Prediction Models in Cancer Care

Andrew J. Vickers, PhD

Abstract

Prediction is ubiquitous across the spectrum of cancer care from screening to hospice. Indeed, oncology is often primarily a prediction problem; many of the early stage cancers cause no symptoms, and treatment is recommended because of a prediction that tumor progression would ultimately threaten a patient’s quality of life or survival. Recent years have seen attempts to formalize risk prediction in cancer care. In place of qualitative and implicit prediction algorithms, such as cancer stage, researchers have developed statistical prediction tools that provide a quantitative estimate of the probability of a specific event for an individual patient. Prediction models generally have greater accuracy than reliance on stage or risk groupings, can incorporate novel predictors such as genomic data, and can be used more rationally to make treatment decisions. Several prediction models are now widely used in clinical practice, including the Gail model for breast cancer incidence or the Adjuvant! Online prediction model for breast cancer recurrence. Given the burgeoning complexity of diagnostic and prognostic information, there is simply no realistic alternative to incorporating multiple variables into a single prediction model. As such, the question should not be whether but how prediction models should be used to aid decision-making. Key issues will be integration of models into the electronic health record and more careful evaluation of models, particularly with respect to their effects on clinical outcomes. CA Cancer J Clin 2011;61:315-326. © 2011 American Cancer Society, Inc.

Introduction: Cancer as a Prediction Problem

Prediction is ubiquitous in oncology; innumerable decisions by patients, family members, oncologists, and other care providers depend on assessing the likelihood of future events. This can be seen across the spectrum of cancer care, from screening to hospice. Screening is recommended for those at an elevated risk of cancer, either because of age (e.g., older but not younger individuals are advised to consider colonoscopy) or risk factors, such as lung imaging in those with a significant smoking history. Prevention is similarly risk stratified; for example, prophylactic mastectomy might be considered for a woman with a strong genetic risk of breast cancer but certainly not for a woman at average risk. Prediction also dominates clinical care of the cancer patient, with decisions about surgery, chemotherapy, and radiotherapy being heavily risk-dependent; simple examples include radical cystectomy versus transurethral resection in bladder cancer, single-agent versus combination chemotherapy for lymphoma, and adjuvant chemotherapy versus monitoring after resection for gastric cancer. For patients with advanced disease, decisions about life-extending versus palliative care inherently involve a survival prediction.

Indeed, it might be argued that cancer care is often primarily about prediction. For many cancers (breast, prostate, colon, and bladder would be obvious examples), early stage disease causes no important symptoms; treatment is recommended because of a prediction that a tumor would progress and threaten a patient’s quality of life or survival.

Recent years have seen attempts to formalize risk prediction in cancer care. In place of informal and implicit prediction algorithms, such as cancer stage, researchers have developed statistical prediction tools that provide a...
quantitative estimate of the probability of a specific event for an individual patient. In this article, I offer an overview of prediction modeling in cancer care. I will start by describing the purported advantages of prediction modeling. I will then discuss evaluation of prediction models, describe some well-known prediction models, and conclude with some current controversies.

On a terminological note, estimation of the future course of disease is often described as prognosis, with some authors making a distinction between prognosis and prediction. In this formulation, prognosis refers to how the patient is likely to progress on standard treatment, such as the probability of recurrence after resection for breast cancer, whereas prediction is reserved for describing how the patient will respond to a particular treatment, such as the relationship between outcomes after treatment with trastuzumab and human epidermal growth factor receptor 2 (HER2)/neu overexpression. Moreover, a distinction is also commonly made between diagnosis and prediction, with the former reserved for states that can be known immediately, such as the presence of cancer on biopsy, and the latter for events that only occur with the passage of time, such as cancer recurrence. For the purposes of this article, I will use prediction to refer to any estimate concerning an unknown future event, including, for example, the result of a planned biopsy. The reason is more than just simplicity of presentation. The scientific, statistical, and practical characteristics of prediction, prognosis, and diagnosis are highly comparable.

**What Is a Prediction Model?**

The simplest way to think about prediction models is in terms of the high school math exercise relating height to shoe size. First, data on height and shoe size are recorded for each member of the class. A graph is then drawn plotting shoe size on the x axis against height on the y axis, placing a dot for each student’s data. Finally, a “line of best fit” is drawn to come close to each of the points. A typical formula for this line might be: \( y = 1.65x + 55 \). Where the line is calculated by formal mathematical methods, this formula is a regression model and can be thought of in terms of a prediction: tell me your shoe size and I can multiply it by 1.65 and add 55 to make a prediction about your height. The \( y \) variable (height in inches) is what we are trying to predict and is known as the dependent variable; the \( x \) variable (shoe size) is known as the predictor.

