects of sample size on classifier design. *IEEE Trans Pattern Anal Mach Intell* 1989;11:873–885.
4. Fukunaga K, Hayes RR. Estimation of classifier performance. *IEEE Trans Pattern Anal Mach Intell* 1989;11:1087–1101.
5. Wagner RF, Chan H-P, Mossoba JT, Sahiner B, Petrick N. Finite-sample effects and resampling plans: applications to linear classifiers in computer-aided diagnosis. *Proc SPIE Conf Medical Imaging* 1997;3034:467–477.
6. Chan HP, Sahiner B, Wagner RF, Petrick N. Classifier design for computer-aided diagnosis: effects of finite sample size on the mean performance of classical and neural network classifiers. *Med Phys* 1999;26:2654–2668. [PubMed: 10619251]
7. Sahiner B, Chan H-P, Petrick N, Wagner RF, Hadjiiski L. Feature selection and classifier performance in computer-aided diagnosis: the effect of finite sample size. *Med Phys* 2000;27:1509–1522. [PubMed: 10947254]
8. Gur D, Wagner RF, Chan H-P. On the repeated use of database for testing incremental improvement of computer-aided detection schemes. *Acad Radiol* 2004;11:103–105. [PubMed: 14746409]
9. Li Q, Doi K. The reduction of bias and variance for the evaluation of computer-aided diagnostic scheme. *Med Phys* 2006;33:868–875. [PubMed: 16696462]
10. Efron, B.; Tibshirani, RJ. *An Introduction to Bootstrap*. Chapman and Hall; New York: 1993.
11. Aoyama M, Li Q, Katsuragawa S, Doi K. Automated computerized scheme for distinction between benign and malignant solitary pulmonary nodules on chest images. *Med Phys* 2002;29:701–708. [PubMed: 12033565]
12. Aoyama M, Li Q, Katsuragawa S, Li F, Sone S, Doi K. Computerized scheme for determination of the likelihood measure of malignancy for pulmonary nodules on low-dose CT images. *Med Phys* 2003;30:387–394. [PubMed: 12674239]
13. MacKay, DJC. PhD Thesis from California Institute of Technology. 1992. Bayesian methods for adaptive models. <http://wol.ra.phy.cam.ac.uk/mackay/thesis.pdf>

14. Kupinski MA, Edwards DC, Giger ML, Metz CE. Ideal observer approximation using Bayesian classification neural networks. *IEEE Trans Med Imaging* 2001;20:886–899. [PubMed: 11585206]
15. Kallergi M. Computer-aided diagnosis of mammographic microcalcification clusters. *Med Phys* 2004;31:314–326. [PubMed: 15000617]
16. Li H, Giger ML, Huo Z. Computerized analysis of mammographic parenchymal patterns for assessing breast cancer risk: Effect of ROI size and location. *Med Phys* 2004;31:549–555. [PubMed: 15070253]
17. Sahiner B, Chan H-P, Roubidoux MA, Helvie MA, Hadjiiski LM, Ramachandra A, Paramagul C, LeCarpentier GL, Nees A, Blane A. Computerized characterization of breast masses on three-dimensional ultrasound volumes. *Med Phys* 2004;31:744–754. [PubMed: 15124991]
18. Nakayama R, Uchiyama Y, Watanabe R, Katsuragawa S, Namba K, Doi K. Computer-aided diagnosis scheme for histological classification of clustered microcalcifications on magnification mammograms. *Med Phys* 2004;31:789–799. [PubMed: 15124996]
19. Nappi J, Frimmel H, Dachman AH, Yoshida H. Computerized detection of colorectal masses in CT colonography based on fuzzy merging and wall-thickening analysis. *Med Phys* 2004;31:860–872. [PubMed: 15125004]
20. Catarious D Jr, Baydush AH, Floyd AE Jr. Incorporation of an iterative, linear segmentation routine into a mammographic mass CAD system. *Med Phys* 2004;31:1512–1520. [PubMed: 15259655]
21. Iordanescu G, Summers RM. Reduction of false positives on the rectal tube in computer-aided detection for CT colonography. *Med Phys* 2004;31:2855–2862. [PubMed: 15543795]
22. Yin T-K, Chiu N-T. A computer-aided diagnosis for locating abnormalities in bone scintigraphy by a fuzzy system with a three step minimization approach. *IEEE Trans on Med Imag* 2004;23:639–654.
23. Paik DS, Beaulieu CF, Rubin GD, Acar B, Jeffrey RB Jr, Yee J, Dey J, Napel S. Surface normal overlap: A computer-aided detection algorithm with application to colonic polyps and lung nodules in helical CT. *IEEE Trans on Med Imag* 2004;23:661–675.
24. Joo S, Yang YS, Moon WK, Kim HC. Computer-aided diagnosis of solid breast nodules: Use of an artificial neural network based on multiple sonographic features. *IEEE Trans on Med Imag* 2004;23:1292–1300.

Figure 1.

The generalization performance levels obtained by use of the resubstitution (GP-RS), leave-one-out (GP-LOO), hold-out (GP-HO), and bootstrap methods (GP-BS).

Figure 2(a).

The generalization performance (GP-RS) and the mean estimated performance (EP-RS), with error bars of one standard deviation, for the 100 trials by use of the resubstitution method. There were large biases between the generalization performance levels and the estimated performance levels, particularly when the sample sizes were small.

Figure 2(b).

The generalization performance (GP-LOO) and the mean estimated performance (EP-LOO), with error bars of one standard deviation, for the 100 trials by use of the leave-one-out method. There was nearly no bias between the generalization performance levels and the estimated performance levels.

Figure 2(c).

The generalization performance (GP-HO) and the mean estimated performance (EP-HO), with error bars of one standard deviation, for the 100 trials by use of the hold-out method. There was nearly no bias between the generalization performance levels and the estimated performance levels.

Figure 2(d).

The generalization performance (GP-BS) and the mean estimated performance (EP-BS), with error bars of one standard deviation, for the 100 trials by use of the bootstrap method. There were large biases between the generalization performance levels and the estimated performance levels, particularly when the sample sizes were small. However, the standard deviation of the estimated performance levels was comparable to that of the RS method, and was smaller than those of the LOO and HO methods.

Important characteristics of the resubstitution, leave-one-out, hold-out, bootstrap, and cross-validation evaluation methods

Evaluation methods	Good generalization performance?	Unbiased estimated performance?	Unique CAD algorithm?	Small variance in estimated performance?	Low computational cost?
Resubstitution	Yes	No	Yes	Yes	Yes
Hold-out	No	Yes	Yes	No	Yes
Leave-one-out	Yes	Yes	No	No	No
Bootstrap	No	No	No	Yes	No
2-fold cross-validation	No	Yes	No	No	Yes
10-fold cross-validation	Yes	Yes	No	No	Yes

Li