Assessment
2020, Vol. 27(4) 840–854
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1073191117752055
journals.sagepub.com/home/asm
SAGE

The Importance of Calibration in Clinical Psychology

Oliver Lindhiem[1], Isaac T. Petersen[2], Lucas K. Mentch[1], and Eric A. Youngstrom[3]

Abstract

Accuracy has several elements, not all of which have received equal attention in the field of clinical psychology. Calibration, the degree to which a probabilistic estimate of an event reflects the true underlying probability of the event, has largely been neglected in the field of clinical psychology in favor of other components of accuracy such as discrimination (e.g., sensitivity, specificity, area under the receiver operating characteristic curve). Although it is frequently overlooked, calibration is a critical component of accuracy with particular relevance for prognostic models and risk-assessment tools. With advances in personalized medicine and the increasing use of probabilistic (0% to 100%) estimates and predictions in mental health research, the need for careful attention to calibration has become increasingly important.

Keywords

calibration, prediction, forecasting, accuracy, bias, confidence, risk-assessment

Probabilistic (0% to 100%) estimates and predictions are being used with increasing frequency in mental health research, with applications to confidence in clinical diagnoses, the probability of future onset of disease, the probability of remission, the probability of relapse, and risk assessment in various domains. Broad categories such as "high risk" or "low risk" are being replaced by the use of the full range of continuous risk assessment to guide the decision-making process (e.g., Baird & Wagner, 2000; Spiegelhalter, 1986). With increasing precision, patients can be assigned personalized probabilistic predictions for a variety of outcomes. Ever-expanding data sets and increasingly sophisticated statistical software allow for individualized probabilistic predictions. However, such predictions can only be useful if they accurately reflect the true underlying probabilities. Accuracy can be defined in a number of different ways, but a critical component of accuracy for personalized probabilistic predictions is calibration. Unfortunately, calibration has been largely neglected by clinical psychologists in favor of other measures of accuracy such as sensitivity, specificity, and area under the receiver operating characteristic (ROC) curve (AUC). In this article, we define calibration, discuss its relationship to other components of accuracy, and highlight several examples in which calibration should be carefully considered by clinical psychologists. We argue that calibration is a critical aspect when evaluating the accuracy of personalized probabilistic predictions. When the goal is to make an optimal clinical decision, a clinician must be concerned with probabilistic prediction (e.g., Spiegelhalter, 1986). In many cases, it is important for probabilistic predictions to be accurate over the full range of 0 to 1 (0% to 100%) and not just accurate for a certain interval. Calibration is relevant both to individual predictions by "experts" and to statistical predictions from mathematical models. Many of the specific concepts and analyses discussed in this article are relevant to both, though some are only relevant to one or the other (e.g., overfitting in mathematical models and cognitive biases in experts).

A Definition of Calibration

Calibration is a specific component of accuracy that measures how well a probabilistic prediction of an event matches the true underlying probability of the event (e.g., Jiang, Osl, Kim, & Ohno-Machado, 2012). That is, given that a particular event occurs with probability p, a well-calibrated predictive modeling technique will produce a corresponding estimate $\hat{p}$ that is "close to" the true value p. In this sense, calibration is closely tied to the statistical notion of bias whereby for a particular parameter of interest θ, the

[1]University of Pittsburgh, Pittsburgh, PA, USA
[2]University of Iowa, Iowa City, IA, USA
[3]University of North Carolina at Chapel Hill, NC, USA

Corresponding Author:
Oliver Lindhiem, Department of Psychiatry, University of Pittsburgh, 3811 O'Hara Street, Pittsburgh, PA 15213, USA.
Email: lindhiemoj@upmc.edu

bias of a particular estimator $\hat{\theta}$ is defined as the expected difference, $E(\theta - \hat{\theta})$. An important distinction, however, is that we are concerned here with predictions (estimates) of probabilities that are generally considered to be the result or output of a more complex model or algorithm. That is, instead of directly observing a series of probabilities $p_1, \ldots, p_n$ from which we hope to produce an estimate $\hat{p}$ of the true value p, we most often observe only a binary response (1 or 0; event did or did not occur) along with a set of predictor variables from which we hope to construct a model that will produce reliable estimates of the true underlying probability that the event will occur at each level of those predictor variables.

Because an event either occurs or it does not, it is impossible to directly measure the true underlying probability of a one-time event. For this reason, the calibration of a statistical model or an individual forecaster can only be measured for a set of predictions. In practice, therefore, good calibration is defined, "in the sense that if a probability of 0.9 is given to each of a series of events, then about 90 percent of those events should actually occur" (Spiegelhalter, 1986; p. 425). This idea of "long-run frequency" is therefore aligned with the classic frequentist interpretation of probability. This means that if we were to examine the set of all days on which a weather forecaster predicts an 80% chance of rain, it should actually rain on approximately 80% of those days and *not* rain 20% of those days. So when, on occasion, a forecaster says there is an 80% chance of rain tomorrow and it does not rain, he or she is not necessarily "wrong." In fact, we would expect this outcome to occur 20% of the time. If, on the other hand, we examined 100 such days on which the forecaster predicted an 80% chance of rain and rain occurred on 95% of those days, we might suspect miscalibration.

Although calibration has received much attention in other fields (e.g., meteorology), it has been relatively ignored in clinical psychology in favor of discrimination (e.g., ROC, AUC, sensitivity, specificity). Discrimination, a key construct in signal detection theory, is a distinct dimension of accuracy from calibration, and both are important to consider when evaluating predictive accuracy. Calibration is closer to the construct of "response bias" in signal detection theory (Macmillan & Creelman, 1990). Response bias is the propensity to favor one response over another in discrimination tasks.[1] It can measure, for example, whether the percentage of Response A matches the actual percentage of Event A. Calibration, is similar to response bias, but for probabilistic models. Like calibration, response bias has received less attention by psychologists than other measures of discrimination such as sensitivity (Macmillan & Creelman, 1990) and is sometimes left out of primers on signal detection theory entirely (e.g., Treat & Viken, 2012). It is important for our predictions to both correctly distinguish between the two outcomes (discrimination) and to agree with the actual rates of outcomes (calibration).[2]

Calibration and Its Relationship to Discrimination

Calibration and discrimination are orthogonal constructs that assess complementary but distinct components of accuracy. It is important for research in clinical psychology to consider both calibration and discrimination separately. To be clinically and/or diagnostically useful in assessing risk, predictive models must be *both* well calibrated *and* have good discrimination (e.g., Steyerberg et al., 2010). Importantly, expertise has been sometimes associated with good discrimination but rarely with good calibration (Koehler, Brenner, & Griffin, 2002). That is, individuals deemed experts may possess an excellent ability to distinguish between and correctly predict final outcomes, but may struggle to assign accurate probabilities to such outcomes. Thus, calibration is an especially important factor to consider in clinical psychology where judgments and predictions are often made by expert clinicians and considering both calibration and discrimination is paramount for clinical psychologists to make the best decision regarding (a) whether to give a person a diagnosis and which diagnosis to give and (b) whether to apply an intervention and which intervention to use.