Medical prediction modeling is directly analogous to this simple classroom example. For mathematical reasons, we do not directly predict probability from the formula, but instead obtain what is known as the logit, the logarithm of the odds of the event. As an example, a formula for the risk of prostate cancer on biopsy given a certain prostate-specific antigen (PSA) level is: \( \log \text{odds of cancer} = \text{PSA level} \times 0.109 - 1.74 \). A man with a PSA of 5 ng/mL therefore has a log odds of \( 5 \times 0.109 - 1.74 = -1.195 \). The formula for converting a log odds into a probability is \( 1 / (1 + e^{-\log \text{odds}}) \), and therefore a PSA of 5 ng/mL gives a predicted probability of 23%. The number used to multiply PSA in the formula (0.109) is known as the coefficient. What is typically reported in scientific articles is the odds ratio; the coefficient is the logarithm of the odds ratio. For predicting the probability of an event in the future, such as the risk of cancer recurrence within 5 years, statisticians often use what is known as Cox modeling. The basic principles are the same: the logarithm of the hazard ratio is the coefficient, which is entered into a formula to calculate risk.

Both the height and prostate cancer models include only a single predictor and are therefore called univariate models. However, it is also possible to use multiple predictors, what is known as multivariable prediction. For example, we can calculate the risk of prostate cancer based on both the PSA level and the digital rectal examination (DRE): \( \log \text{odds of cancer} = 0.096 \times \text{PSA level} + 1.16 \times \text{DRE} - 2.00 \). This formula can be used to show that our patient with a PSA of 5 ng/mL has an 18% risk of cancer if he had a negative DRE, but a 41% risk if the DRE was positive.

**The Rationale for Prediction Modeling**

Perhaps the most well-known prediction model in medicine is the Framingham Risk Calculator. The physician records the patient’s age, gender, total and high-density lipoprotein cholesterol, smoking history, blood pressure, and blood pressure medication; these values are then entered into an online
calculator (available at http://hp2010.nhlbihin.net/atpiii/calculator.asp?usertype=prof) to estimate the probability that the patient will experience a coronary event within 5 years. Paper-based methods for implementing the Framingham model are also available.

The key insight of the Framingham model is that risk depends on multiple factors such that treating individual risk factors in isolation is suboptimal. For example, hypertension is a risk factor for coronary disease, and it is customary to recommend medication to reduce blood pressure to a patient meeting the definition for hypertension (systolic blood pressure of 140 mm Hg or above). This approach is sometimes described as “risk categorization”: patients are placed in categories of “high,” “intermediate,” and “low” risk and treated accordingly.

Figure 1 uses a cancer example to illustrate why risk prediction will generally be preferable to risk categorization. Risk categorization assumes discontinuities in risk at specific cutpoints, with all patients within a risk category at similar risk. This can be a reasonable assumption at low risk but can be misleading at intermediate and high risk. In the graph, for example, risk in patients in the highest category of smoking history varies approximately 3-fold.

**Prediction Modeling Improves Predictive Accuracy**

Medical events are very often associated with multiple risk factors. Considering risk factors in isolation means ignoring relevant information and will typically lead to a decrease in predictive accuracy. In the case of the Framingham model, blood pressure is associated with coronary events, but risk also depends on age, gender, smoking history, and cholesterol level. A young woman with few risk factors other than a systolic blood pressure of 145 mm Hg is at very low risk for a serious cardiac event, and her level of risk is barely affected by a change in blood pressure. Conversely, an older male smoker with elevated cholesterol and a similar blood pressure is at high risk, and he would have a substantially decreased risk if his blood pressure could be reduced. The prediction modeling approach leads to antihypertensive medication being recommended to only one of the 2 patients with hypertension.

There is considerable evidence that incorporating multiple variables into cancer prediction models provides more accurate predictions than simple risk classification. Multivariable models have been shown to be more accurate than clinical stage for predicting disease-specific survival in patients with gastric cancer, recurrence after cystectomy and colectomy, and sentinel lymph node metastasis in patients with melanoma.

It might be pointed out that, in practice, clinicians rarely do consider risk factors in isolation. In the hypertension example given above, a physician would be more likely to treat the older male smoker than the younger, otherwise healthy female nonsmoker even without access to the Framingham prediction model. However, these sorts of subjective predictions would prima facie seem rather challenging, because they require physicians to integrate a
diverse range of quantitative information without allowing bias or wishful thinking to influence their conclusions. There is empirical evidence that the subjective predictions made by physicians are inaccurate. One common focus of research is prediction of survival near the end of life. A systematic review of 8 end-of-life studies found that physicians overestimated life expectancy by an average of 50%, with approximately one-quarter of cancer patients dying at least 4 weeks before their expected date of death.7 It has also been demonstrated that use of a prediction model can improve end-of-life predictions. In a study of hospice patients, physicians were asked to make survival predictions with or without reference to a prediction model. Large differences between predicted and observed survival were reduced by 25% to 50% when the prediction model was used.8

