Prediction Modeling to Assess the Prognostic Significance of a Biomarker Panel

Ann M. O'Hare* and Kenneth E. Covinsky†
*Departments of Medicine, VA Puget Sound Healthcare System and University of Washington, Seattle, Washington; and †Departments of Medicine, VA Medical Center and University of California, San Francisco, San Francisco, California

J Am Soc Nephrol 21: 2017–2019, 2010.
doi: 10.1681/ASN.2010101078

Prediction models are often used to stratify individuals on the basis of their risk for a future event, such as death or development of disease. Such information can be valuable to a wide range of stakeholders, including patients and clinicians as they evaluate potential treatment strategies and plan for the future, researchers as they strive to identify target populations for clinical studies, and policy makers as they evaluate possible health system–level interventions.

In this issue of *JASN*, Fox *et al.*[1] evaluate the contribution of a panel of biomarkers to predicting the risk for developing a low estimated GFR (eGFR) and albuminuria during a 10-year follow-up period. The study was conducted among participants in the Framingham Offspring Study—an extraordinarily well-characterized community cohort with detailed systematic data collection at baseline and over time. The authors selected biomarkers they postulated *a priori* might be associated with the development of renal dysfunction (plasma renin, serum aldosterone, B-type natriuretic peptide (BNP), C-reactive protein, plasminogen activator inhibitor 1, fibrinogen, and homocysteine). In incremental models adjusted for age, gender, and other baseline characteristics, homocysteine and aldosterone were statistically significantly associated with the low eGFR outcome, and both of these measures along with BNP were associated with the albuminuria outcome. Addition of the biomarker panel increased the area under the receiver operating characteristic (ROC) curve (or C-statistic) from 0.81 to 0.82 for the outcome of low eGFR and from 0.73 to 0.75 for the outcome of albuminuria. In both models, the net reclassification improvement index (NRI) was 6.9% and the relative integrated discrimination improvement index (IDI) was statistically significant. To evaluate the practical utility of these findings for real world decision-making, we provide a step-by-step interpretation of relevant model characteristics.[2,3]

Perhaps one of the most commonly used measures of model performance is the area under the ROC curve, or C-statistic. The ROC curve is a plot of the true-positive *versus* false-positive rate across the range of possible threshold values. The area under this curve is the overall probability that the predicted risk for a patient who experiences an event is greater than for a person who does not experience an event. The C-statistic is thus a test of the discriminative accuracy of a risk prediction model or its ability to distinguish individuals who experience the event from those who do not.

In the study by Fox *et al.*,[1] the addition of the biomarker panel to each multivariate model resulted in a modest but statistically significant improvement in the C-statistic. The discriminative accuracy of the final model for albuminuria was fair (as indicated by a C-statistic in the 0.7 to 0.8 range), and that for low eGFR was good (as indicated by a C-statistic in the 0.8 to 0.9 range).

However, some have argued that the C-statistic—which conveys information about how well a model discriminates between individuals who do and do not develop the outcome—is not as helpful to clinicians as other measures of model performance.[2,3] For example, the clinician may be more interested in how close the predicted risk for a group of patients comes to the actual risk, a property referred to as model calibration. In a well-calibrated model, there is good agreement between the observed and predicted risk for the outcome of interest across risk groups. Model calibration is usually assessed using the Hosmer-Lemeshow statistic or extensions of this test. When a model is well calibrated, the value of this statistic will be small and the *P* value will not be statistically significant, although the *P* value should be interpreted with caution because the statistic may be statistically significant in well-calibrated models from large data sets. In the study by Fox *et al.*,[1] all models except for that measuring the relationship of age and gender with microalbuminuria were well calibrated.

The NRI and IDI are used to assess the impact on risk stratification of adding particular variables (or groups of variables) to a prediction model.[4] These measures describe the impact of incremental changes to the model on risk category assignment. Specifically, the NRI describes the percentage of individuals who were classified more accurately (*i.e.*, those who did not experience the event were reclassified to a lower risk group, and those who experienced the event were reclassified to a higher risk group) minus the percentage who were classified less accurately. The IDI is a continuous version of the same measure. In the study by Fox *et al.*,[1] the NRI was 6.9% for both models and the *P* value for the IDI was statistically significant in both cases. Collectively, these findings suggest that addition of the biomarker panel resulted in a modest but statistically significant improvement in the accuracy of risk stratification.

