Calibration in Computer Models for Medical Diagnosis and Prognostication

Lucila Ohno-Machado, MD, PhD

Division of Biomedical Informatics, Department of Medicine, UCSD

Frederic Resnic, MD, MSc

Division of Cardiovascular Medicine, Brigham and Women's Hospital and Harvard Medical School

Michael Matheny, MD, MSc, MPH

GRECC and Center for Health Services Research, Tennessee Valley Health System, Veterans Administration

Division of General Internal Medicine and Public Health

Department of Biomedical Informatics, Vanderbilt University

Mail Contact:

Lucila Ohno-Machado, MD, PhD

Professor and Chief, Division of Biomedical Informatics

9500 Gilman Dr #0671

La Jolla, CA 92093-0671

858-246-0224

machado@ucsd.edu

Predictive models are being developed in virtually every medical specialty for the purposes of diagnosis or prognosis. They provide an individualized estimate of health care related events, such as prognosis in cardiovascular disease, given specific information about an individual (e.g., genotype, family history, past medical history, clinical findings). These models, developed using statistical and machine learning techniques applied to large clinical data sets, are being used by health care professionals as well as patients. Verification that the estimated or predicted event probabilities truly reflect the underlying probability for a particular individual (i.e., evaluating model calibration) is a critical but often overlooked step in model evaluation.

Measuring Calibration

A primitive notion of calibration is often designated as calibration-in-the-large or bias, which is the difference between the average estimate (prediction) and the proportion of observed events. This is the best estimate if one considers a single cluster of patients. Assigning the prior probability of the event as the risk score for every patient would result in a perfectly calibrated-in-the-large model, but it would add no individualized information, hence this index is of limited practical utility for assessment of predictive models, although it can be useful for recalibration.

A fundamental problem in evaluating calibration in medical problems is the lack of a gold-standard to compare the individual risk estimate to. A gold-standard would consist of sufficient numbers of exact replicas of the individual accurately diagnosed or followed without censoring so that the proportion of observed events would be equal to the “true estimate” for the individual. For large enough groups of individuals with *similar* profiles, there could be meaningful approximations of the true probability. However, the way in which we define the similarity of patient profiles is important. Currently, calibration is measured by comparing outcomes in sets

of patients with similar estimated risks, but this *output-space* cluster is dependent on the predictive model and is counter-intuitive.

Output-space similarity

One of the most popular indices to assess calibration of predictive models was developed in the context of logistic regression by Lemeshow and Hosmer (1982). The idea behind the test is simple: if the cases are sorted according to the estimated level of risk and the mean estimate for each decile of risk is very close to the proportion of positive cases within the decile, then one cannot reject the hypothesis that the model is correct. (Hosmer, Taber et al. 1991; Hosmer, Hosmer et al. 1997; Hosmer and Hjort 2002). The sum of [the squared differences between the sum of estimates and number of events in each decile divided by the sum of estimates in that decile] for each outcome is reported to follow a χ^2 distribution with 8 degrees of freedom. If $p < 0.05$, we reject the hypothesis that the model fits the data. The H-L C statistic based on deciles of risk is defined as:

$$C = \sum_{D=0}^{1} \sum_{l=1}^{10} \left[\frac{\left(\pi_{Dl} - o_{Dl} \right)^2}{\pi_{Dl}} \right],$$

where π_{Dl} and o_{Dl} are the sum of estimates in a decile and observed frequencies in the same decile, for cells indexed by group (decile) *l* and outcome *D*. Hosmer and Lemeshow showed via simulations that C is approximately distributed as $\chi 2$ with l-2 = 8 degrees of freedom when the fitted model is the correct one and the estimated expected cell frequencies are sufficiently large. Note that the H-L statistic is model-dependent, since the statistic compares the average estimate in each decile of estimated risk with the proportion of events in that decile. To visualize

the calibration of a predictive model, it is common to plot the average estimate for groups representing either (a) percentiles of estimated risk against the proportion of events in that group, as described above, or (b) pre-defined ranges of the estimates. The latter is common in clinical predictive models.

Input-space similarity

We describe a simulation in which we established in advance 4 tight clusters of "patients" according to 2 variables, x1 and x2. The purpose of this simulation was to illustrate the HL goodness-of-fit statistic and check whether differences in calibration can be determined using this statistic. Bi-normal distributions were generated with identical standard deviations (0.1) and centered at (0,0), (0,1), (1,0), and (1,1), for clusters 1 to 4, respectively, each with 100 patients. The binary outcome for each patient in a cluster was generated from a Bernoulli distribution with probabilities 0.01, 0.4, 0.6, and 0.99 for clusters 1 to 4, respectively. Figure 1 shows the spatial distribution of the clusters. For verification, the four clusters were automatically re-discovered using the Expectation-Maximization algorithm.

The resulting logistic regression model is highly significant. For comparison, we built a neural network with hidden units so that it was capable of finding a non-linear function relating the predictors and outcomes, with results illustrated in Figure 1b. An ideal model would assign the true underlying probability for each case (i.e., 0.01, 0.4, 0.6, and 0.99 depending on which cluster the case belonged to). A neural network with enough parameters was able to get closer to that goal than a semi-linear model such as logistic regression. In this simulation, we intentionally allowed some degree of overfitting of the neural network model to check whether differences could be noted in the H-L statistics.