Researchers have also investigated subjective predictions associated with specific clinical decisions. One study compared clinician and statistical predictions of nonsentinel lymph node status after positive sentinel lymph node biopsy for breast cancer. Clinicians were presented with the characteristics of 33 patients, including details of age, tumor size, nuclear grade, hormone receptor status, and number of positive and negative sentinel lymph nodes. They were then asked to give the probability that the patient would be found to have additional positive lymph nodes on axillary lymph node dissection. Clinicians' predictive accuracy was only slightly better than chance, and far inferior to a predictive model.9

**Prediction Models Allow Incorporation of Novel Predictors**

Several risk factors for cardiovascular disease have been discovered since the Framingham study first reported on smoking, cholesterol, and blood pressure. These include clinical indicators such as family history or kidney disease, and molecular markers such as C-reactive protein. It is not entirely clear how these new risk factors should be used clinically in the absence of a risk prediction model.

A comparable example in oncology concerns prostate biopsy. Risk factors for prostate cancer include elevated PSA, a low free-to-total PSA ratio, and a positive DRE. Some clinicians also believe that PSA velocity is predictive, although this is somewhat controversial.10 In addition, a new marker, PCA3, has recently been developed.11 Assuming that a simple cutpoint is used for each of the 5 risk factors, this gives $2^5 = 32$ combinations (eg, low PSA/normal DRE/low free-to-total PSA ratio/high PSA velocity/low PCA3; low PSA/abnormal DRE/low free-to-total PSA ratio/low PSA velocity/high PCA3, etc). Combining risk factors in this way leads to algorithms of bewildering and increasing complexity. For instance, if 2 new serum measures and 2 genomic markers were discovered, the number of diagnostic categories would increase to 512. One approach to deal with complexity is simply to recommend treatment if any risk factor is present. Such an approach is taken by the National Comprehensive Cancer Network (NCCN) guidelines on prostate cancer detection,12 which suggest biopsy if PSA is elevated, or if DRE positive, or if the free-to-total PSA ratio is low, or if PSA velocity is high. This can be a problem because unless each test has near-perfect specificity, use of multiple tests dramatically increases the number of false-positive results. As an example, imagine that there are 6 independent risk factors for a cancer and that each has a very high specificity of 90%. To avoid biopsy, a patient would need to be negative for all 6 risk factors, the probability of which is $0.9^6 = 0.53$. In other words, if biopsy is recommended if any risk factor is present, approximately one-half of the population would end up with an unnecessary biopsy, even if the risk factors were very specific for prostate cancer.

**Prediction Models Aid the Rational Choice of Cutpoints for Decision-Making**

Although risk is a continuum (from very low to very high risk, with shades of gray in between), clinicians ultimately have to make binary decisions, such as whether or not to treat a patient. This entails consideration of how high a risk is “high enough” to warrant treatment. In the case of heart disease, a systolic blood pressure of 140 mm Hg or greater is typically thought to constitute hypertension, and thus indicates a prescription of an antihypertensive. Alternatively, a Framingham score of 10% or greater risk of a coronary event can be used to determine therapy. The rationale for clinical cutpoints such as 140 mm Hg is often somewhat obscure.13 Moreover, such cutpoints are not amenable to reasoned debate.
If a clinician wanted to lower the cutpoint for medication to 135 mm Hg, it is unclear what arguments could be made for or against. In contrast, a clear rationale can be given for a risk cutpoint, and it is possible to debate what a rational risk cutpoint should be. Decision theory\textsuperscript{14} indicates that a risk cutpoint depends on the relationship between the harms of an event and the harms of unnecessary treatment. A cutpoint of 10%, for example, implies that it is 9 times worse to have an unnecessary coronary event than to undergo years of medical therapy unnecessarily. If new data emerged as to the risks and side effects of antihypertensives, or if dramatically improved treatments for myocardial infarct were developed, this risk-benefit ratio would change, and a different cutpoint might be selected.

In the case of cancer, there are reasons to believe that not only were current clinical cutpoints selected suboptimally, but that clinicians can indeed make rational decisions using risk-based cutpoints. As an obvious example, the use of stage to determine adjuvant chemotherapy is often questionable. For instance, there is no obvious biologic rationale why adjuvant therapy is indicated for gastric cancer if the tumor invades into the muscularis propria layer (T2 disease) but not if it extends only into the muscularis mucosa (T1a disease). Another example comes from cancer screening. For many years, urologists used a PSA cutpoint of 4 ng/mL to determine referral for prostate biopsy. However, the origin of this cutpoint is unknown.\textsuperscript{13} With respect to choice of cutpoints, compare a cited cutpoint for prostate biopsy of approximately 20%\textsuperscript{15} with an approximately 2% cutpoint for lymph node dissection during radical prostatectomy.\textsuperscript{16} These markedly different cutpoints are undoubtedly related to a careful consideration of risks and benefits: prostate biopsy is an invasive procedure, with non-negligible risks, that can often be postponed to allow additional monitoring without undue risk that a cancer will become incurable; in contrast, lymph node dissection is not a particularly morbid procedure, and there are important consequences of failing to resect an affected lymph node both in terms of staging and the likelihood of surgical cure.