Probabilistic models can have good discrimination but poor calibration (Schmid & Griffith, 2005). For example, if a statistical model applied to screening data (e.g., Vanderbilt Assessment Scale; Wolraich, Hannah, Baumgaertel, & Feurer, 1998) based on *Diagnostic and Statistical Manual of Mental Disorders–5th edition* (*DSM-5*; American Psychiatric Association, 2013) symptom counts for attention-deficit/hyperactivity disorder (ADHD) estimates that all children with ADHD (e.g., confirmed diagnosis based on Schedule for Affective Disorders and Schizophrenia for School-Age Children [K-SADS]; Kaufman et al., 1997) have a probability of .51 of having ADHD, and that all children without ADHD (e.g., also confirmed based on K-SADS) have a .49 probability of having ADHD, the model would have perfect discrimination but very poor calibration. Models that are well calibrated are also not necessarily useful at classification tasks (e.g., Spiegelhalter, 1986). An example of this is a model that always predicts the base rate of an event. For example, one could take a "bet the base rate" model (e.g., Youngstrom, Halverson, Youngstrom, Lindhiem, & Findling, 2017) that uses the *DSM-5* prevalence estimate of ADHD in children of 5%. If the model, based on the base rate, assigns a prediction that each child has a 5% probability of having ADHD, the model will have perfect calibration (in the sense that the predicted risk matches the observed rate exactly and sometimes referred to as "mean" calibration; see Van Calster et al., 2016) but low discrimination (the estimates do not differentiate between children who have ADHD and those who do not). In this case, the model would be perfectly calibrated but does not provide any new information above and beyond the base rate. Such a model would indeed be useless in a classification task. But the model could still be enormously useful to an individual patient for a prognostic task. If a patient is diagnosed with a potentially fatal disease, knowing the base rate for recovery would be immensely useful. For example, it would make an important difference to a patient if the base rate of recovery is .99 and not .20.

Risk assessment tools can also be well-calibrated for one range of the instrument but poorly calibrated for another range. For example, Duwe and Freske (2012) describe the Minnesota Sex Offender Screening Tool–3, which has good calibration for values below 40% but for values above 40%, the instrument overestimates the risk of sexual recidivism. Because 99% of offenders have scores below 40%, the instrument can still be very useful. It is also useful to know that for the 1% of offenders who have scores above 40%, the results overestimate the actual risk of recidivism. Similar instances of miscalibration just for the high-risk range of risk calculators are not uncommon (e.g., Fazel et al., 2016).

Metrics to Evaluate Calibration

Meteorologists have had an interest in calibration for over 100 years, at least as far back as 1906 (Lichtenstein, Fischhoff, & Phillips, 1982). In 1950, a meteorologist at the U.S. Weather Bureau named Glenn Brier (1950) proposed an approach to verify the accuracy of probabilistic forecasts of binary weather events. Brier was concerned with forecasters who "hedge" and forecast a value other than what he or she actually thinks. Early weather forecasters would hedge because consumers of weather forecasts (the public) were more critical if it rained when no rain was forecast (and perhaps caught without an umbrella) than vice versa (Lichtenstein et al., 1982). Brier proposed a statistic, commonly referred to as a Brier score, for a set of probabilistic (0.0-1.0) predictions for binary events coded "1" or "0." Given a particular probabilistic prediction $\hat{p}$ where the true underlying probability is p, the Brier score for such a prediction is defined as $(p-\hat{p})^2$. The Brier score for a set of forecasts is simply the mean squared error, so that given a set of predictions $\hat{p}_1,\ldots,\hat{p}_n$ with true probabilities $p_1,\ldots,p_n$, the Brier score is

$$\frac{1}{n}\sum_{i=1}^{n}\left(p_i-\hat{p}_i\right)^2.$$

A Brier score for single forecast of 80% (.80) for an event that occurs ("1") would be $(1-.80)^2 = .04$. For a forecast of 40% (.40) for an event that does not occur ("0"), the single Brier score would be $(0-.40)^2 = .16$. The average Brier score for the two predictions is (.04 + .16)/2 = .10. Two or more Brier scores can be averaged for a set of predictions to derive an overall measure of accuracy for evaluative purposes (e.g., comparing two individual forecasters on a set of predictions). Lower scores indicate better accuracy but there are no established standards as the definition of a "good" Brier score depends on the base rate for the event that is being forecast as well as the difficulty of the forecast. As a benchmark, flipping a fair coin (average prediction of 0.5) for a set of forecasts would result in a Brier score of 0.25. The Brier score remains a popular measure of overall accuracy and is used widely with diverse applications ranging from medical research (Rufibach, 2010) to forecasting tournaments (e.g., the Good Judgment Project; Tetlock, Mellers, Rohrbaugh, & Chen, 2014).

It is important to highlight that a Brier score is not a measure of calibration per se, but rather a measure of overall accuracy. However, Brier scores are important because they can be decomposed into components that specifically assess discrimination and calibration (Rufibach, 2010; Spiegelhalter, 1986). Spiegelhalter's z-test statistic measures the calibration aspect of the Brier score (Redelmeier, Bloch, & Hickam, 1991; Rufibach, 2010). Given a set of observations (binary outcomes, 0 or 1) $y_1,\ldots,y_n$ along with associated predicted probabilities $\hat{p}_1,\ldots,\hat{p}_n$, Spiegelhalter's z-test statistic is defined as,

$$z=\frac{\sum_{i=1}^{n}\left(y_i-\hat{p}_i\right)\left(1-2\hat{p}_i\right)}{\sqrt{\sum_{i=1}^{n}\left(1-2\hat{p}_i\right)^2\hat{p}_i\left(1-\hat{p}_i\right)}}$$

and asymptotically follows a standard normal distribution. The null hypothesis of the associated statistical test is that the model is well-calibrated, so statistically significant scores (i.e., $z<-1.96$ or $z>1.96$) generally indicate poor calibration. As with any z test, the interpretation of p values should take into consideration the sample size and resulting statistical power. Conventional alpha levels (e.g., .05 or .01) may not be appropriate for very small or very large samples. Regardless of the sign of the z value (positive or negative), larger absolute values of z indicate a greater degree of miscalibration.

