Invited Commentary

Invited Commentary: Clinical Usefulness of the Framingham Cardiovascular Risk Profile Beyond Its Statistical Performance

Ralph B. D’Agostino, Sr.,* and Michael J. Pencina

* Correspondence to Dr. Ralph B. D’Agostino, Sr., Mathematics and Statistics Department, Boston University, 111 Cummington Street, Boston, MA 02215 (e-mail: ralph@bu.edu).

Initially submitted March 12, 2012; accepted for publication March 13, 2012.

Framingham risk functions (FRFs) have been developed for over 50 years. There have been numerous applications of them and within the last few decades they have been used in drug treatment guidelines. The Adult Treatment Panel III explicitly used a coronary heart disease FRF in their guidelines for cholesterol drug treatment. Evaluation of these functions has traditionally involved discrimination and calibration measures. One major goal of the FRFs is to see if they are valid in non-Framingham settings. In this issue of the Journal, Khalili et al. (Am J Epidemiol. 2012;176(3):177–186) apply a recent global cardiovascular disease FRF to the data from their Tehran Lipid and Glucose Study and demonstrate that the cardiovascular disease FRF performs extremely well: as good as the best risk function generated from the Tehran data. The FRF is transportable to the Tehran data without need for any calibration adjustment. The investigators then move beyond the traditional discrimination and calibration evaluations, look for utility, and apply the decision theory concepts of net benefit fraction. This application makes assumptions about treatment guidelines. There are both useful and negative aspects of this application, and caution is advised against a too enthusiastic acceptance of it to evaluate prediction rules for primary prevention of cardiovascular disease.

cardiovascular diseases; decision support techniques; epidemiologic methods; forecasting; Kaplan-Meier estimate; proportional hazards models; risk

Abbreviations: ATP III, Adult Treatment Panel III; FRF, Framingham risk function.

One of the early major goals of the Framingham Heart Study was to generate assessment tools capable of assessing cardiovascular disease risk for an individual, taking into account the multiple cardiovascular disease risk factors that a person may carry (risk factors such as sex, age, systolic blood pressure, total and high density lipoprotein cholesterol, smoking behavior, and diabetes). For over 50 years, the Framingham Heart Study has generated such assessment tools, often called Framingham risk functions (FRFs). Of particular interest are primary event FRFs where, for a person free of cardiovascular disease, his/her 10-year risk of developing cardiovascular disease can be produced solely on the basis of the risk factors listed above.

The performance of these FRFs has been evaluated by using measures of discrimination and calibration (1). Discrimination measures evaluate the ability of the FRF to separate those who will go on to develop a cardiovascular disease event versus those who will not. Calibration measures evaluate the ability of the FRF to generate accurately the correct probability that a person will develop a cardiovascular disease (1). The evaluation methods for discrimination are the area under the receiver operating characteristic curve (often called the “AUC”) and, for calibration, the Nam-D’Agostino \( \chi^2 \) test (1–3). These measures supply a quantification of how well the FRFs will perform, require no input beyond the data, and do not depend on how the FRF will be used.

FRFs have been produced with good to excellent performance as measured by discrimination and calibration (4–6). In addition, they have been shown to be valid in settings very different from Framingham such as in the US areas outside of Framingham, Spain, China, and India, at times requiring a simple calibration adjustment (7–9).

Khalili et al. (10) take a recent cardiovascular disease FRF (6) and evaluate its validity (transportability) to the
Tehran Lipid and Glucose Study. They follow the validation procedures involving discrimination and calibration as presented by D'Agostino et al. (1). The evaluation involves seeing how well the FRF performs on the Tehran data in addition to comparing its performance with the best function that can be produced with the Tehran data using the same cardiovascular disease risk factors as are in the FRF. The FRF performs extremely well, matching the performance of the best risk functions computed on the Tehran data. The FRF does not even need calibration adjustments. The analysis of Khalili et al. appears to be excellent, and the results are very gratifying with their implication that there is universality across nations in cardiovascular disease risk.

Khalili et al. (10) take their analysis to another step. Over the years, the FRFs have been used not only to quantify cardiovascular disease risk but also as a tool in treatment guidelines. The Adult Treatment Panel III (ATP III) (11) used a FRF that quantified the 10 years’ risk of hard coronary disease, consisting of coronary death or myocardial infarction, in patients free of cardiovascular disease as a component for setting guidelines for treatment with cholesterol medication. Part of the ATP III cholesterol treatment guidelines was the recommendation that a patient with a 10-year risk of hard coronary disease over 20% should receive treatment. Khalili et al. used this to study the utility of the cardiovascular disease FRF on the Tehran study population. Specifically, they quantify what would be the net benefit fraction if treatment were recommended for those with a 10-year risk of cardiovascular disease beyond 20%. The net benefit function is a weighted function of true positives and false positives. The formula is given explicitly in their article. They also investigate what would happen if the cutpoint for treatment were different from 20%. The value of 10% is of particular interest to them.