In reviewing the model parameters described, it is important to remember that the clinical utility of any prediction model is contingent on its ability to assign patients to clinically meaningful risk groups.[3] In calculating the NRI and IDI, patients were assigned to groups with a $<3\%$, 3 to 6%, and $>6\%$ 8-year risk for developing each outcome.

Published online ahead of print. Publication date available at www.jasn.org.

Correspondence: Dr. Ann M. O'Hare, Nephrology Division, VA Medical Center, 1660 S. Columbian Way, Seattle, WA 98109. Phone: 206-277-3192; Fax: 206-764-2022; E-mail: ann.ohare@va.gov

Whether these risk thresholds prove to be clinically meaningful will depend on the context in which the model is applied. Consider the clinician who must weigh the risks and benefits of starting a new medication intended to lower his or her patient's risk for developing albuminuria or a low eGFR. Ideal risk thresholds would distinguish between patients who will benefit from taking a medication to prevent these outcomes from those who will not. The models presented here would at best distinguish between those with a <3 *versus* $>6\%$ risk for developing this outcome during the ensuing 8-year period. By adding the biomarker panel to the standard multivariable model, 6.9% more patients were classified more accurately than were classified less accurately across these strata. For clinical decisions of this sort, a much wider separation of risk thresholds is often desirable.[5–7] Of course, the optimal degree of risk separation will depend on factors such as the potential harms of the intervention, the importance of preventing the outcome, and the effectiveness of the intervention. Similarly, the magnitude of the NRI must be weighed against the incremental cost and burden of ascertaining the additional measures to be included in the model. Similar principles apply to population-level interventions such as screening to identify those at risk for developing albuminuria or a low eGFR, although optimal risk thresholds for screening for these outcomes, which have not been established, may differ from those needed to support clinical decision-making.

Perhaps the most important attribute of a useful prediction model is a clinically significant outcome. Although the outcomes selected for these models—a low eGFR[8] and microalbuminuria[9]—represent the criteria used to define chronic kidney disease,[10] they carry less clinical significance than hard outcomes such as mortality and progression to ESRD,[11–13] for which patients with chronic kidney disease are at risk. Nevertheless, these intermediate measures may serve as useful surrogate outcomes in situations in which more significant clinical outcomes are rare, as is often the case for ESRD.

However, this practice can be problematic if risk for the intermediate outcome does not track with that for the hard outcome. Use of an incident eGFR of <60 ml/min per 1.73 m^2 as an outcome may be particularly problematic because the Modification of Diet in Renal Disease (MDRD) equation does not yield accurate estimates for true GFR at levels close to 60 ml/min per 1.73 m^2.[14] Furthermore, the clinical significance of very moderate reductions in eGFR (*e.g.*, 45 to 59 ml/min per 1.73 m^2) may be uncertain in some groups, such as the elderly.[15,16]

Another important piece of information needed to evaluate the real-world utility of a prediction model is whether the model has been validated in a sample other than the one in which it was developed.[3] Validating a prediction model in a different sample tests for model overfitting and may provide information on the generalization of the model to other populations that perhaps differ in the distribution of risk factors and in the relationship of these risk factors to the outcome. Because the model described in the study by Fox *et al.*[1] has not yet been validated, it is possible that these results may be overly optimistic and may not generalize to other populations.

It is important to recognize that risk prediction models are not intended to identify potential mechanistic or causal associations. Candidate variables are generally selected because they are known or expected to be associated with the outcome. Model building is aimed at identifying a parsimonious set of candidate variables to stratify patients efficiently into groups at high and low risk for the outcome. It is best not to make causal inferences about variables that are either selected or not selected by the modeling process. A prediction model is best judged by how accurately it accomplishes its intended purpose of risk stratification and by pragmatic judgments of its usability.

The study by Fox *et al.*[1] provides an instructive example of how prediction models can be used to evaluate the usefulness of a biomarker panel for real-world decision-making. Such an approach will become increasingly valuable as an ever-increasing number of novel potential biomarkers, including genetic markers, become commercially available.

DISCLOSURES

A.M.O. receives royalties from UpToDate.