Prediction models also allow individualized decision-making. A point related to choice of cutpoints is treatment individualization. In the case of heart disease, imagine that a patient with hypertension is averse to medication. One might imagine that a physician would accede to careful monitoring if the patient’s blood pressure were 141 mm Hg, but insist on treatment if the patient’s blood pressure were 195 mm Hg. The problem then becomes how high a blood pressure is high enough to override concerns about drug tolerability. There is simply no basis for discussion. In contrast, an estimate of risk can be considered against the disbenefits of treatment and a rational decision taken. As an example from oncology, Sonpavde describes a case of a breast cancer patient experiencing poor tolerability to adjuvant chemotherapy. The physician counsels the patient that, on the basis of a prediction model, adjuvant chemotherapy is associated with a 2% decrease in her risk of recurrence. The woman decides that this small benefit is not worth the side effects and decides to forgo further treatment.\textsuperscript{17}

Despite the clear and obvious advantages of prediction modeling for decision-making, the use of cutpoints remains disturbingly common in oncology, even for novel markers based on cutting-edge technology. For example, the prostate cancer marker PCA3 is based on transcription-mediated amplification of messenger RNA, yet is reported as a score ranging from 4 to 125, with patients scoring 35 or more deemed to be at high risk.

**Evaluation of Prediction Models**

Two key aspects of prediction model performance are discrimination and calibration. Discrimination refers to the ability of a prediction model to distinguish between patients. A typical measure of discrimination is the concordance index, which is often described in the diagnostic setting as the area under the receiver operating characteristic (ROC) curve or AUC. The concordance index gives the probability that, for a randomly selected pair of patients, the model gives a higher probability to the patient who had the event, or who had shorter survival. As such, the concordance index is measured on a scale ranging from 0.5 (no better than chance) to 1 (perfect prediction). Calibration refers to whether a prediction for an individual patient is close to his or her true risk. A model is well calibrated if for every 100 patients given a risk of $p\%$, close to $p$ have the event. Calibration is typically assessed by plotting expected against observed probabilities; good models are close to the
The population is divided into equally spaced categories of predicted risk ("quantiles"). The observed risk within each category is then calculated and plotted as a point estimate and 95% confidence interval. The dashed line is a regression line comparing predicted and observed risk. The solid 45° line indicates perfect calibration. A statistical model to predict prostate cancer recurrence within 5 years based on stage, grade, and prostate-specific antigen was created using a data set on patients treated during the stage shift (1987-1995). The figure on the left shows that this model is well-calibrated when applied to a cohort treated up to 1995, with close concordance between predicted and observed risk; the panel on the right shows poor calibration when the model is applied to patients treated after the stage shift (1996 and beyond). The model overestimates risk for most patients; for example, of the patients given a risk close to 20%, only approximately 10% actually developed recurrence. This miscalibration occurs for a number of reasons, including stage shift, changes in grading practice, and improvements in surgical technique.
Methodologists have recently demonstrated that traditional approaches to model evaluation have serious limitations and have developed novel statistical methods as alternatives. With respect to discrimination, the concordance index has limited clinical interpretability. If a model has a concordance index of, for example, 0.70, is that good or bad? Or take the common question of whether a novel molecular marker is of value. If use of a marker increases the concordance index of a model from 0.70 to 0.72, is that sufficient to make it worth measuring the marker? There is a comparable problem with calibration: how much miscalibration is “too much” to entail that a model should not be used? A final problem concerns how to integrate calibration and discrimination. For example, if model A had superior discrimination but poorer calibration than model B, it is unclear which model should be used in clinical practice.

Recent methodologic developments have moved beyond calibration and discrimination to evaluate models in more clinically relevant terms. Reclassification metrics assume that different patient decisions are taken on either side of a risk threshold and then examine whether these decisions would be good or bad ones. As a simple example, take a model to predict the risk of prostate cancer in men with elevated PSA. Table 1 provides hypothetical data from a study of 2 models assuming that men with a 20% or greater risk of prostate cancer are referred to biopsy. It can be seen that use of the model dramatically decreases the number of unnecessary biopsies at the cost of a small increase in missed cancers. To claim that the model was not of benefit, and that we should continue the current approach of biopsying all men with elevated PSA, suggests that it would be worth biopsying 450 men to find just 25 prostate cancers.

