Citation: Clinical and Translational Gastroenterology (2013) 4, e44; doi:10.1038/ctg.2013.19
 2155-384X/13

www.nature.com/ctg

CLINICAL/NARRATIVE REVIEW

A Primer on Predictive Models

Akbar K. Waljee, MD, MS[1,2], Peter D. R. Higgins, MD, PhD[1] and Amit G. Singal, MD, MS[3,4]

Prediction research is becoming increasing popular; however, the differences between traditional explanatory research and prediction research are often poorly understood, resulting in a wide variation in the methodologic quality of prediction research. This primer describes the basic methods for conducting prediction research in gastroenterology and highlights differences between traditional explanatory research and predictive research.

Clinical and Translational Gastroenterology (2013) **4,** e44; doi:10.1038/ctg.2013.19; published online 26 December 2013
Subject Category: Clinical Review

INTRODUCTION

Prediction research, which aims to predict future events or outcomes based on patterns within a set of variables, has become increasingly popular in medical research.[1] Accurate predictive models can inform patients and physicians about the future course of an illness or the risk of developing an illness and thereby help guide decisions on screening and/or treatment. For example, predictive models have been developed in gastroenterology to predict the risk of disease flares for inflammatory bowel disease and risk of hepatocellular carcinoma among patients with cirrhosis.[2,3]

There are several important differences between traditional explanatory research and prediction research. Explanatory research typically applies statistical methods to test causal hypotheses using *a priori* theoretical constructs (e.g., hepatocellular carcinoma surveillance underutilization is related to provider-level factors[4]). In contrast, predictive research applies statistical methods and/or data mining techniques, without preconceived theoretical constructs, to predict future outcomes (e.g., predicting the risk of hospital readmission[5]).[6] Although predictive models may be used to provide insight into causality of pathophysiology of the outcome, causality is neither a primary aim nor a requirement for variable inclusion.[6] Noncausal predictive factors may be surrogates for other drivers of disease, with tumor markers as predictors of cancer progression or recurrence being the most common example. Unfortunately, a poor understanding of the differences in methodology between explanatory and predictive research has led to a wide variation in the methodologic quality of prediction research.[7] The aim of this primer is to describe basic methods for conducting prediction research, which can be divided into three main steps: developing a predictive model, independently validating its performance, and prospectively studying its clinical impact.

TYPES OF PREDICTIVE MODELS

Although prediction research in medicine has traditionally used a Bayesian framework approach, with statistical techniques such as regression models, data mining techniques such as machine learning algorithms are a form of artificial intelligence that are being used with increasing frequency.[8] Machine learning has been previously used to predict behavior or outcomes in business, such as identifying consumer preferences for products based on prior purchasing history. A number of different techniques to develop predictive algorithms exist, using a variety of prediction analytic tools/software and have been described in extensive detail elsewhere.[8,9] Some examples include neural networks, support vector machines and decision trees. Decision trees, for example, use techniques such as classification and regression trees, boosting and random forest to predict various outcomes. The analysis can be conducted using free software environments such as "R"[10] as well as vendor applications.

Machine learning algorithms, such as random-forest approaches,[11,12] have several advantages over traditional explanatory statistical modeling, such as lack of a predefined hypothesis, making it less likely to overlook unexpected predictor variables or potential interactions. Approaching a predictive problem without a specific causal hypothesis can be quite effective when many potential predictors are available (increasingly common with electronic health records) and when there are interactions between predictors, which are common in biological and social causative processes. Predictive models using machine learning algorithms may therefore facilitate recognition of clinically important risk and variables in patients with several marginal risk factors that may otherwise not be identified. In fact, many examples of discovery of unexpected predictor variables exist in the machine learning literature.[2,3]

[1]Department of Internal Medicine, University of Michigan, Ann Arbor, Michigan, USA; [2]Veterans Affairs Center for Clinical Management Research, Ann Arbor, Michigan, USA; [3]Department of Internal Medicine, UT Southwestern Medical Center, Dallas, Texas, USA and [4]Department of Clinical Sciences, University of Texas Southwestern, Dallas, Texas, USA
Correspondence: Akbar K. Waljee, MD, MS, Veterans Affairs Center for Clinical Management Research, Ann Arbor VA Medical Center, 2215 Fuller Road, 111D, Ann Arbor, Michigan 48105, USA. E-mail: awaljee@med.umich.edu
Received 20 June 2013; revised 26 September 2013; accepted 6 November 2013