Another commonly used measure of calibration is the Hosmer–Lemeshow (HL) goodness-of-fit statistic (Hosmer & Lemeshow, 1980; Schmid & Griffith, 2005). To calculate the statistic, one first divides the set of observations into G groups where G is selected by the user. The default for G is 10 for most software programs but typically ranges from 2 to 10 depending on the sample size and range of predictions. For a given value of G, the HL statistic is defined as,

$$H=\sum_{i=0}^{1}\sum_{g=1}^{G}\frac{\left(OBS_{ig}-EXP_{ig}\right)^2}{EXP_{ig}}=\sum_{g=1}^{G}\frac{\left(OBS_{0g}-EXP_{0g}\right)^2}{EXP_{0g}}+\frac{\left(OBS_{1g}-EXP_{1g}\right)^2}{EXP_{1g}}$$

where OBS_{ig} denotes the number of events in group g with outcome i (where $i = 0$ or 1) and EXP_{ig} is the expected number of events in group g with outcome i, determined by

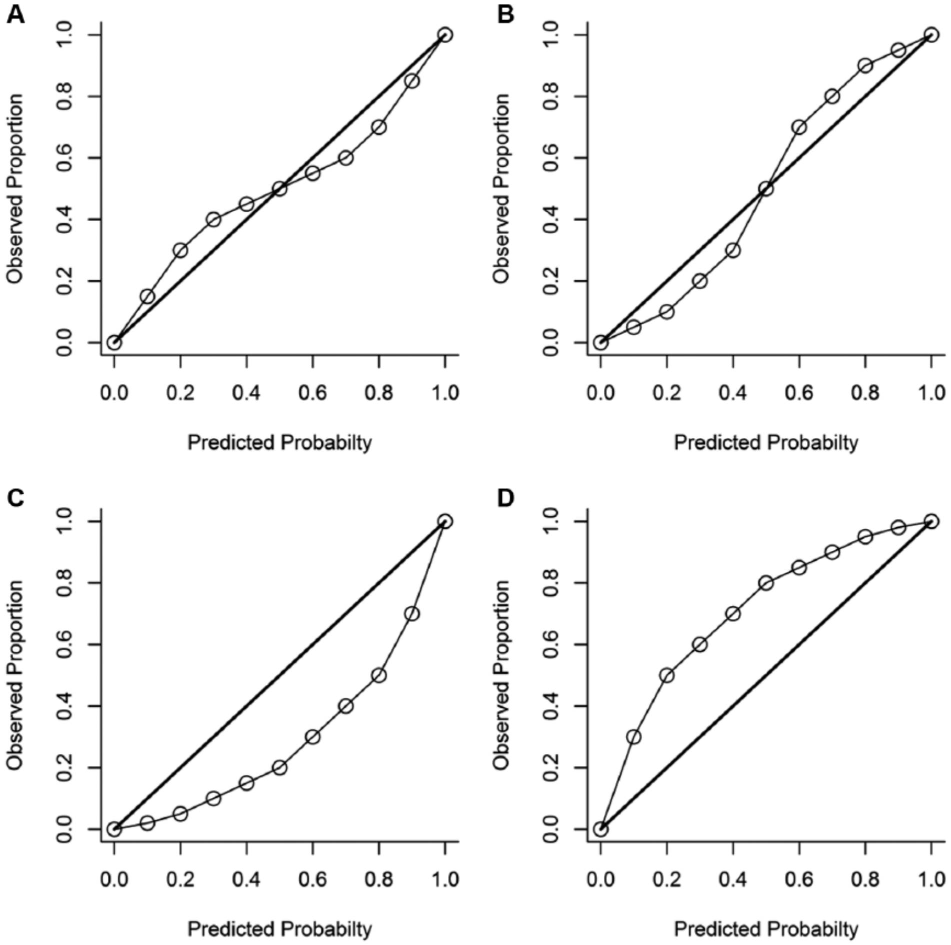

Figure 1. Common patterns of miscalibration: (A) Overextremity, (B) Underextremity, (C) Overprediction, and (D) Underprediction.
Note. Perfect calibration is represented by the diagonal line.

taking the average predicted probability across all observations in group g . The HL statistic has an approximate chi-squared distribution with $G - 2$ degrees of freedom and is designed to measure the fit between predicted proportions and those actually observed. Like Spiegelhalter's z-test statistic, large values of the HL statistic (low corresponding p values) indicate poor calibration. A common benchmark for poor calibration for the HL statistic is $p < .05$, but like all inferential statistics this is somewhat arbitrary and may not be appropriate for very small or very large data sets. The HL statistic has other disadvantages including dependence on the selected G, low statistical power for small samples, and no information on the direction or pattern on miscalibration (Steyerberg, 2008).

Due to the limitations of summary statistics, calibration should ideally be evaluated using more than one statistic in addition to a graphical approach (e.g., Griffin & Brenner, 2004; Lichtenstein et al., 1982). Graphical approaches are essentially qualitative and can be used to supplement summary statistics such as Spiegelhalter's z-test statistic or the HL statistic. The typical graphical approach involves plotting predicted probabilities (x-axis) against actual probabilities (y-axis) using a smoothing technique. Such calibration plots typically include a diagonal reference line signifying perfect calibration. An early approach described by Copas (1983a) is to use a smoothed histogram of predicted probabilities and observed proportions. A very similar approach is to use a locally weighted scatterplot smoothing (LOWESS) plot with predicted probabilities on the x-axis and LOWESS-smoothed actual probabilities (i.e., observed proportions) on the y-axis (e.g., Duwe & Freske, 2012). One benefit of graphical techniques is that patterns of miscalibration can easily be identified. See Figure 1 for sample calibration plots showing common patterns of miscalibration.