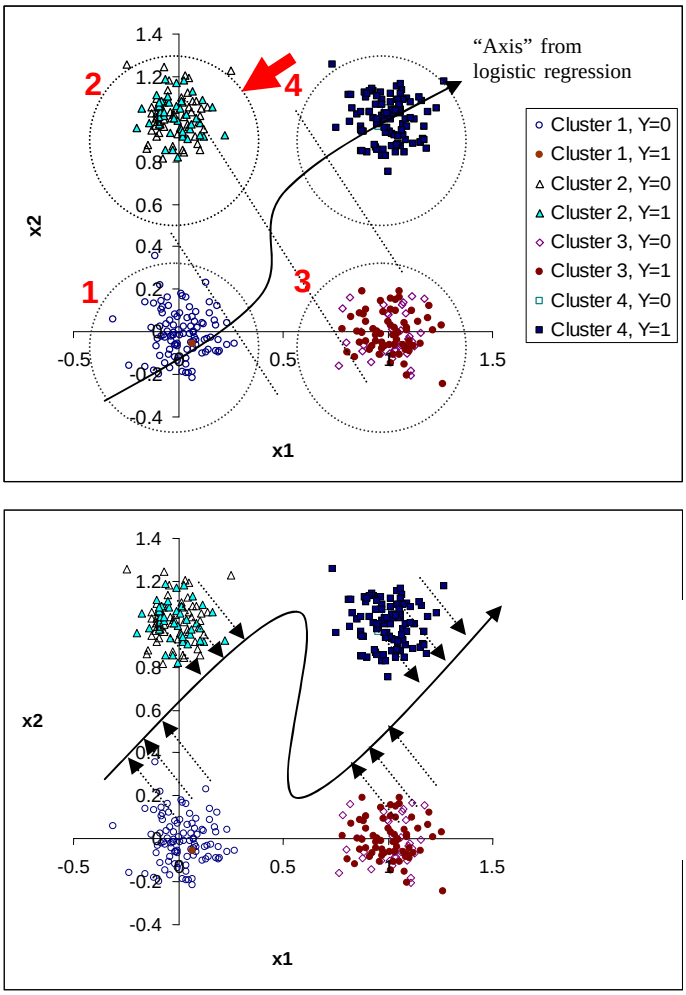

Figure 1. Simulation with four pre-defined non-overlapping bi-normal clusters of individuals with known underlying probability of an event (0.01, 0.40, 0.60, and 0.99 for clusters 1 to 4, respectively). Top Panel: Two dimensional data are projected into one dimension by the logistic regression model. Dotted diagonal lines divide quartiles of risk as determined by the logistic regression model. A patient from cluster #2 indicated by the arrow has an estimate closest to the average estimate in cluster #3 than to that of cluster #2. Confidence in its estimate should be lower than that for a patient in the middle of one of the clusters. The input-space clusters, as opposed to the quartiles of risk, are easy to explain: Cluster #1 has patients with "low x1 and low x2", Cluster #2 has patients with "low x1 and high x2", and so on.
Bottom Panel: Projection of the points into an "axis" for neural network estimates. The neural network model fits better the true probabilities for clusters 2 and 3 than does the logistic regression model.

In Table 1, we display descriptive statistics for the estimates obtained by the two types of models.

Cluster	Proportion of Events	LR Mean	NN Mean	LR Std Dev	NN Std Dev	LR Minimum	NN Minimum	LR Maximum	NN Maximum
1	0.01	0.0338	0.0219	0.0172	0.0069	0.0066	0.015	0.0949	0.059
2	0.42	0.4129	0.4819	0.1080	0.0207	0.1955	0.431	0.8013	0.584
3	0.64	0.6291	0.6852	0.1149	0.0146	0.2873	0.647	0.8507	0.732
4	0.98	0.9740	0.9908	0.0127	0.0011	0.9301	0.985	0.9954	0.992

Table 1. Descriptive statistics of Logistic Regression (LR) and Neural Network (NN) estimates according to input-space clusters. Note that the neural network model's estimates do not overlap between clusters (i.e., the minimum estimate for cluster #3 is greater than the maximum estimate for cluster #2).

The H-L C statistic for the logistic regression model was 6.43 (p=0.59), hence we would not reject the hypothesis that the model is calibrated. Although the neural network model had a less favorable H-L C of 11.773 (p = 0.16), its overall errors were smaller.

In this example, the neural network was better able to approximate the true underlying probability of the event in clusters #2 and #3, as could be seen by the ranges of estimates in these clusters, as well as their maximum residuals. However, comparison of HL-C and inspection of the calibration plot in Figure 2 would not indicate that neural networks would be a better model in this case.

