Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Biomedical Informatics 38 (2005) 345–346

Journal of Biomedical Informatics

www.elsevier.com/locate/yjbin

Guest Editorial

Clinical machine learning

Investigators have used predictive models in clinical medicine increasingly for risk stratification, diagnosis, and prognostic classification of patients. In this special issue, we focus on clinical applications of different types of classification models, ranging from naïve Bayes classifiers to novel architectures of artificial neural networks. The clinical domains in which these models are compared are diverse and illustrative of the practical use of machine learning in medicine. Although all papers present clinical applications, they either introduce new algorithms or present comprehensive comparisons of modeling techniques that are critical for the development of practical models.

This issue starts with two articles that describe model comparisons in two clinical domains: community acquired pneumonia [1] and interventional cardiology [2]. Cooper et al. [1] provide a comprehensive comparison of 11 different models to predict dire outcomes for patients with community acquired pneumonia. The learning algorithms they utilized span a broad range of techniques. The goal is to construct a model that can assist clinicians in determining which patients should be admitted to the hospital. The article shows that there may be small but significant differences in classification performance. The authors defend the idea that, for highly prevalent conditions such as community acquired pneumonia, a small increase in discrimination may account for large differences in health care costs and therefore methods that result in small increases in model performance should not be overlooked.

There are several clinical predictive models published in the medical literature. Clinicians may wonder whether these models, which usually result from multi-center studies involving large number of patients, are applicable to their patients. Matheny et al. [2] compare published models for predicting in-hospital complications after interventional cardiology procedures, including a model developed by the authors using local data. The authors show that, with few exceptions, the published models attain a high level of discrimination, but their calibration is low. The local model has high discrimination and calibration. This has important implications for the selection of models for individual counseling.

Variable selection is an extremely important step in model formulation, and is utilized in all studies reported in this issue. Dimensionality reduction by variable selection is particularly important when researchers want to define a small number of important markers for the outcome of interest. A comparison of variable selection strategies in the context of different types of Bayesian classifiers for prognosis of porto-systemic shunt in cirrhotic patients is presented by Gomez et al. [3]. The authors compare variable selection approaches that are independent of the predictive model per se (i.e., filter approaches) with those that are guided by the performance of the model (i.e., wrapper approaches). Although the latter are expected, in theory, to produce more accurate models, the authors show that this is not necessarily true in practice. They note, however, that in their data set, the wrapper approaches utilized fewer variables and may therefore be preferable in practical applications.

Practical utilization of prediction models depends not only on their classification performance or the number of variables utilized, but also on their applicability at the point of care. In many cases, it is important that these models be made available to clinicians who do not have immediate access to a computer. For this purpose, Dreiseitl et al. [4] describe a case study regarding the implementation of a predictive model based on logistic regression applied to paper-based nomograms. The paper-based tool provides decision support for clinicians to estimate the probability of malignancy of various types of nevi. In this article, the authors show that there is no critical loss of information when adapting the logistic regression model to a paper-based nomogram, and provide guidelines on how to build such adapted models.

All articles in this issue depict a model's discrimination ability using receiver operating characteristic (ROC) curves. The area under the ROC curve (AUC)

1532-0464/$ - see front matter © 2005 Published by Elsevier Inc.
doi:10.1016/j.jbi.2005.05.009

is a good index of model discrimination. However, the precision with which AUCs are calculated depends on assumptions regarding the distribution of the data. Zou et al. [5] present a novel approach to non-parametric estimation of AUCs, and provide examples of its utilization in two large clinical studies. They discuss how different conclusions may be reached, depending on how the AUCs are calculated.

In the methodology review that closes this special issue, Lasko et al. [6] present a brief tutorial on ROC analysis, point to literature on the utilization of the method, and summarize recent advances in this area. The tutorial can serve as a reference for readers who are not familiar with issues in calculating AUCs and their variances, who need to explain the technique's rationale to collaborators or students, or who want to have a quick reference to the formulae for calculating and comparing AUCs.

Although this issue does not provide examples of all machine learning algorithms that have been applied to medical data, it presents comparative examples that include the types of models that have been used most often in clinical domains. The results of the comparisons illustrate that no particular type of model can be considered the best in all cases, and that changes in model parameters (e.g., variable selection procedures) and their application in a new population may have a greater effect in model performance than the choice of algorithm (e.g., logistic regression, artificial neural networks).

References

[1] Cooper GF, Abraham V, Aliferis CF, Aronis JM, et al. Predicting dire outcomes of patients with community acquired pneumonia. J Biomed Inform 2005;38(5):347–66.

[2] Matheny ME, Resnic FS, Ohno-Machado L. Discrimination and calibration of mortality risk prediction models in interventional cardiology. J Biomed Inform 2005;38(5):367–75.

[3] Blanco R, Inza I, Menino M, Quiroga J, Larrañaga P. Feature selection in Bayesian classifiers for the prognosis of survival of cirrhotic patients treated with TIPS. J Biomed Inform 2005;38(5):376–88.

[4] Dreiseitl S, Harbauer A, Binder M, Kittler H. Evaluating Patient self-assessment: a case study from questionnaire to nomographic decision aid. J Biomed Inform 2005;38(5):389–94.

[5] Zou KH, Resnic FS, Talos I, Goldberg-Zimring D, et al. A global goodness-of-fit test for receiver operating characteristic curve analysis via the bootstrap method. J Biomed Inform 2005;38(5):395–403.

[6] Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L. The use of receiver operating characteristic curves in biomedical informatics. J Biomed Inform 2005;38(5):404–15.

Lucila Ohno-Machado
Decision Systems Group
Department of Radiology
Brigham and Women's Hospital
Division of Health Sciences and Technology
Harvard-MIT, Boston
MA 02115, USA
E-mail address: machado@dsg.bwh.harvard.edu

Available online 5 July 2005