Published in final edited form as:

Eur Urol. 2010 April ; 57(4): 571–573. doi:10.1016/j.eururo.2009.12.026.

Prediction Models in Urology: Are They Any Good, and How Would We Know Anyway?

Andrew Vickers

Department of Epidemiology and Biostatistics Memorial Sloan-Kettering Cancer Center 1275 York Avenue Box 44 New York, NY 10065 USA

1. Models, models everywhere

Prediction models constitute one of the most common categories of research papers in urology. Indeed, a review of literature published in July 2007 found >100 different prediction tools for prostate cancer (PCa) alone [1]. Even within the very narrow research area of Lughezzani et al's paper [2] on outcomes after radical prostatectomy, models have been published to predict recurrence based on preoperative [3] and postoperative [4] data; recurrence at 2 yr [5]; recurrence at 5 yr incorporating transforming growth factor- β 1 [6]; recurrence at 5 yr including surgeon experience in the model [7]; the trifecta of continence, potency, and freedom from recurrence [8]; PCa-specific death at 10 yr [9]; and so on and so forth. PCa prediction modeling has become so complex that a “meta-literature” has developed around choosing appropriate prediction tools [10,11].

My guess is that few of the published prediction models are much good, and even if they were, we wouldn't know it.

2. Prediction models 101: the importance of independent validation

The first problem with most prediction models is the lack of independent validation. One of the most basic concepts of prediction modeling is that models need to be developed and tested on different data sets. This approach is partly to avoid a problem known as *overfit*. As a trivial example, 2 d ago, I spent 45 min traveling home from work; yesterday, the trip took me 47 min. A model based on this data set (time: 46 min) predicts travel time in the “training” data set to within 30 s. Today, however, there was train trouble, and my journey took 1 h, meaning that my model was useless when applied in practice.

There are statistical methods to correct for overfit, but in my view, they do little more than breed a false sense of security. This is because a model developed based on one set of patients might be inappropriate for another set of patients, not just for statistical reasons but due to substantive differences in the data. Just as my travel time data set is a poor predictor for my colleague, who walks to work from a few blocks away, a data set of patients treated before the stage shift in PCa may not work well for patients treated today. A Gleason 7 in 1991 is not the

© 2009 European Association of Urology. Published by Elsevier B.V. All rights reserved.

T: 646.735.8142 F: 646.735.0011 vickersa@mskcc.org

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conflicts of interest

The author has nothing to disclose.

same as a Gleason 7 in 2009 [12]; moreover, many contemporary surgeons have much greater levels of surgical experience than surgeons did 20 yr ago, and surgical experience has a large influence on recurrence risk [13]. Similarly, a “PCa risk calculator” based on data from a population-based screening study in Europe [14] may not apply to a patient presenting at a US clinic, who has a different history of prostate-specific antigen testing, who may receive more extended biopsy, and who has likely received a work-up to exclude benign disease.

3. Calibration? Discrimination? So what?

The second major problem with the typical paper describing a urology prediction model is that the results are often very difficult to interpret. Papers on prediction models typically report a measure of discrimination such as the area under the curve or its equivalent for time-to-event data, the concordance index. These measures can be interpreted as the probability that, for a randomly selected pair of patients, the patient with the event (ie, shorter survival) is given a higher risk by the prediction model. Measures of discrimination therefore range from 0.5 (chance) to 1 (perfect discrimination). But just how high does discrimination need to be in order to be “high enough” to justify use of a prediction model? Is 0.70 sufficient? What about 0.60? Moreover, despite often being described as a measure of “predictive accuracy,” in fact, discrimination is entirely unaffected by whether or not a prediction is a good one. The wellknown Kattan nomogram [15], for example, gives the probability of recurrence after radical prostatectomy. For the purposes of this paper, I would like to propose a new “Vickers score.” To calculate a Vickers score, you divide the Kattan score by 100 and add 0.99. Thus, a patient with a recurrence risk from the Kattan nomogram of 10% would be given a Vickers risk of 99.1%, and a patient with a Kattan risk of 20% would be given a Vickers risk of 99.2%. The key point is that the Vickers score and the Kattan nomogram would have the same concordance index because patients given a risk of 99.2% do indeed tend to have a higher risk of recurrence than those given a risk of 99.1%. I could claim that my model had “high predictive accuracy,” even though I would tell a low-risk patient that he would almost definitely recur.