Additional measures and tests of calibration are specific to external validation (i.e., testing the model performance on an external data set). Models generally tend to be well calibrated when applied to the original data set, sometimes referred to as "apparent calibration" (Harrell, 2015; Steyerberg, 2008). This is due partly to overfitting and presents an overly optimistic view of how the model will perform when applied to a new data set (Babyak, 2004; Moons et al., 2012). Internal validation methods such as *k*-fold and leave-one-out cross-validation can detect calibration problems caused by overfitting but are insufficient if there are important differences (e.g., base rates or regression coefficients) between the original data set and an external data set (Bleeker et al., 2003). Internal validation may be sufficient for "temporal transportability" in which a model is applied to future patients at the same clinic but not for "geographic transportability" in which a model is applied to a new clinic (König, Malley, Weimar, Diener, & Ziegler, 2007). External validation is critical if a model developed in one setting is to be used in another setting (Bleeker et al., 2003; König et al., 2007; Moons et al., 2012; Youngstrom et al., 2017). Miscalibration is common when a model is applied to a new data set, evidenced by differences in the intercept and/or slope of the calibration plots between the original data set and external data set (Harrell, 2015; Steyerberg, 2008). Differences in the intercept indicate problems with overprediction or underprediction, whereas differences in the slope indicate problems with overextremity or underextremity. A recalibration test can be used to determine whether the calibration plot for the model deviates from an intercept of 0 and slope of 1. If necessary, the new model can be recalibrated to achieve "logistic calibration" with intercept of 0 and slope of 1 (Harrell, 2015; Van Calster et al., 2016).

Patterns of Miscalibration

Terms such as "overconfidence" are often used to describe miscalibration but can be misleading due to their imprecision (Griffin & Brenner, 2004; Lichtenstein et al., 1982; Moore & Healy, 2008). There are two common ways for a set of predictions to be "overconfident." The first is for the set of predictions to be consistently higher than the actual proportion of outcomes across the full range of predictions. In this scenario, the calibration curve is entirely below the reference line of a calibration plot. This pattern is often referred to as "overprediction," although the term "specific overconfidence" has also been used. The second is for the set of predictions to be consistently too close to 0 or 1. In this scenario, the calibration curve is above the reference line of a calibration plot for values below 0.5 and below the reference line for values above 0.5, often looking like a backward "s" (see Figure 1). This pattern is often referred to as "overextremity" and is typical of statistical overfitting. The term "generic overconfidence" has also been used to describe this pattern.

Similarly, there are two common ways for a set of predictions to be "underconfident." The first is for the set of predictions to be consistently lower than the actual proportion of outcomes across the full range of predictions. In this scenario, the calibration curve is entirely above the reference line of a calibration plot. This pattern is often referred to as "underprediction," although the term "specific underconfidence" has also been used. The second is for the set of predictions to be consistently too close to 0.5. In this scenario, the calibration curve is below the reference line of a calibration plot for values below 0.5 and above the reference line for values above 0.5 (somewhat "s" shaped). This pattern is often referred to as "underextremity," although the term "generic underconfidence" has also been used.

Illustrations

Table 1 shows generated estimates from four hypothetical models. Model A has both good discrimination and good calibration. With a cut-score (also referred to as a cutoff or threshold) of 0.50, the model can accurately predict all 10 events (AUC = 1.00). The model also has good calibration. Events that were forecast in the 90% to 100% likelihood range (0.90-1.00) occurred 100% of the time. Events that were forecast in the 0% to 10% likelihood range (0.00-0.10) occurred 0% of the time. Formal calibration statistics also indicate good calibration (Spiegelhalter's $z = -.73$, $p = .4680$; HL goodness-of-fit = 0.53, $p = .9975$). See Figure 2 for calibration plot of Model A.

Model B has good discrimination but poor calibration. As with Model A, Model B can accurately predict all 10 events with a cut-score of .50 (AUC = 1.00). However, the model has poor calibration. Events that were forecast at 60% likelihood (.60) occurred 100% of the time. Events that were forecast at 40% likelihood (.40) occurred 0% of the time. Formal calibration statistics also indicate poor calibration (Spiegelhalter's $z = -2.58$, $p = .0098$; HL goodness-of-fit = 6.67, $p = .0357$). See Figure 2 for calibration plot of Model B.

Model C has poor discrimination but good calibration. With a cut-score of .50, the model can only predict the events at the level of chance (AUC = .50). However, the model has good calibration. Events that were forecast in the 45% to 55% likelihood range (.45 to .55) occurred 50% of the time. Formal calibration statistics also indicate good calibration (Spiegelhalter's $z = .04$, $p = .9681$; HL goodness-of-fit = 0.00, $p = 1.0000$). See Figure 2 for calibration plot of Model C.

Model D has both poor discrimination and poor calibration. As with Model C, Model D can only predict the events at the level of chance with a cut-score of .50 (AUC = .50). The model also has poor calibration. Events that were forecast in the 5% to 15% likelihood range (.5 to .15) occurred 50% of the time. Events that were

Table 1. Discrimination and Calibration of Four Hypothetical Models.

	Model A	Model B	Model C	Model D	Actual event
Prediction 1	0.94	0.60	0.49	0.89	Yes (1)
Prediction 2	0.96	0.60	0.51	0.91	Yes (1)
Prediction 3	0.95	0.60	0.50	0.90	Yes (1)
Prediction 4	0.94	0.60	0.50	0.11	Yes (1)
Prediction 5	0.96	0.60	0.50	0.09	Yes (1)
Prediction 6	0.04	0.40	0.49	0.89	No (0)
Prediction 7	0.06	0.40	0.51	0.91	No (0)
Prediction 8	0.05	0.40	0.50	0.90	No (0)
Prediction 9	0.04	0.40	0.50	0.11	No (0)
Prediction 10	0.06	0.40	0.50	0.09	No (0)
Brier score (MSE)	0.0026	0.1600	0.2500	0.4101	
Hosmer–Lemeshow	0.53	6.67	0.00	17.98	
p	.9975	.0357	1.0000	.0030	
Spiegelhalter's z	−0.7258	−2.5800	0.0400	4.2268	
p	.4680	.0098	.9681	.0000	
AUC	1.0000	1.0000	0.5000	0.5000	
Discrimination	Good	Good	Poor	Poor	
Calibration	Good	Poor	Good	Poor	

Note. MSE = mean squared error; AUC = area under the ROC curve.

forecast in the 85% to 95% likelihood range (.85 to .95) occurred 50% of the time. Formal calibration statistics also indicate poor calibration (Spiegelhalter's z = 4.23, p = .0000; HL goodness-of-fit = 17.98, p = .0030). See Figure 2 for calibration plot of Model D.

