Translating Probability Models into Clinical Decisions

Frank E Harrell Jr
Professor of Biostatistics and Statistics
Division of Biostatistics and Epidemiology
Department of Health Evaluation Sciences
University of Virginia School of Medicine
Charlottesville VA 22908 USA
fharrell@virginia.edu

Department of Biomedical Informatics
Ljubljana University,
Ljubljana, Slovenia
Organizer: Janez Stare

June 1997

Abstract

Models for predicting the probability of a positive diagnosis or other event are becoming increasingly popular. For example, the binary logistic regression model is frequently used for predicting the probability that a given patient has a certain disease. Logistic regression model coefficients are estimated using the method of maximum likelihood. Coefficients which maximize the likelihood of the observed data are optimum in some ways. The method of maximum likelihood optimizes a "logarithmic probability scoring rule" or $\sum_{i=1}^{n}[Y_i \log P_i + (1 - Y_i)\log(1 - P_i)]$, where Y is 1 if the event occurred, 0 if not, and P is the probability that $Y = 1$ as dictated from the assumed model. Thus the process of estimating model coefficients optimizes a sensitive measure of predictive accuracy that uses probabilities as continuous measures.

A separate approach to prediction is *classification*, which can be done after the fact by dichotomizing predicted probabilities, or the predictive instrument can be derived to optimize some (cost) function of "false positive" and "false negative" classifications. If one likes the philosophy that maximally accurate probability estimation should be the central goal, then maximum likelihood and its variants should be favored over classification methods. Then when an empirical model is developed, its predicted probabilities need to be validated for discrimination and calibration accuracy.

Even when a probability model has been developed by the analyst, it is all too common for her to try to transform the problem into a classification task. Then it is common to choose as a measure of predictive ability for binary logistic models the fraction of correctly classified responses. Here one chooses a cutoff on the predicted probability of a positive response and then predicts that a response will be positive if the predicted probability exceeds this cutoff. There are a number of reasons why this measure should be avoided:

1. It's highly dependent on the cutpoint chosen for a "positive" prediction.

2. You can add a highly significant variable to the model and have the percent classified correctly actually decrease. Classification error is a very insensitive and statistically inefficient measure since if the threshold for "positive" is say 0.75, a prediction of 0.99 rates the same as one of 0.751.

3. It gets away from the purpose of fitting a logistic model. A logistic model is a model for the probability of an event, not a model for the occurrence of the event. For example, suppose that the event we are predicting is the probability of being struck by lightning. Without having any data, we would predict that you won't get struck by lightning. However, you might develop an interesting model that discovers real risk factors that yield probabilities of being struck that range from 0.000000001 to 0.001.

4. If you make a classification rule from a probability model, you are being presumptuous. Classifications need to be deferred until both the physician and the patient are able to put the diagnosis in context. Different patients have different thresholds for treatment.

5. The classification accuracy, unlike proper scoring rules, is not maximized when the predicted probabilities are the population probabilities.

A predicted probability is the best way to summarize a probability regression model. Instead of imposing an arbitrary cutoff in declaring a prediction positive ($Y = 1$), the use of a probability can allow the threshold to vary (as it always does in practice) by such factors as age and availability of health care resources. Suppose, for example, that a model is developed to assist physicians in diagnosing coronary artery disease. Physicians sometimes say that they want a binary decision model, but when you study their behavior you'll find that if you give them a probability, they will apply different thresholds for treating different patients or for ordering other diagnostic tests. Even though the age of the patient may be a strong predictor of the probability of disease, the physician will often use a lower threshold of disease likelihood for treating a young patient. It is important to note age is one of the strongest predictors of coronary disease, and its affect has been taken into account in the model for the probability of disease. The usage of different thresholds for treatment for patients of different ages is above and beyond how age effects the probability of disease.

For another example, consider two infants who arrive at a clinic on the same day. If one had a predicted probability of serious infection of 0.6 while the other was 0.95, and if hospital beds were scarce, it would make sense to hospitalize the one with a prediction of 0.95. Even if the clinic staff were uncomfortable with the use of probabilities, it would be advisable to rank the infants by risk each day and to select infants for admission in decreasing order of risk down to some lower threshold, subject to available beds.