The problem with the Vickers score is known as *miscalibration*: A model is well calibrated if for every 100 patients given a risk of $x\%$, close to $x\%$ actually have the event. Calibration is normally illustrated graphically in terms of a calibration plot, such as Figure 2 in Lughezzani et al's paper [2]. The problem is, how much miscalibration is too much? Numerical estimates of calibration are questionable and tend not to be reported, leaving investigators to describe calibration in rather vague terms (eg, “because the dots are relatively close to the dashed line, the predictions calculated with use of our nomogram approximate the actual outcomes” [15]).

4. A new approach

Lughezzani et al's paper [2] breaks the mold of urology prediction models in two key ways. First, it constitutes an independent validation study. Three previously published models were applied to an entirely new set of data, so the results give us a good idea of how the models would perform in clinical practice. Second, the investigators go beyond discrimination and calibration, using a decision analytic technique to evaluate the prediction models. The basis of decision analysis is to evaluate prediction models by their implications for decision making: If a model helps you make better decisions, it is a good one and should be used in practice. The key thing about decisions is that they have consequences, and these consequences need to be weighted in some way. As a simple example, in the case of prostate biopsy, it is far worse to miss a cancer than to avoid an unnecessary biopsy. Decision analysis puts a number on just how much worse, and that approach allows the analyst to compare different models and state which would lead to better results if used in clinical practice.

“Putting a number” on consequences can be a hard thing to do. To know how much worse it is to miss a cancer than to conduct an unnecessary biopsy, we would need good data for the

marginal increase in risk associated with a delay in PCa diagnosis as well as estimates for the pain, inconvenience, and risk of biopsy. Lughezzani et al use a simple shortcut based on a number used implicitly or explicitly by all of us every day: the threshold probability [2]. As an illustration, if a man had a threshold probability for prostate biopsy of 20%, he would agree to biopsy if his risk were $\geq 20\%$ but not if it were $< 20\%$. The interesting thing about threshold probability is that it tells us about the relative value of true and false positives. Colleagues of mine say that they would do a lymph node dissection during a radical prostatectomy, even if the patient had only a 2% chance of lymph node invasion, on the ground that a lymphadenectomy does not present important risks to the patient but leaving cancer behind would be a disaster. Alternatively, few urologists would do more than four or five biopsies to find one cancer (a risk threshold of around 15[en]20%) because biopsy is an invasive procedure and a PCa not detected today is unlikely to become incurable very quickly.

The decision analysis presented by Lughezzani et al in Figure 3 [2] has two main advantages. First, it gives a result based on clinical outcomes rather than an abstraction like the concordance index. (When is it the task of a clinician to guess which of two patients had an event?) Second, it gives a clear result: The Cancer of the Prostate Risk Assessment (CAPRA) model has the highest net benefit across a wide range of threshold probabilities and so will lead to the best decisions in clinical practice. There is no need to subjectively judge calibration curves or to decide whether one model's better calibration is offset by its poorer discrimination.

I cannot agree with the conclusions of Lughezzani et al [2] (although I have reason to believe that these were foisted on them by recalcitrant reviewers). This is because I do not think it is at all difficult to “reconcile concordance index, calibration and decision curve analysis findings.” The Stephenson model suffers from poor calibration, D'Amico from moderate discrimination, so CAPRA comes out best in the decision analysis. Similarly, I do not believe that discrimination alone can be used as a benchmark. After all, the grossly inaccurate Vickers score comes out pretty well in terms of discrimination. Critically, however, the authors have given us the information to judge for ourselves, based on a sound study design. Let's hope that future evaluations of prediction models follow suit.