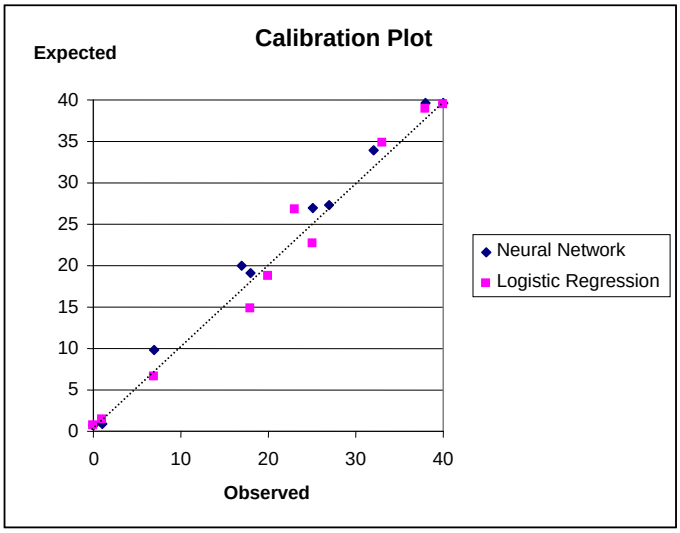

Figure 2. Calibration plot for logistic regression and neural network models based on deciles of risk. There is no apparent superiority of one model versus another.

Implications for Medical Decisions

In clinical practice, there are important implications for incorrect estimates. For example, the widely employed clinical practice guideline from the Adult Treatment Panel III (NCEP, 2002) utilizes cardiovascular risk estimates similar to the ones available in online calculators to recommend particular treatment regimens. For non-calibrated estimates, this may result in inappropriate use of medication to manage cholesterol levels. Computer-based post-marketing tools for surveillance of new medications and medical devices utilize models that adjust risk for the population being treated (Matheny et al, 2006). These models depend on the accuracy of these estimates to trigger appropriate alerts for unsafe technologies and drugs. For non-calibrated estimates, this risk adjustment may result either in a large number of false positives or false negatives. Both situations incur large costs to the healthcare system. It is therefore critical to assess the calibration of estimates before employing these models in clinical settings.

We and others have shown in different domains that the calibration of medical diagnostic and prognostic models can vary significantly according to the population in which they are applied (Hukkelhoven et al, 2006; Matheny et al, 2005; Ohno-Machado et al, 2006), as opposed to the model's discrimination, often measured by areas under the Receiver Operating Characteristic curve. While there are some efforts to recalibrate models for different populations and study reclassification rates, web-based calculators that estimate individualized risk do not yet take this issue into account, and may present incorrect estimates to a given individual. We have proposed some methods to take into account input-space clusters in predictive models (Osl et al 2008; Robles et al 2008), but much remains to be done in order to inform healthcare workers and the public about the potential shortcomings of this facet of personalized medicine. As new molecular-based biomarkers for a variety of health conditions are currently being developed at an accelerated pace and used in multidimensional models to diagnose or prognosticate these conditions, it becomes even more important to develop accurate methods to assess the quality of estimates derived from predictive models.

Acknowledgements

The authors acknowledge support from the National Library of Medicine, NIH, FDA, and VA grants R01LM009520 (LO), R01 LM008142 (FR), HHSF 223200830058C (FR), VA HSR&D CDA2-2008-020 (MM).

References

Hosmer, D. W. and N. L. Hjort. "Goodness-of-fit processes for logistic regression: simulation results." Stat Med. 2002;21(18): 2723-38.

Hosmer, D. W., T. Hosmer, et al. "A comparison of goodness-of-fit tests for the logistic regression model." Stat Med, 1997;16(9): 965-80.

Hosmer, D. W., S. Taber, et al. "The importance of assessing the fit of logistic regression models: a case study." Am J Public Health. 1991;81(12): 1630-5.

Hukkelhoven, C. W., A. J. Rampen, et al. "Some prognostic models for traumatic brain injury were not valid." J Clin Epidemiol. 2006;59(2): 132-43.

Lemeshow, S. and D. W. Hosmer, Jr. "A review of goodness of fit statistics for use in the development of logistic regression models." Am J Epidemiol. 1982;115(1): 92-106.

Matheny ME, Ohno-Machado L, Resnic FS. Discrimination and calibration of mortality risk prediction models in interventional cardiology. J Biomed Inform. 2005;38(5):367-75.

Matheny, M. E., L. Ohno-Machado, et al. "Monitoring device safety in interventional cardiology." J Am Med Inform Assoc. 2006;13(2): 180-7.

NCEP. Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. *Circulation.* Dec 17 2002;106(25):3143-3421.

Ohno-Machado L, Resnic FS, Matheny ME. Prognosis in Critical Care. Annual Review of Biomedical Engineering. 2006;8:567-599.

Osl M, Ohno-Machado L, Dreiseitl S. Improving calibration of logistic regression models by local estimates. AMIA Annu Symp Proc. 2008: 535-9.

Robles V, Bielza C, Larranaga P, Gonzales S, Ohno-Machado L. Optimizing logistic regression coefficients for discrimination and calibration using estimation of distribution algorithms. TOP, 2008;16(2):345-66.