REFERENCES

1. Fox CS, Gona P, Larson MG, Selhub J, Tofler G, Hwang S-J, Meigs JB, Levy D, Wang TJ, Jacques PF, Benjamin EJ, Vasan RS: A multi-marker approach to predict incident CKD and microalbuminuria. *J Am Soc Nephrol* 21: 2143–2149, 2010
2. Cook NR: Statistical evaluation of prognostic *versus* diagnostic models: Beyond the ROC curve. *Clin Chem* 54: 17–23, 2008
3. Janes H, Pepe MS, Gu W: Assessing the value of risk predictions by using risk stratification tables. *Ann Intern Med* 149: 751–760, 2008
4. Pencina MJ, D'Agostino RB, Vasan RS: Review: Statistical methods for assessment of added usefulness of new biomarkers. *Clin Chem Lab Med* August 18, 2010 [epub ahead of print]
5. D'Agostino RB Sr, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, Kannel WB: General cardiovascular risk profile for use in primary care: The Framingham Heart Study. *Circulation* 117: 743–753, 2008
6. Lee SJ, Lindquist K, Segal MR, Covinsky KE: Development and validation of a prognostic index for 4-year mortality in older adults. *JAMA* 295: 801–808, 2006
7. Walter LC, Brand RJ, Counsell SR, Palmer RM, Landefeld CS, Fortinsky RH, Covinsky KE: Development and validation of a prognostic index for 1-year mortality in older adults after hospitalization. *JAMA* 285: 2987–2994, 2001
8. Astor BC, Levey AS, Stevens LA, Van Lente F, Selvin E, Coresh J: Method of glomerular filtration rate estimation affects prediction of mortality risk. *J Am Soc Nephrol* 20: 2214–2222, 2009
9. Ninomiya T, Perkovic V, de Galan BE, Zoungas S, Pillai A, Jardine M, Patel A, Cass A, Neal B, Poulter N, Mogensen CE, Cooper M, Marre M, Williams B, Hamet P, Mancia G, Woodward M, Macmahon S, Chalmers J: Albuminuria and kidney function independently predict

cardiovascular and renal outcomes in diabetes. *J Am Soc Nephrol* 20: 1813–1821, 2009
10. Levey AS, Coresh J, Balk E, Kausz AT, Levin A, Steffes MW, Hogg RJ, Perrone RD, Lau J, Eknoyan G: National Kidney Foundation practiceguidelines for chronic kidney disease: Evaluation, classification, and stratification. *Ann Intern Med* 139: 137–147, 2003
11. Hallan SI, Ritz E, Lydersen S, Romundstad S, Kvenild K, Orth SR: Combining GFR and albuminuria to classify CKD improves prediction of ESRD. *J Am Soc Nephrol* 20: 1069–1077, 2009
12. Matsushita K, Selvin E, Bash LD, Franceschini N, Astor BC, Coresh J: Change in estimated GFR associates with coronary heart disease and mortality. *J Am Soc Nephrol* 20: 2617–2624, 2009
13. Hemmelgarn BR, Manns BJ, Lloyd A, James MT, Klarenbach S, Quinn RR, Wiebe N, Tonelli M; Alberta Kidney Disease Network: Relation between kidney function, proteinuria, and adverse outcomes. *JAMA* 303: 423–429, 2010
14. Levey AS, Stevens LA, Schmid CH, Zhang YL, Castro AF 3rd, Feldman HI, Kusek JW, Eggers P, Van Lente F, Greene T, Coresh J: A new equation to estimate glomerular filtration rate. *Ann Intern Med* 150: 604–612, 2009
15. Glassock RJ, Winearls C: CKD in the elderly. *Am J Kidney Dis* 52: 803, author reply 803–804, 2008
16. Anderson S, Halter JB, Hazzard WR, Himmelfarb J, Horne FM, Kaysen GA, Kusek JW, Nayfield SG, Schmader K, Tian Y, Ashworth JR, Clayton CP, Parker RP, Tarver ED, Woolard NF, High KP: Prediction, progression, and outcomes of chronic kidney disease in older adults. *J Am Soc Nephrol* 20: 1199–1209, 2009

See related article, "A Multi-Marker Approach to Predict Incident CKD and Microalbuminuria," on pages 2143–2149.