When the medical staff can deal neither with probabilities nor with risk rankings, an (arbitrary) dichotomous decision rule may be needed for simplicity. The problem here is how to most easily estimate the probability P based on patient characteristics X, so that a reasonable dichotomization can be derived. Nomograms (Pryor et al. 1983, Am J Med 75:771-780; Spanos et al. 1989, JAMA 262:2700-2707) can be easily used for even complex nonlinear models. Once P is estimated, the dichotomous decision could be $Y = 1$ if $P \geq c$, $Y = 0$ if $P < c$. The cutoff c can be chosen according to which type of classification error is more serious. This is a more direct approach than setting the sensitivity and specificity. If $c = 0.8$, the probability that an infant is normal if she is classified as "serious infection" is at most 0.2. On the other hand, an infant who is classified as normal because $P < 0.80$ actually can have a probability of 0.79 of having a serious infection if her true P is 0.79.

If some situations, the procedure can be simplified further if only one cutoff c is considered. Letting $d = \mathrm{logit}(c) = \log \frac{c}{1-c}$, we classify a subject as $Y = 1$ if the predicted logit from a binary logistic model is $\geq d$. For simplicity, let us assume for now that $c = 0.5$ so that $d = 0$. A simple model would be $P = \frac{1}{1+\exp-(a+bf(X))}$ for a single continuous variable X with appropriate linearizing transformation $f(X)$. If $f(X) = X$, we only need to solve for $a + bX > 0$ to find the threshold for X beyond which Y is predicted as 1. This threshold is $-\frac{a}{b}$ from the fitted coefficients. Suppose that one wishes to diagnose pneumonia on the basis of the respiration rate r and the presence or absence of a cough. The predicted logit might be $a + b \times \mathrm{cough} + h \times r$, and for infants without cough we declare $Y = 1$ if $r > -\frac{a}{h}$; for those with cough we use $r > -\frac{a+b}{h}$. For example, we may declare a positive diagnosis if $r > 90$ without cough or $r > 80$ with cough. If there is an interaction between cough and r, a similar rule will result.

If the model contains two continuous factors X_1 and X_2, a single $x - y$ plot with a line beyond which $Y = 1$ can be drawn. X_1–specific tables of X_2 cutoffs can also be made.

What if the model contains a series of indicators $X1, \ldots, X_5$ with a predicted log odds of disease of $-3 + .9X_1 + 1.1X_2 + .95X_3 + 1.05X_4 + 1X_5$? This model can be simplified to $-3 + m$ where m is the number of the five signs present. The prediction would be $Y = 1$ if $m \geq 3$. Now consider the model $\mathrm{logit} = -3 + m + .03(r - 70)$ where r is respiration rate. This model predicts $Y = 1$ if $m \geq 4$ or if $m = 3$ and $r \geq 70$ or if $m = 2$ and $r \geq 103$ or if $m = 1$ and $r \geq 137$.

Outline

- Strategy for estimating the probability of an event or positive diagnosis on the basis of patient characteristics
- What does a best–fitting model fit?
- Components of predictive accuracy
- Computation of $\hat{P}$
- Problems in dichotomizing $\hat{P}$: Why it should not be done by the analyst or in a publication
- What if you really need to dichotomize?