DEVELOPING A PREDICTIVE MODEL

The first step in developing a predictive model, when using traditional regression analysis, is selecting relevant candidate predictor variables for possible inclusion in the model; however, there is no consensus for the best strategy to do so.[13] A backward-elimination approach starts with all candidate variables, and hypothesis tests are sequentially applied to determine which variables should be removed from the final model, whereas a full-model approach includes all candidate variables to avoid potential overfitting and selection bias. Previously reported significant predictor variables should typically be included in the final model regardless of their statistical significance but the number of variables included is usually limited by the sample size of the data set.[14] Inappropriate selection of variables is an important and common cause of poor model performance in this situation. As described above, variable selection is less of an issue using machine learning techniques given that they are often not solely based on predefined hypotheses. There are several other important issues related to data management when developing a predictive model, such as dealing with missing data and variable transformation; however, these topics are beyond the scope of this primer and addressed elsewhere.[15–17]

VALIDATING A PREDICTIVE MODEL

For a prediction model to be valuable, it must not only have predictive ability in the derivation cohort but must also perform well in a validation cohort.[7,18] A model's performance may differ substantially between derivation and validation cohorts for several reasons including overfitting of the model, missing important predictor variables, interobserver variability of predictors leading to measurement errors, and differences in the patient cohort case mix.[18] Therefore, model performance in the derivation cohort may be overly optimistic and is not a guarantee that the model will perform equally well in new patients. For example, external validation of the HALT-C predictive model for hepatocellular carcinoma was recently demonstrated to have a significantly worse performance in an external validation cohort.[3] Unfortunately, the majority of published prediction research focuses solely on model derivation, and validation studies are scarce.[1,18]

Validation can be performed using internal or external validation. A common approach to internal validation is to split the data set into two portions—a "training set" and "validation set". If splitting the data set is not possible given the limited available data, measures such as cross validation or bootstrapping can be used for internal validation.[19] Machine learning algorithms, more specifically the random-forest approach, uses an alternative approach called—"in-bag" and "out-of-bag" sampling.[11] In a random-forest approach, the initial cohort is divided into two groups—"in-bag" and "out-of-bag" samples. The in-bag sample is created using random sampling with replacement from the initial cohort, creating a sample equivalent in size to the initial cohort. The out-of-bag sample is composed of the unsampled data from the initial cohort, and typically includes about one-third of the initial cohort. The "out-of-bag" cohort can serve as an internal validation cohort for the model derived using the "in-bag" sample. However, internal validation nearly always yields optimistic results given that the derivation and validation data sets are very similar (as they are from the same cohort). Although external validation is more difficult as it requires data collected from similar patients in a different setting or a different center, it is always preferred to internal validation.[1,18]

When a validation study shows disappointing results, researchers are often tempted to reject the initial model and to develop a new predictive model using the validation cohort data. For example, there are over 60 published predictive models for breast cancer. This approach neglects the information captured from prior studies and predictive models. There are several methods to update prior predictive models with data from the patients of the validation cohort, but these are unfortunately rarely utilized.[1]

ASSESSING THE PERFORMANCE OF A PREDICTIVE MODEL

When assessing model performance, it is important to remember that explanatory models are judged based on strength of associations, whereas predictive models are judged solely based on their ability to make accurate predictions. The performance of a predictive model is assessed using several complementary tests, which assess overall performance, calibration, discrimination, and reclassification (Table 1).[20] Performance characteristics should be determined and reported for both the derivation and validation data sets.

The overall model performance can be measured using R^2, which characterizes the degree of variation in risk explained by the model.[21] The adjusted R^2 has been proposed as a better measure, as it accounts for the number of predictors and helps to prevent overfitting. Brier scores are a similar measure of performance, which are used when the outcome of interest is categorical instead of continuous.[22] Calibration is the difference between observed and predicted event rates for groups of patients and is assessed using the Hosmer–Lemeshow test.[23] Discrimination is the ability of a model to distinguish between patients who do and do not experience the outcome of interest, and it is most commonly assessed using receiver operating characteristic (ROC) curves.[24] However, ROC analysis alone is relatively insensitive for assessing differences between good predictive models;[25] therefore, several relatively novel performance measures have been proposed. The net reclassification improvement and integrated discrimination improvement are measures used to assess changes in predicted outcome classification between two models.[20,26] Although it is common for prediction research studies to report results from ROC analysis, the other measures of model performance, calibration, and reclassification are seldom reported.[7,20]

STUDYING THE CLINICAL IMPACT OF A PREDICTIVE MODEL

The performance of a predictive model may suffer when applied in clinical practice compared with testing in derivation or validation data sets owing to differences in the patient

Table 1 Performance characteristics for a predictive model (measures of predicitve error)