Examples of Calibration in Psychology

Cognitive Psychology

The domain in psychology where calibration has possibly received the greatest attention is the cognitive psychology literature on overconfidence. In this context, the probabilistic estimates are generated by human forecasters rather than mathematical models. One of the most common findings on calibration in this context is that, although there are some exceptions, people (including experts) tend to be overconfident in their judgments and predictions. Overconfidence is studied in various ways. One way is by asking people to make a judgment/prediction (e.g., "Will it rain tomorrow? [YES/NO]") followed by asking them to rate their confidence (as a percentage) in their answer. Confidence in a dichotomous judgment (yes/no) expressed as a percentage (0% to 100%) is mathematically identical to a probabilistic prediction (0.00-1.00) of a dichotomous event. When making such probability judgments, a person would be considered well-calibrated if his or her responses match the relative frequency of occurrence (i.e., judgments of 70% are correct 70% of the time). The most persistent finding in the overconfidence literature is that of widespread overprecision (overextremity) of judgments/predictions. That is, people tend to make predictions with too extreme probability judgments (Moore & Healy, 2008). Researchers have examined the reasons for overconfidence. Calibration can improve in response to feedback (e.g., Bolger & Önkal-Atay, 2004), suggesting that overconfidence may result, in part, from cognitive biases. Anchoring and adjustment occurs when someone uses insufficient adjustment from a starting point or prior probability, known as the anchor (Tversky & Kahneman, 1974). Other research suggests additional cognitive biases may be involved in overconfidence, including the confirmation bias (Hoch, 1985; Koriat, Lichtenstein, & Fischhoff, 1980) and base rate neglect (Koehler et al., 2002).

In addition, although many experts have been shown to have poor calibration in their predictions or judgments including clinical psychologists (Oskamp, 1965), physicians (Koehler et al., 2002), economists (Koehler et al., 2002), stock market traders and corporate financial officers (Skala, 2008), lawyers (Koehler et al., 2002), and business managers (Russo & Schoemaker, 1992), there are some cases of experts showing good calibration. For instance, experts in weather forecasting (Murphy & Winkler, 1984), horse race betting (Johnson & Bruce, 2001), and playing the game of bridge (Keren, 1987) have shown excellent calibration (but see Koehler et al., 2002, for exceptions to these exceptions). The reasons for high calibration among experts in these domains likely include that they receive clear, consistent, and timely outcome feedback (Bolger & Önkal-Atay, 2004). It is unlikely, however, that clinical psychologists receive timely outcome feedback regarding their diagnostic decisions or other clinical predictions, suggesting that clinical psychologists are likely to be poorly calibrated in their judgments and predictions

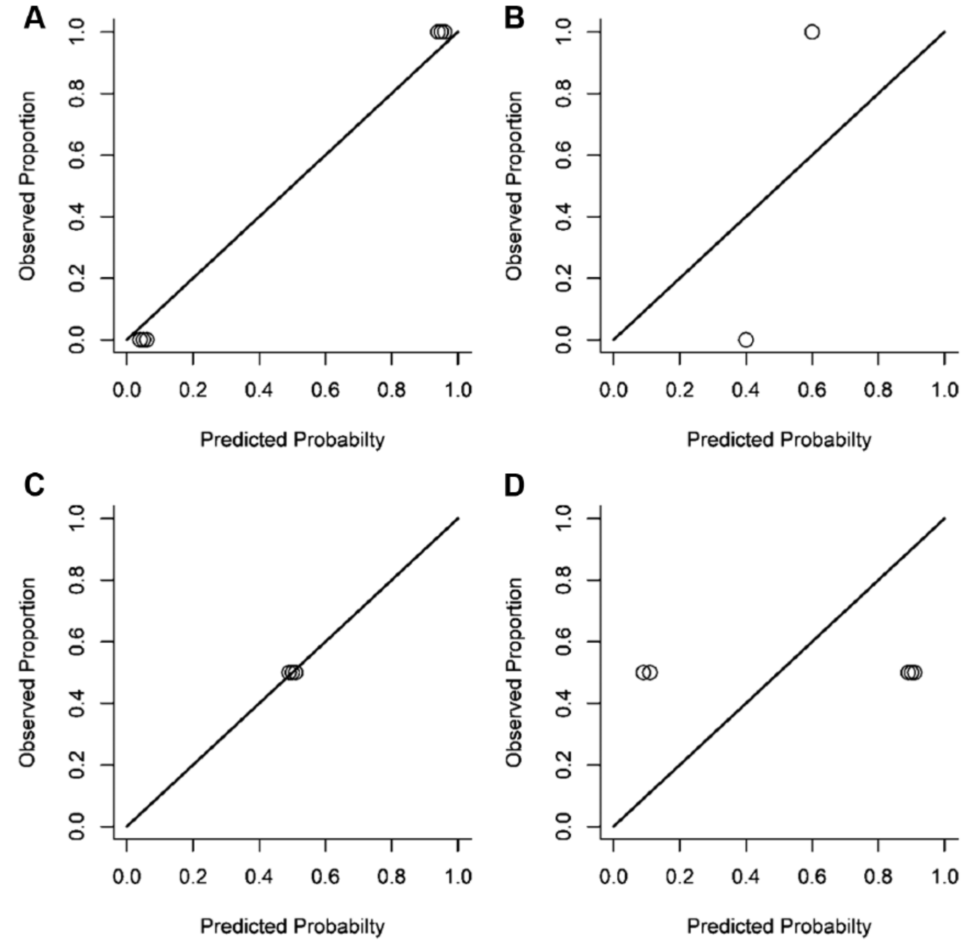

Figure 2. Calibration plots for Model A (good discrimination and good calibration), Model B (good discrimination and poor calibration), Model C (poor discrimination and good calibration), and Model D (poor discrimination and poor calibration).
Note. Perfect calibration is represented by the diagonal line.

(consistent with findings from Oskamp, 1965 and assertions by Smith & Dumont, 1997). Nevertheless, overprecision/overextremity is common even among experts (Moore & Healy, 2008), and a meta-analysis of five domains of experts (including weather forecasters) showed that all domains of experts show systematic miscalibration (Koehler et al., 2002).