Strategy for Estimating $\Pr[Y = 1|X]$

- Select a model (e.g., binary or ordinal logistic model or survival model)
- Clinical guidance
- How does each potential predictor relate to Y?
  - Stratify outcomes by intervals of X
  - Better: nonparametric regression (generalization of moving average)
  - Piecewise polynomials (cubic spline functions)
- Check for effect modification (interaction)
- Guard against overfitting
  - Limit list of predictors
  - Reduce groups of related variables into summary scores
  - Shrinkage

Criteria for Best–Fitting Model

- For binary logistic model we use method of maximum likelihood
- Solve for coefficients that maximizes likelihood of observed data
- Maximum likelihood estimates (MLEs) are optimum in some ways
- Optimizes a "logarithmic probability scoring rule" or $\sum_{i=1}^{n}[Y_i \log P_i + (1 - Y_i)\log(1 - P_i)]$,
  - Y=1 (event), 0 (no event)
  - P_i = predicted probability that $Y = 1$ as dictated from the model
- Estimating model coefficients optimizes a sensitive measure of predictive accuracy
- P is used as a continuous measure

- Some researchers instead optimize a “cost function” after predictions are dichotomized
- E.g., score based on weights for “false negatives” and “false positives”
- This is inefficient and is not optimized when predicted probabilities are population values
- MLE is preferred

Components of Predictive Accuracy

- Discrimination: ability of predictions to separate good from bad outcomes
  - Area under "receiver operating characteristic" curve (c)
  - Somers' D_{xy} rank correlation between predicted and observed outcome ($D_{xy} = 2(c - \frac{1}{2})$)
- Calibration: how close is predicted vs. observed to a 45° line

Computation of $\hat{P}$

- Calculator
- If only one continuous and one categorical predictor (e.g., age and sex) can use a graph
- Nomogram
- Table