Acknowledgments

Funding support

This paper was supported in part by funds from David H. Koch provided through the Prostate Cancer Foundation, the Sidney Kimmel Center for Prostate and Urologic Cancers, and P50-CA92629 SPORE grant from the National Cancer Institute to Dr. P. T. Scardino.

References

1. Shariat SF, Karakiewicz PI, Roehrborn CG, Kattan MW. An updated catalog of prostate cancer predictive tools. *Cancer* 2008;113:3075–99. [PubMed: 18823041]
2. Lughezzani G, Budäus L, Isbarn H, et al. Head-to-head comparison of the three most commonly used preoperative models for prediction of biochemical recurrence after radical prostatectomy. *Eur Urol*. In press.
3. Stephenson AJ, Scardino PT, Eastham JA, et al. Preoperative nomogram predicting the 10-year probability of prostate cancer recurrence after radical prostatectomy. *J Natl Cancer Inst* 2006;98:715–7. [PubMed: 16705126]
4. Stephenson AJ, Scardino PT, Eastham JA, et al. Postoperative nomogram predicting the 10-year probability of prostate cancer recurrence after radical prostatectomy. *J Clin Oncol* 2005;23:7005–12. [PubMed: 16192588]
5. Walz J, Chun FK, Klein EA, et al. Nomogram predicting the probability of early recurrence after radical prostatectomy for prostate cancer. *J Urol* 2009;181:601–7. discussion 607[en]8. [PubMed: 19084864]

6. Shariat SF, Karam JA, Walz J, et al. Improved prediction of disease relapse after radical prostatectomy through a panel of preoperative blood-based biomarkers. *Clin Cancer Res* 2008;14:3785–91. [PubMed: 18559597]
7. Kattan MW, Vickers AJ, Yu C, et al. Preoperative and postoperative nomograms incorporating surgeon experience for clinically localized prostate cancer. *Cancer* 2009;115:1005–10. [PubMed: 19156928]
8. Eastham JA, Scardino PT, Kattan MW. Predicting an optimal outcome after radical prostatectomy: the trifecta nomogram. *J Urol* 2008;179:2207–10. discussion 2210[en]1. [PubMed: 18423693]
9. D'Amico AV, Moul J, Carroll PR, Sun L, Lubeck D, Chen MH. Cancer-specific mortality after surgery or radiation for patients with clinically localized prostate cancer managed during the prostate-specific antigen era. *J Clin Oncol* 2003;21:2163–72. [PubMed: 12775742]
10. Nguyen CT, Kattan MW. Development of a prostate cancer metagram: a solution to the dilemma of which prediction tool to use in patient counseling. *Cancer* 2009;115(suppl):3039–45. [PubMed: 19544545]
11. Ross RW, Kantoff PW. Predicting outcomes in prostate cancer: how many more nomograms do we need? *J Clin Oncol* 2007;25:3563–4. [PubMed: 17704399]
12. Egevad L. Recent trends in Gleason grading of prostate cancer: I. Pattern interpretation. *Anal Quant Cytol Histol* 2008;30:190–8. [PubMed: 18773736]
13. Vickers AJ, Bianco FJ, Serio AM, et al. The surgical learning curve for prostate cancer control after radical prostatectomy. *J Natl Cancer Inst* 2007;99:1171–7. [PubMed: 17652279]
14. van den Bergh RC, Roobol MJ, Wolters T, van Leeuwen PJ, Schroder FH. The Prostate Cancer Prevention Trial and European Randomized Study of Screening for Prostate Cancer risk calculators indicating a positive prostate biopsy: a comparison. *BJU Int* 2008;102:1068–73. [PubMed: 18715246]
15. Kattan MW, Wheeler TM, Scardino PT. Postoperative nomogram for disease recurrence after radical prostatectomy for prostate cancer. *J Clin Oncol* 1999;17:1499–507. [PubMed: 10334537]