Koehler et al. (2002) found that physicians routinely ignored the base rate in their judgments and predictions. The researchers found that the physicians' pattern of miscalibration was strongly related to the base rate likelihood of the target event (consistent with findings from Winkler & Poses, 1993), and to a lesser extent, to the discriminability of the relevant hypotheses (i.e., the strength of evidence for making a correct decision). Specifically, they found that physicians showed underprediction when the base rate was high and discriminability was high (e.g., ICU survival), fair calibration when the base rate was low and discriminability was high (e.g., myocardial infarction), and overprediction when the base rate was very low and discriminability was low (e.g., pneumonia). The authors interpreted these findings as suggesting that physicians' decisions were strongly related to the support for the relevant hypotheses (e.g., the representativeness heuristic, availability heuristic, and confirmation bias) rather than to the base rate or discriminability of the hypothesis. For a discussion of physicians' errors in probabilistic reasoning by neglecting base rates, see Eddy (1982).

We suspect it is likely that clinical psychologists, similar to other health care professionals (and people in general), show poor calibration because of neglecting the base rate likelihoods of diagnoses and events. Miscalibration is particularly likely when base rates are low (e.g., the diagnosis of ADHD in primary care) leading to overprediction or when base rates are high (e.g., noncompliance in 2-year-olds) leading to underprediction (Koehler et al., 2002).

Clinical Psychology

We are aware of very few studies that have examined calibration in clinical psychology. As described earlier, some studies have examined the calibration of predictive models of criminal recidivism (Duwe & Freske, 2012; Fazel et al., 2016). Other studies have examined the calibration of predictive models of depression and suicide. Perlis (2013) examined predictive models of treatment resistance in response to pharmacological treatment for depression. Calibration was evaluated using the HL goodness-of-fit test and concluded that for such data sets, logistic regression models were often better calibrated than predictive models based on machine learning approaches (naïve Bayes, random forest, support vector machines). The author hypothesized that this finding may be at least partially attributed to the fact that such machine learning approaches construct models that maximize discrimination regardless of calibration. Thus, as discussed earlier, when building models and evaluating their accuracy, it is important for prediction models to consider both discrimination and calibration.

One other recent study (Lindhiem, Yu, Grasso, Kolko, & Youngstrom, 2015) evaluated the calibration of the posterior probability of diagnosis index. The posterior probability of diagnosis index measures the probability that a patient meets or exceeds a diagnostic threshold (Lindhiem, Kolko, & Yu, 2013). In a pediatric primary-care sample, the index performed well in terms of discrimination (AUC) but was not well calibrated (Lindhiem et al., 2015). With too many predictions below .05 and above .95, the calibration plot (backward "S") was characteristic of an "overconfident" model. A recalibration method was proposed to recalibrate the index to improve calibration while maintaining the strong discrimination.

Finally, a recent study describing the development of a risk calculator to predict the onset of bipolar disorder in youth evaluated the tool for calibration (Hafeman et al., 2017). Similar to some of the recidivism/relapse prediction tools described earlier, the risk calculator was well calibrated in the validation sample across a fairly narrow range of predictions (0.00 to 0.24). External validation was not examined to test the performance of the tool in a clinical setting.

Methods to Improve Calibration

Unless properly accounted for, many predictive modeling techniques can produce predictions with adequate discrimination but poor calibration (Jiang et al., 2012). Left unchecked, many statistical models including machine learning methods and logistic regression are susceptible to overfitting, wherein a model is constructed, that is, overly biased to the particular data at hand and therefore does not generalize well (i.e., produces inaccurate predictions) on new data (Babyak, 2004). As noted above, overextremity is a type of overconfidence in which predictions are consistently too close to 0.0 or 1.0. An example is a model for which predictions of events at 95% probability actually occur 75% of the time, while predictions of events at 5% probability actually occur 25% of the time. If a model is to be used for individualized purposes, such as making a prediction for a particular patient, it is critical to ensure that the model predictions have adequate calibration. Calibration can be examined and potentially improved either during the model development phase or later when the model is applied to new data.

Overfitting during the model development phase often causes the calibration slope to be less than one when applied to an external data set (Harrell, 2015). The value of the new slope, referred to as the "shrinkage factor" can be applied to the model to correct this type of miscalibration. For example, regression coefficients can be "preshrunk" so that the slope of the calibration plot will remain one on external validation (Copas, 1983b; Harrell, 2015). If an external data set is not available and the shrinkage factor cannot be determined, internal cross-validation techniques are generally considered the next-best option. Internal cross-validation involves partitioning the data set into *k* nonoverlapping groups of roughly equal size; one group is then held out, while a model is constructed using the remaining *k*-1 groups and the accuracy is evaluated by making predictions on the hold-out set and comparing with the observed values. The process is repeated for each of the *k* groups yielding *k* estimates of accuracy that are averaged to provide a final measure. This entire process can then be repeated on different predictive modeling methods to compare performance and/or on the same method under a variety of slight alterations, often governed by a tuning parameter that is designed to trade off overfitting and underfitting. Under this setup, the same modeling approach is applied a number of times, each time with a different value of the tuning parameter thereby yielding different sets of predicted values. Internal cross-validation errors are computed for each level of the tuning parameter and final predictions are generally taken as those that minimize the error. This method of tuning a particular model aims to strike a balance in using enough information in the data to make useful insights without being overly biased to the data in the sample at hand (overfitting) and is essential for learning methods that are highly sensitive to tuning parameter values such as "lasso" (Tibshirani, 1996) or gradient boosting (Friedman, 2001). As noted earlier, internal cross-validation may be sufficient to ensure that a model has "temporal transportability" (e.g., same clinic but with future patients) but insufficient to ensure "geographic transportability" (e.g., different clinic; König et al., 2007). There is no substitute for external validation if a model is to be applied in a new setting (Bleeker et al., 2003; König et al., 2007; Moons et al., 2012).

Other traditional methods to improve the calibration of predictive models include binning, Platt scaling, and isotonic regression, but there is evidence that these approaches can fail to improve calibration (Jiang et al., 2012). Jiang et al. have proposed a method called adaptive calibration of predictions, an approach that uses individualized confidence intervals, to calibrate probabilistic predictions. In this method, all of the predictions from a training data set that fall within the 95% confidence interval of the target prediction (the estimate to be calibrated) are averaged, with this new value replacing the original estimate.

