Published in final edited form as:
Circ Cardiovasc Genet. 2010 February 1; 3(1): 1–2. doi:10.1161/CIRCGENETICS.110.937862.

Prediction Models That Include Genetic Data

Paola Sebastiani, PhD and **Thomas T. Perls, MD, MPH**
Department of Biostatistics (P.S.), Boston University School of Public Health; and Department of Medicine (T.T.P.), Geriatrics Division, Boston University School of Medicine and Boston Medical Center, Boston, Mass

Keywords

editorial; genetics; risk factors

Can alternative modeling approaches that integrate genetic data help to improve the prediction of risk for common diseases? In this issue of *Circulation: Cardiovascular Genetics,* Stengård et al[1] set out to answer this question with regard to several genetic variations of the *APOE* gene and risk for ischemic heart disease (IHD). In their study, they included 3686 women and 2772 men with no medical history of IHD from the Copenhagen City Heart Study and sought to correlate IHD events to risk factors such as abnormal lipid levels, hypertension, diabetes, smoking history, and various *APOE* genotype data during a mean of 6.5 years of follow-up.

The traditional statistical approach to this analysis would be to use Cox proportional hazard modeling to map the hazard of developing IHD to a linear combination of significant risk factors. Stengård et al adopted an interesting alternative procedure that is consistent with the intuition that risk factors may have different effects in subjects with different unmeasured exposures (for example different genetic backgrounds and/or other unknown variables). Therefore, rather than looking for the risk factors that have a homogeneous effect on the hazard for IHD, the authors used the rule induction algorithm PRIM to discover subgroups of subjects with varying combinations of phenotypic risk factors for IHD and *APOE* alleles $\varepsilon 2$, $\varepsilon 3$, and $\varepsilon 4$. They then used the PRIM again to further segregate these subgroups according to additional genotypes in the 5′ promoter region of the *APOE* gene and determined how this additional information changed the risk of IHD. Rule induction is one of the most popular approaches to data mining due to its comprehensibility.[2] The method generates a set of "if-then" rules from the data that can be used either as a summary of interesting patterns discovered in the data or as a classification rule to predict the outcome of new subjects. Many rule induction algorithms have been proposed such as classification and regression trees (CART),[3] the algorithm C 4.5 introduced by Quinlan to induce more parsimonious classification and regression trees,4 and more recently PRIM.5 The original CART algorithm implements a recursive partition of the space of input variables to stratify subjects into different risk sets defined by combinations of values of the input variables. The recursive partition is continued until all subjects are allocated without uncertainty to one of the mutually exclusive partitions. Because of the forcing nature of this strategy, the original CART often produces too many rules and overfits the data. The C4.5 algorithm has, among other improvements over CART, a pruning step that reduces the

Correspondence to Paola Sebastiani, PhD, Department of Biostatistics, Boston University School of Public Health, 801 Massachusetts Ave, Boston, MA 02118. sebas@bu.edu.

Disclosures
None.

number of rules and limits the overfitting of the data. Both CART and C4.5 are “bottom-up” algorithms that incrementally add one variable at a time to define the set of rules. PRIM is a top-down strategy that starts with all the variables and peels away as little as possible at each step to determine a parsimonious partition of the samples.

Algorithmic differences aside, the idea of rule induction is to stratify samples into different risk sets in an unsupervised (nonhypothesis driven) way. The rules summarized in Tables 3 and 4 of the article by Stengård et al are the initial risk sets discovered from the data using PRIM and show the gender-specific effect of some risk factors on the hazard for IHD. For example, the *APOE* alleles define different risk sets in men (Table 4, MS 3 and MS 4), but not in women; similarly, high-density lipoprotein is an important risk factor in men but not women, and increased triglyceride levels is an important risk factor in women but not men. These findings highlight the property of the rule induction approach to automatically identify significant risk sets from the data. The authors go a step further and apply the PRIM algorithm to the subsamples identified in Tables 3 and 4 to verify whether the additional genetic information of genotypes of the 5′ *APOE* promoter can further dissect the 7 risk sets into more specific ones. The results of their analysis in Table 5 show that this additional genetic information can indeed stratify female subjects in their sample into more specific risk sets and suggest that variations of the 5′ *APOE* promoter correlate with different hazards for IHD in a complex, nonlinear way.