Interpretation of the classification table is a little more difficult when comparing the standard model with the model incorporating the new predictor; it is not immediately clear whether it would be justified to biopsy 35 men to find one cancer. A simple decision analytic solution has recently been proposed. This is based on the decision-theoretic principle that the probability cutpoint at which a physician would advise biopsy (pt) is informative of how the physician and patient weigh the harms of false-positive results (unnecessary biopsy) in comparison with the harms of false-negative results (delayed cancer diagnosis). Specifically, these relative harms are equivalent to the odds at the threshold. Therefore, a cutpoint of 20% implies that a delayed cancer is 4 times more harmful than an unnecessary biopsy. This can be used in a formula for a “net benefit”: true positive results = false positive results × p0 ÷ (1 - p0). The net benefit can be thought of as directly analogous to profit. Imagine that an investor buys goods in France using euros and sells them in the United States for dollars. Just as net profit is dollars - euros × exchange rate, net benefit is gain - loss × conversion factor. The interpretation of net benefit is very simple: the model with the highest net benefit should be chosen. This is again directly analogous to profit: investors choose the most profitable strategy, irrespective of the size of the difference in profit. Table 1 shows calculation of net benefit for each possible strategy (biopsy all men, biopsy on the basis of the standard prediction model, biopsy on the basis of the model plus new marker, and no biopsy). It is clear that the model including the new marker has the highest net benefit and will therefore lead to the best clinical outcomes.

**TABLE 1. Hypothetical Data From a Prostate Cancer Study**

| APPROACH                  | NO. OF BIOPSIES PER 1000 MEN WITH ELEVATED PSA | TRUE-POSITIVE RESULTS (CANCERS FOUND) | FALSE-POSITIVE RESULTS (UNNECESSARY BIOPSIES) | NET BENEFIT CALCULATION | NET BENEFIT |
|---------------------------|-----------------------------------------------|---------------------------------------|-----------------------------------------------|-------------------------|-------------|
| Biopsy all men            | 1000                                          | 250                                   | 750                                           | 250 - 750 × 0.2 ÷ (1 - 0.2) | 62.5        |
| Standard prediction model  | 525                                           | 225                                   | 300                                           | 225 - 300 × 0.2 ÷ (1 - 0.2) | 150         |
| Prediction model with additional marker | 490                                         | 220                                   | 270                                           | 220 - 270 × 0.2 ÷ (1 - 0.2) | 152.5       |
| No biopsy                 | 0                                             | 0                                     | 0                                              | 0 - 0 × 0.2 ÷ (1 - 0.2) | 0           |

Abbreviations: PSA, prostate-specific antigen.
Different physicians and patients can disagree about the appropriate risk threshold. A man averse to invasive procedures might require a high probability (eg, 30%) before he would accede to biopsy; a conservative physician might err on the safe side and use a lower threshold (such as 15%). It is possible to plot net benefit for different risk thresholds in what is known as a decision curve. Figure 4 shows an example decision curve. A new prediction model has the highest net benefit across a range of reasonable threshold probabilities. Referring to biopsy on the basis of this model will therefore lead to better patient outcomes than the alternative strategies of biopsying all men, biopsying no men, or biopsying on the basis of the standard model.

Examples of Cancer Prediction Models

The Prostate Cancer Prevention Trial Risk Calculator

Each year, many millions of men undergo PSA testing for the early detection of prostate cancer. It has been amply demonstrated that risk rises continuously with PSA; there is no discontinuity at a particular PSA cutpoint such that men just above the cutpoint have importantly higher risk than men just below it. It is also known that other factors, such as age, race, family history, DRE, and prior negative biopsy, are predictive of risk. Moreover, there is widespread agreement that patient preferences about prostate biopsy vary widely, with some men having above-average anxiety about prostate cancer, and others having relatively greater aversion to biopsy.

A prediction model for prostate cancer meets each of these 3 challenges: risk is modeled as a continuous function of PSA; the risk prediction incorporates other risk factors such as family history; and the output of the prediction model is a probability that can be discussed with the patient, and a decision made taking into account personal preference. The Prostate Cancer Prevention Trial (PCPT) risk calculator was developed using data from the placebo arm of a randomized trial of finasteride for prostate chemoprevention. The great advantage of this study is that patients underwent a protocollized prostate biopsy at the end of the study irrespective of indication, thus avoiding verification bias. Several studies based on US populations have confirmed the accuracy of the PCPT risk calculator. However, not all studies have been favorable, and a recent meta-analysis of multiple prostate biopsy cohorts suggests that the properties of the model are likely to be somewhat cohort-dependent.