Another method to recalibrate probabilistic estimates uses Bayes' theorem and a training data set (Lindhiem et al., 2015). The general formula is as follows:

$$p(\text{event} \mid \text{estimate}) = \frac{p(\text{estimate} \mid \text{event})\, p(\text{event})}{p(\text{estimate})}$$

where p(event | estimate) is the recalibrated estimate, "event" is the actual occurrence of the event in the training data set (coded "0" or "1"), and "estimate" is the original model estimate (ranging from 0.00 to 1.00). p(event) is the base rate of the event in the training data set. A k-nearest neighbor algorithm is used to smooth the estimates because not all model estimates can be represented in the training data set. The method has been shown in some instances to enhance the calibration of probabilistic estimates without reducing discrimination.

There are also various methods for improving the calibration of "expert" predictions based on clinical judgment (Lichtenstein et al., 1982). It is well-documented that diagnosticians, including mental health professionals, tend to be overconfident in their professional judgments (e.g., Smith & Dumont, 1997). Methods to correct for this predictable pattern of miscalibration include instruction in the concept of calibration, warnings about overconfidence, and feedback about performance (e.g., Stone & Opel, 2000). The evidence for the effectiveness of these and other methods is mixed, and depends on numerous factors including the difficulty of subsequent prediction tasks.

A Clinical Example

Next, we illustrate the importance of calibration for developing predictive models to screen for mental health diagnoses, using the ABACAB data set (N = 819; see Youngstrom et al., 2005). ABACAB was designed as an assessment project with specific aims including establishing the base rate of bipolar spectrum and other disorders in community mental health, examining the accuracy of clinical diagnoses compared with research diagnoses based on a consensus review process with trained interviewers using recommended semistructured methods, and assessing the diagnostic accuracy of rating scales and risk factors that were recused from the diagnostic formulation process (see Youngstrom et al., 2005, for additional details).

For this example, we used several supervised learning techniques to predict the likelihood of a bipolar disorder diagnosis from brief screening data. The outcome variable was a consensus diagnosis of any bipolar disorder based on the K-SADS (Kaufman et al., 1997). The predictor variables were 11 items from the parent-completed Mood Disorder Questionnaire (Wagner et al., 2006). We explored a variety of supervised learning techniques to assign predicted probabilities for a bipolar diagnosis based on the Mood Disorder Questionnaire items, including both classic statistical approaches (naïve Bayes and logistic regression) as well as more modern tree-based methods (classification and regression tree [CART] and random forests).

Naïve Bayes

Naïve Bayes assigns a probability of class membership to each observation based on Bayes rule under the assumption that predictor variables (features) are independent conditional on belonging to a particular class (e.g., Kononenko, 1990).

Logistic Regression

Logistic regression, a classic parametric statistical method, is a generalized linear model in which the log odds of the response are assumed to be a linear function of the features (e.g., McCullagh & Nelder, 1989).

Classification and Regression Tree

We used a single regression tree built according to the popular CART methodology (Breiman, Friedman, Stone, & Olshen, 1984). In this approach, the feature space is recursively partitioned by splitting observations into response-homogeneous bins. After construction, the probability that a particular observation's response is "1" is predicted by averaging the response values (0's and 1's) located within the corresponding bin.

Random Forest

Finally, random forest (Breiman, 2001) represents an extension of CART that involves resampling of the single CART tree in which several (in our case, 500) regression trees are constructed—each built with a bootstrap sample of the original data—and then decorrelated by randomly determining subspaces in which splits may occur. Final predictions are obtained by taking a simple average of the predictions generated by each individual tree.

Model Building, Internal Cross-Validation, and Model Comparisons

Each of the four predictive models were built and evaluated using the full data set and also with fivefold internal

Table 2. Discrimination and Calibration for Four Predictive Models.

	Naïve Bayes	CART	Random forest	Logistic regression
Full data set (no cross-validation)				
Brier score (MSE)	.2045	.1202	.0893	.1153
Hosmer–Lemeshow	153809.83	0.00	34.83	13.84
p	.0000	1.0000	.0001	.1802
Spiegelhalter's z	41.8198	0.0000	−3.9950	0.2616
p	.0000	1.0000	.0001	.7936
AUC	.8283	.8047	.9229	.8380
With fivefold cross-validation				
Brier score (MSE)	.2085	.1366	.1316	.1266
Hosmer–Lemeshow	138421.81	81.62	36.06	11.49
p	.0000	.0000	.0001	.3208
Spiegelhalter's z	42.6386	2.9713	2.6733	2.3604
p	.0000	.0000	.0075	.0183
AUC	.8149	.7194	.7734	.8380

Note. CART = classification and regression tree; MSE = mean squared error; AUC = area under the ROC curve.

cross-validation. Both approaches were taken to illustrate the importance of internal cross-validation in evaluating model performance. Model results are summarized in Table 2 and Figures 3 and 4. Significant *p* values for both Spiegelhalter's *z* and HL chi-square test suggest that naïve Bayes and random forest were poorly calibrated when using both the full data set and fivefold internal cross-validation. The calibration plots—Figures 3 and 4—also reflect poor calibration. Both Spiegelhalter's *z* and the HL chi-square test indicate perfect calibration with the CART-based regression tree when the full data set was used to train the model, but poor calibration using fivefold internal cross-validation. The perfect calibration in the full data set is an artifact of the CART algorithm and highlights the importance of evaluating model performance using internal cross-validation (for which CART produced poorly calibrated predictions). Logistic regression was generally well-calibrated in both the full data set and using fivefold internal cross-validation, as evidence by the calibration plots and statistics. Although the *p* value for Spiegelhalter's *z* was below .05 using fivefold internal cross-validation, the HL chi-square remained nonsignificant ($p = .3208$).

When to Prioritize Calibration Over Discrimination

The statistical analyses performed must match the research question or practical problem that it is intended to address (Spiegelhalter, 1986). In general terms, discrimination analyses are most relevant for classification tasks, whereas calibration is important for predictive/prognostic tasks.