The choice of the rule induction algorithm is subjective and for the relatively low-dimension problem described in the article of Stengård et al, all 3 algorithms described earlier are likely to discover the same set of rules. An important question is whether we really need another data mining method for this purpose. The authors argue that regression models make assumptions that may limit their applicability to genetic risk modeling of complex traits, and indeed several recent attempts to build genetic risk models using traditional statistical methods have failed to show that additional genetic information can substantially improve the prediction of risk of common diseases such as diabetes[6] or cardiovascular disease.[7] The failure of these and other attempts may be a consequence of the regression modeling approach that can easily reach saturation with a handful of variables, and the use of nontraditional modeling tools such as the one adopted by Stengård et al can prove to be very fruitful.[8–11] With the growing literature asserting accurate risk prediction models,[12] it is also important to establish basic guidelines to evaluate accuracy.[13] Stengård et al use sensitivity, specificity, and percent predicted values in the same data used for rule induction to demonstrate the “additional predictive value” of genetic data. However, this evaluation does not rule out data overfitting, and the possibility that the rules discovered in the Copenhagen City Heart Study may have no value in other populations. Testing sensitivity and specificity in new data with similar characteristics to the discovery set should become a standard step of the model building procedure and the assertion of accuracy.[14]

Finally, the authors make a case for better stratification to better tailor prevention and treatment. However, the value of the rules discovered in the analysis of Stengård et al falls short of apparent clinical utility. The stratification of the female subsample FS 1 into 2 sets defined by the genotypes of the 5′ *APOE* promoter does not help the clinician decide between intervention options. This situation might change in the future if genotype specific treatments have a differential effect on IHD-related outcomes. If one wishes to entertain the possibility of genotype specific effects, then it would also be useful to consider a different implementation of the 2-step PRIM in which the genetic data are used in the first step of the procedure to partition the subjects into subsamples characterized by different genetic profiles and then additional, modifiable risk factors are used to further stratify the subsamples. This approach could suggest direct intervention and prove immediate clinical utility of genetic data.

Acknowledgments

Sources of Funding

This work was supported in part by grant AG025727 from the National Institute on Aging (to T.P.) and RC2HL101212 from the National Heart, Lung and Blood Institute (to P.S.).

References

1. Stengård JH, Dyson G, Frikke-Schmidt R, Tybjærg-Hansen A, Nordestgaard BG, Sing CF. Context-dependent associations between variation in risk of ischemic heart disease and variation in the 5′ promoter region of the *Apolipoprotein E* gene in Danish women. Circ Cardiovasc Interv 2010;3:22–30.
2. Langley P, Simon HA. Applications of machine learning and rule induction. Commun ACM 1995;38:54–64.
3. Breiman, L.; Friedman, J.; Olshen, R.; Stone, C. Classification and Regression Trees. Belmont, CA: Wadsworth; 1984.
4. Quinlan, JR. C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann; 1993.
5. Friedman JH, Fisher NI. Bump hunting in high-dimensional data. Statistics Comput 1999;9:123–143.
6. Meigs JB, Shrader P, Sullivan LM, McAteer JB, Fox CS, Dupuis J, Manning AK, Florez JC, Wilson PW, D'Agostino RB Sr, Cupples LA. Genotype score in addition to common risk factors for prediction of type 2 diabetes. N Engl J Med 2008;359:2208–2219. [PubMed: 19020323]
7. Paynter NP, Chasman DI, Buring JE, Shiffman D, Cook NR, Ridker PM. Cardiovascular disease risk prediction with and without knowledge of genetic variation at chromosome 9p21.3. Ann Intern Med 2009;150:65–72. [PubMed: 19153409]
8. Rodin AS, Boerwinkle E. Mining genetic epidemiology data with Bayesian networks I: Bayesian networks and example application (plasma apoE levels). Bioinformatics 2005;21:3273–3278. [PubMed: 15914545]
9. Sebastiani P, Ramoni MF, Nolan V, Baldwin CT, Steinberg MH. Genetic dissection and prognostic modeling of overt stroke in sickle cell anemia. Nat Genet 2005;37:435–440. [PubMed: 15778708]
10. Huang M, Dinney CP, Lin X, Lin J, Grossman HB, Wu X. High-order interactions among genetic variants in DNA base excision repair pathway genes and smoking in bladder cancer susceptibility. Cancer Epidemiol Biomarkers Prev 2007;16:84–91. [PubMed: 17220334]
11. Cui J. Overview of risk prediction models in cardiovascular disease research. Ann Epidemiol 2009;19:711–717. [PubMed: 19628409]
12. Accad M. Statistics and the rise of medical fortunetellers. Tex Heart Inst J 2009;36:508–509. [PubMed: 20069074]
13. Kraft P, Wacholder S, Cornelis MC, Hu FB, Hayes RB, Thomas G, Hoover R, Hunter DJ, Chanock S. Beyond odds ratios–communicating disease risk based on genetic profiles. Nat Rev Genet 2009;10:264–269. [PubMed: 19238176]
14. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 2008;9:356–369. [PubMed: 18398418]