The Gail Model

A common medical rule of thumb is that treatment should focus on those patients with above-average risk. This naturally begs the question of how risk is assessed, and what counts as average. The Gail model predicts a woman’s risk of invasive breast cancer within 5 years on the basis of age, race, family history, age at first menses, age of first live birth, and relevant medical history (eg, prior negative breast biopsy). An online version of the prediction model is available at http://www.cancer.gov/bcrisktool/. The Gail model was used as an inclusion criterion in several trials of breast cancer chemoprevention, where patients were deemed eligible if their risk on the Gail model was at or above the median (1.66%). The positive results of these trials have lead to the recommendation that “clinicians discuss chemoprevention with women at high risk for breast cancer” and “use information [from the Gail model] to help individual patients considering tamoxifen therapy estimate the potential benefit.”

FIGURE 4. Hypothetical Decision Curve Analysis for Cancer Prediction Models. A new model (thick black line) has the highest net benefit across a wide range of risk thresholds when compared with the standard model (dashed line), a strategy of treating all patients (straight black line), or a strategy of treating no patients (thick gray line). Use of the new model would therefore improve medical decision-making. The standard model is sometimes inferior to a strategy of just treating all patients, providing evidence of miscalibration.
The Kattan Nomogram

Prediction models require a complex calculation to derive a predicted risk from values of the relevant predictors. This is likely why prediction models have not historically played an important role in medicine; in comparison with determining whether a patient is above or below a certain threshold (such as PSA of 4 ng/mL), a prediction model requires the clinician to ascertain the values of multiple predictors, multiply each predictor by a different coefficient, sum, add a constant, and finally transform the resulting log odds to a probability. Information technology has recently made such calculations a trivial task. Before the widespread availability of Internet-connected computers in the clinic, a graphical calculating device often described as a nomogram allowed ready calculation of predicted probabilities.

One of the first widely publicized nomograms for medical prediction was that of Kattan et al for the prediction of recurrence after radical prostatectomy, based on clinical stage, tumor grade, and PSA.35 The clinician uses the figure to calculate points for each predictor; these are summed and the total number of points compared against risk.

Adjuvant! Online

The results of randomized trials are often given as an absolute risk reduction or the number needed to treat (NNT). The results of a typical trial of adjuvant therapy for breast cancer might be described in terms of a 25% death rate in the surgery-only group versus 20% in women also undergoing adjuvant chemotherapy, giving an absolute risk reduction of 5% and an NNT of 20. What is not widely recognized, however, is that this NNT is an average across risk groups. As such, individual women can have dramatically different risk reductions depending on whether, for example, they have a large, high-grade, estrogen receptor-positive tumor that has spread to the lymph nodes, in comparison with a small, low-grade, lymph node-negative, hormone receptor-negative tumor.

The Adjuvant! Online model is an online risk calculator that uses routinely available breast cancer characteristics such as tumor size and lymph node status to provide women with 3 critical numbers: the absolute reduction in the risk of recurrence associated with chemotherapy, hormonal therapy, and combination therapy. A screenshot of the Adjuvant! Online calculator (which is available at https://www.adjuvantonline.com) is shown in Figure 5.
These numbers can be used in patient counseling.\textsuperscript{41} For example, a woman who expects to gain only a small benefit with adjuvant therapy may decide to forgo treatment; conversely, a woman who is averse to chemotherapy may change her mind when presented with information that she would experience a large decrease in risk. There is direct evidence that women can understand the Adjuvant! Online interface and accordingly make appropriate treatment decisions. Women with breast cancer were randomized to a consultation with or without data from Adjuvant! Online. Those with little to gain from adjuvant treatment tended to use less therapy in the Adjuvant! Online group compared with controls; conversely, women with a large increase in predicted survival from Adjuvant! Online tended to use more therapy.\textsuperscript{42} There is also evidence that the predictions from Adjuvant! Online correspond with actual patient outcomes.\textsuperscript{43} Although initially focusing on breast cancer, the Adjuvant! Online Web site now includes models for colon and lung cancer.

The ACCENT Model

Adjuvant therapy after colectomy has many similarities to that for breast cancer: a common cancer, clear evidence that adjuvant therapy is of value on average, and wide differences in benefit depending on patient risk. Gill et al have published a simple table whereby patients and clinicians can look up the absolute risk reduction associated with chemotherapy for different colon cancer risk groups. For example, for patients with lymph node-negative, low-grade T3 disease, chemotherapy is associated with a predicted 6\% absolute increase in disease-free survival at 5 years; in contrast, chemotherapy for a patient with lymph node-positive, high-grade T4 disease improves predicted survival by 18\%.\textsuperscript{44} An updated version of the tool is available online at http://www.mayoclinic.com/calcs/colon/index-ccacalc.cfm.