Aspect	Measure	Outcome measure	Description
Overall performance	R^2	Continuous	Average squared difference between predicted and observed outcome
	Adjusted R^2	Continuous	Same as R^2, but penalizes for the number of predictors
	Brier score	Categorical	Average square distances from the predicted and the observed outcomes
Discrimination	ROC curve (c-statistic)	Continuous or categorical	Overall measure of how effectively the model differentiates between events and non-events
	C-index	Cox-model	
Calibration	Hosmer–Lemeshow test	Categorical	Agreement between predicted and observed risks
Reclassification	Reclassification table	Categorical[a]	Number of individuals that move from one category to another by improving the prediction model
	NRI		A quantitative assessment of the improvement in classification by improving the prediction model
	IDI		Similar to NRI but using all possible cutoffs to categorize events and non-events

IDI, Integrated discrimination index; NRI, net reclassification index.
[a]Can be performed for continuous data as well if a risk cutoff is assigned.

population and case mix.[27] The distribution of predictive factors and outcomes are often different when broadly applied to general populations, rather than the carefully selected populations in which the model was derived and validated. Furthermore, high model performance does not necessarily guarantee provider acceptance and uptake in clinical practice.[1] For example, providers may not use a predictive model because they feel that the application of the model is not sufficiently user-friendly or that the model itself does not have sufficient face validity.

Predictive models are developed with the goal of providing estimates of outcome probabilities to complement provider clinical intuition. They should ideally recommend decisions instead of simply providing risk estimates for an outcome. Predictive models that estimate risk without recommending particular decisions are less likely to change provider behavior and outcomes than those that translate risk into a decision recommendation.[27] With the growing implementation of electronic health records, predictive models can serve as the basis for electronic decision support tools with real-time risk assessments. Implementation of the predictive algorithm could be used to identify high-risk individual cases and transmit annotated data back to the provider, facilitating changes to their clinical assessment. If properly validated in several different populations, predictive algorithms could also form the basis for publicly available online risk calculators. Electronic predictive models are particularly attractive, as they can optimize user-friendliness and may be introduced quickly and cheaply, after implementation of an electronic health record system.

Impact studies serve to study the effect of predictive models on provider behavior and patient outcomes.[28] This is often done using a design that compares outcomes between providers provided with output from the predictive model to a control group without the predictive model. Although this is best done using a site-randomized controlled trial approach, this may also be assessed using a pre-post study design. A potential intermediate step using decision modeling techniques or Markov modeling can be used to estimate the potential consequences and benefits of using a predictive model. If this analysis does not reveal improved patient outcomes, this would obviate the need for formal impact studies.

EXAMPLE OF PREDICTIVE MODELING

An example of the analytic tools used in predictive modeling can be found in a recent publication examining the performance characteristics of predictive models for development of hepatocellular carcinoma among patients with cirrhosis.[3] In this study, the performance of a traditional regression model is compared with that of machine learning algorithms. This study highlights a couple of important concepts. First, external validation is crucial. Internal validation overestimated the performance of the models, and each has substantially worse performance when externally validated. Second, it is important to use a wide range of complementary methods to assess predictive model performance, not just ROC curve analysis. The machine learning algorithm and traditional regression analysis models had similar c-statistics using ROC curve analysis, but the machine learning algorithm, using random forest, outperformed the traditional regression model when using net reclassification improvement, integrated discrimination improvement, and misclassification tables.

CONCLUSIONS

Although predictive models cannot replace clinical judgment, they can provide objective estimates about the future course of an illness and serve as important adjuncts in clinical practice. For example, predictive models have been used to risk stratify patients with regard to readmission risk, allowing for early interventions to reduce readmissions. Although low-risk patients could be considered for early discharge, high-risk patients might be triaged to specialized hospital services, intensive outpatient case management, and earlier clinic visits post discharge. Such applications may be particularly important to maximize cost-effectiveness under the Accountable Care Organization model.[29] However, predictive models must be properly developed and also validated in a separate cohort using modern assessment of their performance.

Finally, the clinical impact of these predictive models must be prospectively assessed once implemented in clinical practice.

TAKE HOME POINTS

Prediction research may serve as an important adjunct to clinical practice.

Prediction research involves developing a predictive model, independently validating its performance, and prospectively studying its clinical impact.

CONFLICT OF INTEREST

Guarantors of the article: Akbar K. Waljee, MD, MS and Amit G. Singal, MD, MS.

Specific author contributions: Akbar K. Waljee and Amit G. Singal were involved in study concept and design, drafting of the manuscript, critical revision of the manuscript for important intellectual content, and study supervision. Peter D.R. Higgins - involved in critical revision of the manuscript for important intellectual content.

Financial support: Waljee's research is funded by a VA HSR&D CDA-2 Career Development Award 1IK2HX000775-01. Singal's research is funded by an ACG Junior Faculty Development Award and grant number KL2TR000453. The content is solely the responsibility of the authors and does not necessarily represent the official views of UT-STAR, the University of Texas Southwestern Medical Center at Dallas and its affiliated academic and health care centers, the National Center for Advancing Translational Sciences, the Veterans Affairs, or the National Institutes of Health.