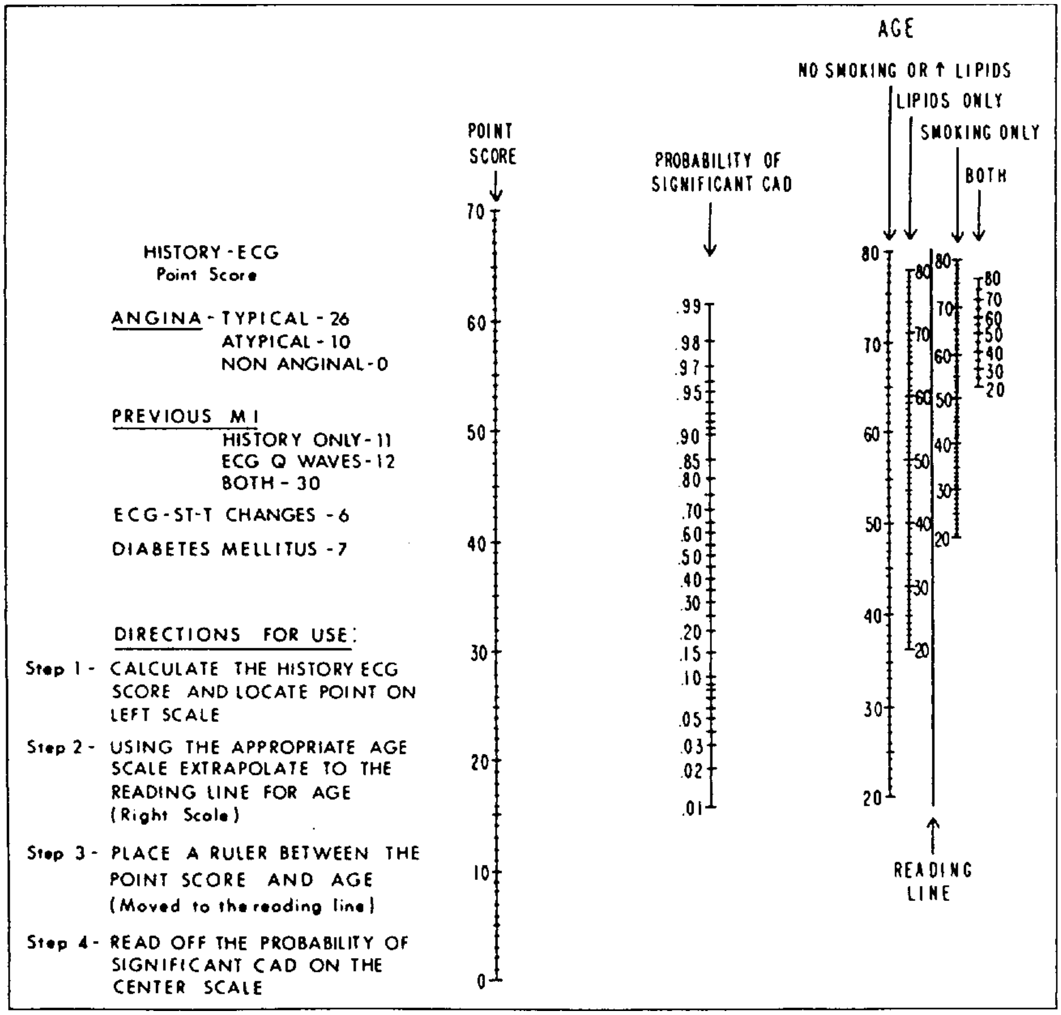

Figure 1: A nomogram for estimating the likelihood of significant coronary artery disease (CAD) in women. Depiction of a fitted binary logistic regression model. Categorical predictors have their points added manually. ECG = electrocardiographic; MI = myocardial infarction (Pryor et al. 1983). Presence of important age × risk factor interactions is handled by constructing separate age scales for each level of the interacting factor. Here, interaction means a change in the slope (regression coefficient) for age depending on which risk factors are present. A change in slope implies stretching or shrinking the scale on the age axis. A better way to interpret this is that the effect of the risk factors declines with age.

Problems Caused by Dichotomizing $\hat{P}$

- Choose a cutoff c on $\hat{P}$
- Predict $Y = 1$ if $\hat{P} > c$, $Y = 0$ otherwise
- Use proportion of "correct" predictions as an accuracy measure (one minus *classification error*)
- Results highly dependent on c
- Can sometimes add highly significant variable to the model and have the proportion classified correctly actually decrease
- Classification error is a very insensitive and statistically inefficient measure.
  Example: If $c = 0.75$, $\hat{P} = 0.99$ rates the same as $\hat{P} = 0.751$.
- Approach is not consistent with the goal of logistic modeling

- Meaningful to predict the probability of a rare outcome, whereas classifying them as present/absent may not be relevant
- Classification is presumptuous— should defer until physician and patient can put the diagnosis in context.
  Different physicians and patients have different thresholds for treatment.
- Example: Diagnosis of coronary artery disease
  Pr(disease) depends on age, sex, angina, risk factors, etc.
  Most physicians send younger patients to cardiac catheterization more readily, even after Pr(disease) has *fully taken age into account*.
- Example: Two infants arrive at a clinic, one with Pr(serious infection)=0.6, other with 0.95.
  Hospital beds scarce → hospitalize one with 0.95.
  Could rank all infants arriving for treatment by descending Pr(disease)
  Note that threshold for treatment varies daily.

What if Dichotomization is Mandated

- If model has any continuous predictors or if it has many categorical ones, will need to compute $\hat{P}$ to be able to classify $Y = 1$ if $\hat{P} > c$
- Example 1: $\hat{P} = [1 + \exp -(a + bf(x))]^{-1}$
  $\hat{P} > c \rightarrow \text{logit} = \log(\frac{\hat{P}}{1-\hat{P}}) > \log(\frac{c}{1-c}) = d$
  $\rightarrow a + bf(x) > d$ or $f(x) > \frac{d-a}{b}$
- Example 2: cough=0 or 1, r=respiration rate
  $\text{logit} = a + b \times \text{cough} + h \times r$
  No cough $\rightarrow r > \frac{d-a}{h}$
  Cough $\rightarrow r > \frac{d-a-b}{h}$
- Even if classify, useful to compute $\hat{P}$
  - $\hat{P} > c \rightarrow \hat{Y} = 1$
  - $\Pr[Y = 0|\hat{Y} = 1, \hat{P}] = 1 - \hat{P}$
    false positive rate
  - $\Pr[Y = 1|\hat{Y} = 0, \hat{P}] = \hat{P}$
    false negative rate

- Example: $c = 0.8, \hat{P} = .79$, false negative rate=0.79