Current Controversies in Prediction Modeling

Too Many Models, Not Enough Independent Validation

Recent years have seen a glut of prediction models. A search on MEDLINE for “cancer” with either “prediction model,” “prognostic model,” or “nomogram” retrieves close to 10,000 hits. One review identified over 100 different prediction models for prostate cancer alone.\textsuperscript{45} Numerous models are available online including, amusingly, www.nomograms.org, www.nomogram.org, and www.cancernomograms.com. The ubiquity of prediction models has led to a “meta-literature” to help clinicians distinguish useful from less useful models.\textsuperscript{46,47} One plausible explanation for the high incidence of prediction modeling in the literature is that an investigator does not need to conduct an experiment, or even develop an interesting scientific question. The data set is fed into statistical software and out pops a paper.

With so many models being developed on a continuous basis, it should come as no surprise that only a minority have been subject to independent validation. The vast majority of models are created and validated on the same data set, with little further analysis. Lack of external validation is due at least in part to the failure of researchers to make available to others the mathematical formulae underlying their models. A counterexample to the general trend of “once-off” articles on models is instructive. A urologic cancer research team at the University of Montreal has published a number of studies in which competing models are tested head-to-head on a single external validation set.\textsuperscript{16,48-50} For example, the group tested the Kattan nomogram, the Cancer of the Prostate Risk Assessment (CAPRA) score, and the D’Amico risk classification for prediction of recurrence after radical prostatectomy.\textsuperscript{50} That such a natural and obvious approach is sufficiently rare to be noteworthy indicates a contemporary overemphasis on the creation of new models in place of the evaluation of existing models.

Poor Integration Into Clinical Practice

Information technology has been critical to the rapid growth of prediction modeling, whether through facilitating the collation and distribution of large data sets, rapid implementation of complex regression formulae for model creation, or Web-based tools to run models for individual patients. Yet there is surprisingly little integration of models into electronic health record systems and scant attention to how models can be used effectively in the clinical workflow. For example, it is not unusual for physicians to access a patient’s data electronically, then go to a Web site and retype the patient’s data into the
Web site interface to obtain a prediction. Similarly, it would seem relatively straightforward to integrate prediction modeling into the laboratory report. For example, instead of simply reporting a PSA level of 5 ng/mL, the report could also give: “Predicted probability of prostate cancer: DRE negative: 18%; DRE positive: 41%.” Yet this has rarely been attempted.

Do Models Do More Good Than Harm?
The literature is almost completely devoid of studies investigating the clinical implications of models. As a simple illustration, the Oncotype DX breast cancer assay (Genomic Health, Redwood City, Calif) gives a predicted probability of breast cancer recurrence on the basis of RNA expression in tumor samples. This probability is used in decisions about adjuvant therapy. There are unambiguous data that the Oncotype DX score does indeed correlate with recurrence risk.51 However, take a group of 100 women given a low risk of recurrence by Oncotype DX, who decide to forgo adjuvant therapy as a result. At least some of these women will develop recurrences, and a proportion of these would have been prevented by adjuvant therapy. As such, some women will die of breast cancer as a direct result of using Oncotype DX. Of course, these deaths would be offset by a large reduction in the use of adjuvant therapy, which has side effects and risks of its own. To know whether the increase in deaths is worth the decrease in treatment, each would have to be carefully quantified. To the best of my knowledge, this has never been attempted.51

One example where a prediction model was evaluated in terms of clinical outcome concerned a study of a model predicting recurrence after radical cystectomy.52 The authors calculated the number of patients treated and recurrences avoided if referral to adjuvant therapy was based on the statistical model compared with conventional decisions based on TNM stage criteria. Across various assumptions for the benefits of adjuvant therapy, in terms of relative risk reduction, and harms, in terms of the maximum number of patients an oncologist would treat to prevent one recurrence, it could be clearly demonstrated that use of the model would lead to improved clinical outcomes.

Yet even this study was based on a theoretical model of physician behavior. Impact studies are practical assessments of the real-world effects of prediction modeling on medical decision-making and patient outcomes.53,54 Such studies are all but unknown in oncology. The Adjuvant! Online trial described above,32 where use of adjuvant therapy was compared between breast cancer patients receiving the results of a prediction model and those randomized to usual care control, is a rare counterexample.

Conclusions
There can be no reasonable doubt that prediction models are here to stay. Given the burgeoning complexity of the diagnostic and prognostic information available to oncologists, a trend that will only increase as our understanding of the human and cancer genomes improves, there is simply no realistic alternative to incorporating multiple variables in a single prediction model. It certainly seems implausible that the TNM staging system could be expanded indefinitely to incorporate new markers. As such, the question should not be whether but how prediction models should be used to aid decision-making in cancer care. Key issues will be integration of models into the electronic health record and more careful evaluation of models, particularly with respect to their effects on clinical outcomes.