Potential competing interests: None.

1. Toll DB, Janssen KJ, Vergouwe Y *et al.* Validation, updating and impact of clinical prediction rules: a review. *J Clin Epidemiol* 2008; **61**: 1085–1094.
2. Waljee AK, Joyce JC, Wang SJ *et al.* Algorithms outperform metabolite tests in predicting response of patients with inflammatory bowel disease to thiopurines. *Clin Gastroenterol Hepatol* 2010; **8**: 143–150.
3. Singal AG, Mukherjee A, Higgins PD *et al.* Machine learning algorithms outperform conventional regression models in identifying risk factors for hepatocellular carcinoma in patients with cirrhosis. *Am J Gastroenterol* 2013; **108**: 1723–1730.
4. Singal AG, Yopp AC, Gupta S *et al.* Failure rates in the hepatocellular carcinoma surveillance process. *Cancer Prev Res (Phila)* 2012; **5**: 1124–1130.
5. Singal AG, Rahimi RS, Clark C *et al.* An automated model using electronic medical record data to identify patients with cirrhosis at high risk for readmission. *Clinical Gastroenterol and Hepatol* 2013; **11**: 1335–1341.
6. Moons KG, Royston P, Vergouwe Y *et al.* Prognosis and prognostic research: what, why, and how? *BMJ* 2009; **338**: b375.
7. Hemingway H, Riley RD, Altman DG. Ten steps towards improving prognosis research. *BMJ* 2009; **339**: b4184.
8. Waljee AK, Higgins PD. Machine learning in medicine: a primer for physicians. *Am J Gastroenterol* 2010; **105**: 1224–1226.
9. Siegel CA, Siegel LS, Hyams JS *et al.* Real-time tool to display the predicted disease course and treatment response for children with Crohn's disease. *Inflamm Bowel Dis* 2011; **17**: 30–38.
10. Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013, ISBN 3-900051-07-0, http://www.R-project.org/.
11. Breiman L. Random forests. *Machine Learning* 2001; **45**: 5–32.
12. Liaw A, Wiener M. Classification and regression by random Forest. *R News* 2002; **2**: 18–22.
13. Royston P, Moons KG, Altman DG *et al.* Prognosis and prognostic research: developing a prognostic model. *BMJ* 2009; **338**: b604.
14. Greenland S. Modeling and variable selection in epidemiologic analysis. *Am J Public Health* 1989; **79**: 340–349.
15. Ibrahim JG, Chu H, Chen MH. Missing data in clinical studies: issues and methods. *J Clin Oncol* 2012; **30**: 3297–3303.
16. Kaambwa B, Bryan S, Billingham L. Do the methods used to analyse missing data really matter? An examination of data from an observational study of Intermediate Care patients. *BMC Res Notes* 2012; **5**: 330.
17. Waljee A, Mukherjee A, Singal A *et al.* Comparison of modern imputation methods for missing laboratory data in medicine. *BMJ Open* 2013; **3**: pii: e002847; doi:10.1136/bmjopen-2013-002847.
18. Altman DG, Vergouwe Y, Royston P *et al.* Prognosis and prognostic research: validating a prognostic model. *BMJ* 2009; **338**: b605.
19. Steyerberg EW, Harrell FE Jr, Borsboom GJ *et al.* Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol* 2001; **54**: 774–781.
20. Steyerberg EW, Vickers AJ, Cook NR *et al.* Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010; **21**: 128–138.
21. Gerds TA, Cai T, Schumacher M. The performance of risk prediction models. *Biom J* 2008; **50**: 457–479.
22. Czado C, Gneiting T, Held L. Predictive model assessment for count data. *Biometrics* 2009; **65**: 1254–1261.
23. Hosmer DW, Hosmer T, Le Cessie S *et al.* A comparison of goodness-of-fit tests for the logistic regression model. *Stat Med* 1997; **16**: 965–980.
24. Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. *Biometrics* 2005; **61**: 92–105.
25. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 2007; **115**: 928–935.
26. Pencina MJ, D'Agostino RB Sr, Demler OV. Novel metrics for evaluating improvement in discrimination: net reclassification and integrated discrimination improvement for normal variables and nested models. *Stat Med* 2012; **31**: 101–113.
27. Moons KG, Altman DG, Vergouwe Y *et al.* Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ* 2009; **338**: b606.
28. Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Ann Intern Med* 2006; **144**: 201–209.
29. Meek JA. Affordable Care Act: predictive modeling challenges and opportunities for case management. *Prof Case Manag* 2012; **17**: 15–21; quiz 22–23.