The authors point out that decision curve analysis cannot be used instead of cost-effectiveness studies, and we agree with them. However, the question remains as to how useful decision curve analysis is in settings similar to the one explored here. The analyses presented by the authors raise the following questions. First, the value of 20% relates to estimating the 10-year probability of hard coronary disease. Cardiovascular disease is much broader, and a cutpoint for treatment would presumably be much higher. More important and very much to the point, the ATP III guidelines did not use solely the cutpoint of 20% for recommending treatment. If the 10-year probability was over 20%, treatment was recommended. In addition, if the 10-year probability was below 20%, then, depending upon how much below it was, the treatment decision would be based on the value of low density lipoprotein cholesterol. This additional provision is not accounted for in the calculation of net benefit function and leads to the net benefit function’s missing all the true positives that would have been recommended treatment. Unlike the discrimination and calibration measures, the net benefit function oversimplifies the true decision rule for treatment and comes up with a quantification that is not a reliable assessment of this rule. One can argue that the actual treatment rule used in the net benefit function could be made closer to a more practical and realistic rule. However, this is easier said than done. Treatment guidelines, such as those of the ATP III panel, often involve careful consideration of the events under consideration and cost trade-offs. For example, the 20% cutoff was based upon the fact that 20% was the average probability of someone with a hard coronary event having a second hard coronary event within 10 years. It is questionable whether taking this 20% and applying it to cardiovascular disease as the outcome make sense. Further, as stated above, ignoring the fact that a 10-year estimate below 20% does not correspond to no treatment implies that the calculated net benefit function corresponds to an oversimplified classification rule rather than the one recommended in practice. Finally, the authors suggest a threshold selection based on the value of net benefit function achieved in reference to “treat none” or “treat all” strategies. This seems to go against a key premise of decision analysis, where a threshold is a function of known or assumed misclassification costs.

So then, the question is what should be done. Should the net benefit function not be computed? We do not agree with this drastic suggestion. Rather, we recommend that the risk functions should first be evaluated with robust self-contained measures such as discrimination and calibration, as was done by the authors. Then, if we know that, in a given setting, the treatment decisions will be based solely on whether the predicted risks exceed some threshold, we can compute the net benefit function. However, we should avoid making oversimplifying assumptions, so that we can easily assess utility with decision rules that do not correspond to reality. Rather, we should perform careful analyses, such as computing the net benefit function, for actual treatment rules. Even better, we should do serious analyses that introduce costs, life-years saved, quality-adjusted life-years, etc. However, as we perform these, we should realize that we are not evaluating the risk function, but rather we are evaluating the consequences of how it will be used in the real world. We fully endorse the distinction between prediction models and prediction rules made by one of the authors (12). However, developing simple treatment rules so that we can compute “utilities” may generate interpretable, yet misleading assessments. The authors have shown the validity of the cardiovascular disease FRF to Tehran data. The question now is how to develop a sensible treatment rule using it, not to compute a utility assessment that does not correspond to reality.

On another point, Khalili et al. (10) suggest that a cutoff of 10% should be used for treating women. This reflects the fact that women do carry less risk than men. An alternative approach would be to consider people with low 10-year risks and compute their long-term risk, such as a 30-year risk (13). Many women will have a low 10-year risk, very much related to their age. It is important to sort out those who are potentially carrying a large long-term risk from those who are not. Treatment recommendations can be based on both the short-term risk and long-term risks. Net benefit function can be computed from such, again mirroring what would be recommended clinically.

Our bottom-line conclusion here is that simple, interpretable performance measures such as discrimination and calibration should be primary tools for evaluating the performance of risk models. Transportability and validation,
with or without recalibration, should also be essential. Utility measures should be computed on realistic treatment rules, not simplified ones. Other elements such as costs should also be considered, but again on realistic treatment rules or to develop realistic and usable treatment rules.

ACKNOWLEDGMENTS

Author affiliations: Department of Mathematics and Statistics, Boston University, Boston, Massachusetts (Ralph B. D’Agostino, Michael J. Pencina); Framingham Study, Framingham, Massachusetts (Ralph B. D’Agostino, Michael J. Pencina); and Department of Biostatistics, Boston University, Boston, Massachusetts (Ralph B. D’Agostino, Michael J. Pencina).

Conflict of interest: none declared.

REFERENCES