References
1. Vickers AJ, Cronin AM, Roobol MJ, et al. The relationship between prostate-specific antigen and prostate cancer risk: the Prostate Biopsy Collaborative Group. Clin Cancer Res. 2010;16:4374-4381.
2. Kattan MW, Karpeh MS, Mazumdar M, Brennan MF. Postoperative nomogram for disease-specific survival after an R0 resection for gastric carcinoma. J Clin Oncol. 2003;21:3647-3650.
3. Ferrone CR, Kattan MW, Tomlinson JS, Thayer SP, Brennan MF, Warshaw AL. Validation of a postresection pancreatic adenocarcinoma nomogram for disease-specific survival. J Clin Oncol. 2005;23:7529-7535.
4. International Bladder Cancer Nomogram Consortium, Bochner BH, Kattan MW, Vora KC. Postoperative nomogram predicting risk of recurrence after radical cystectomy for bladder cancer. J Clin Oncol. 2006;24:3967-3972.
5. Weiser MR, Landmann RG, Kattan MW, et al. Individualized prediction of colon cancer recurrence using a nomogram. J Clin Oncol. 2008;26:380-385.
6. Wong SL, Kattan MW, McMasters KM, Coit DG. A nomogram that predicts the presence of sentinel node metastasis in melanoma with better discrimination than the American Joint Committee on Cancer staging system. Ann Surg Oncol. 2005;12:282-288.
7. Glare P, Virik K, Jones M, et al. A systematic review of physicians’ survival predictions in terminally ill cancer patients. BMJ. 2003;327:195-198.
CA: A Cancer Journal for Clinicians

10. Etzioni RD, Ankerst DP, Weiss NS, Inoue LY, et al. Comparative effectiveness of routine PSA screening in the United States: a decision analytic study. *J Natl Cancer Inst* 2001;93:90-101.

11. Atkinson DG, Royston P, Weiss NS, Inoue LY, Thompson IM. Is prostate-specific antigen velocity in early detection of prostate cancer? A critical appraisal of the evidence. *J Clin Oncol* 2007;25:1510-1515.

12. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the effect of individual predictors of cardiovascular risk: the role of reclassification measures. *Ann Intern Med* 2009;150:795-802.

13. Thompson IM, Ankerst DP, Chi C, et al. Operational definition of prostate specific antigen in men with an initial PSA level of 3.0 ng/ml or lower. *JAMA* 2005;294:66-70.

14. Vickers AJ, Gupta A, Savage CJ, et al. Assessment of prostate cancer risk: results from the Prostate Cancer Prevention Trial. *J Natl Cancer Inst* 2006;98:529-534.

15. Eyre SJ, Ankerst DP, Wei JT, et al. Validation in a multiple urology practice cohort of the Prostate Cancer Prevention Trial risk calculator for predicting prostate cancer detection. *J Urol* 2009;182:2653-2658.

16. Plaza P, Wu TV, Siderovski DP, et al. Population-based validation of the Prostate Cancer Prevention Trial risk calculator for predicting prostate cancer in a screened population. *Urology* 2006;68:1152-1155.

17. Kaplan DJ, Boorjian SA, Ruth K, et al. Validation of the Prostate Cancer Prevention Trial Risk calculator in a high-risk screening population. *BJU Int* 2010;105:334-337.

18. Nguyen CT, Yu C, Mousa A, Kattan MW, Jones JS. Performance of prostate cancer prevention trial risk calculator in a contemporaneous cohort screened for prostate cancer and diagnosed by extended prostate biopsy. *J Urol* 2010;183:529-533.

19. Gail MH, Brinton LA, Byar DP, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst* 1989;81:1879-1886.

20. Fisher B, Costantino JP, Wickerham DL, et al. Tamoxifen for the prevention of breast cancer: report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study. *J Natl Cancer Inst* 1998;90:1371-1388.

21. Ross RW, Kantoff PW. Predicting outcomes of prostate cancer metagrame: a solution to the dilemma of which prediction tool to use in patient counseling. *Cancer* 2009;115(13 suppl):3039-3045.

22. Zorn KC, Capitanio U, Jeldres C, et al. Multi-institutional external validation of seminal vesicle invasion nomograms: head-to-head comparison of Gallina nomogram versus 2007 Partin tables. *Int J Radiat Oncol Biol Phys* 2009;73:1461-1467.

23. Benchikh A, Savage C, Cronin A, et al. A panel of kallikrein markers can predict outcome of prostate biopsy following clinical work-up: an independent validation study from the European Randomized Study of Prostate Cancer screening. France. *BMC Cancer* 2010;10:635.

24. Vickers A, Gupta A, Savage CJ, et al. A panel of kallikrein marker predicts prostate cancer in a large, population-based cohort followed for 15 years without screening. *Clin Cancer Res* 2011;20:255-261.

25. Bolla M, van Poppel H, Eisenberger MA, et al. Ten-year results of a randomized trial comparing 3 versus 5 years of adjuvant hormonal therapy in patients with locally advanced prostate cancer. *J Clin Oncol* 2010;28:2493-2498.