Discrimination Analyses for Classification Tasks

Discriminant analysis is appropriate when the goal is to assign a group of patients into established categories (Spiegelhalter, 1986). Mental health diagnoses have traditionally been treated as dichotomous categories (e.g., presence/absence of diagnosis), lending conveniently to discriminant analysis. Analyses are typically done using ROC analyses and the associated AUC statistic (sometimes also called the *c* statistic or *c*-index). An ROC curve is a plot of sensitivity (true positive rate) on the *y*-axis and 1 − specificity (false positive rate) on the *y*-axis across the full range of cut-scores. The AUC statistic is a measure of the area under the ROC curve. A purely random model would have an AUC of 0.5 and a perfect model would have an AUC of 1.0. The AUC statistic is therefore oftentimes a useful metric for evaluating the accuracy of diagnostic testing (Cook, 2007). In this case, there is objective, and naturally dichotomous, group membership. The goal is to correctly assign group membership to as many patients as possible.

Calibration Analyses for Predictive Tasks

Calibration becomes important when estimating the true underlying probability of a *particular* event, such as is the case in risk assessment (Cook, 2007). Examples include risk calculators designed to estimate the likelihood of violent reoffending among inmates released from prison (e.g., Fazel et al., 2016). In these instances, stochastic (i.e., random) processes are assumed to be at play and predictions must therefore be made in probabilistic terms. As noted earlier, a model that correctly predicts all events ("1"s) to occur with probability greater than 0.5 and all nonevents ("0"s) to occur with probability less than 0.5 will have perfect discrimination even though the individual predictions may have little meaning. If correct overall classification is all that matters, then the individual predicted probabilities are of no consequence. On the other hand, when evaluating the accuracy of a probabilistic prediction

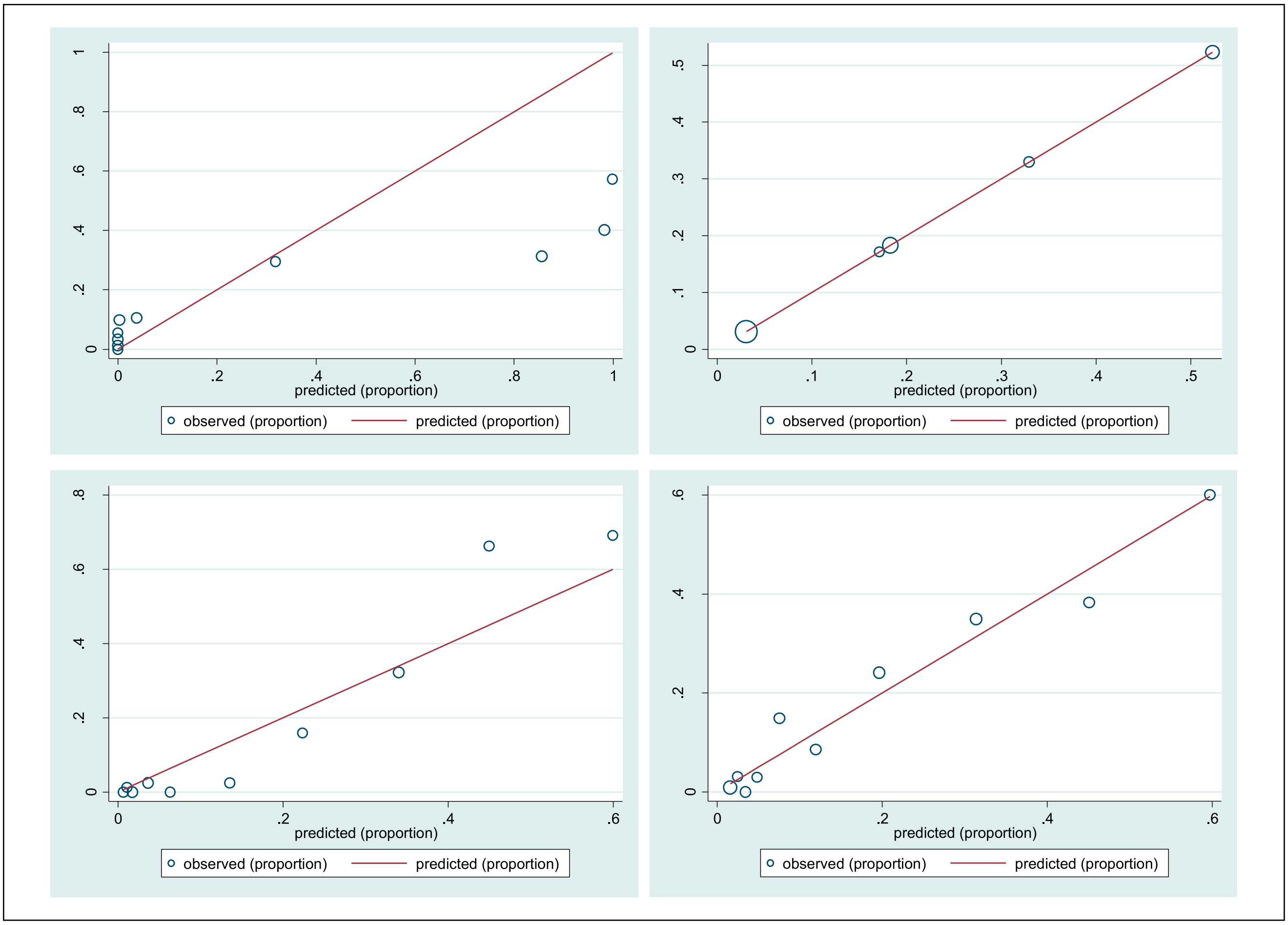

Figure 3. Calibration plots for four models using the full data set (no internal cross-validation): Naïve Bayes (top, left), CART (top, right), random forest (bottom, left), and logistic regression (bottom, right).
Note. CART = classification and regression tree. Perfect calibration is represented by the diagonal line.

of a specific clinical outcome for a particular patient, model fitting and discriminant analyses alone are not appropriate (Spiegelhalter, 1986). Calibration is also important when model output will otherwise be used by a physician and/or patient to make a clinical decision (e.g., Steyerberg et al., 2010). This is because the individual predicted probabilities now have practical significance (and may be of paramount concern to the individual patient). It is now important that a value of 0.25, for example, can be interpreted to mean that the patient has a 25% chance of developing a disease. In other words, prediction requires the assessment of absolute risk and not just relative risk (e.g., beta weights, odds ratios). Examples include risk calculators designed to estimate the likelihood of violent reoffending among inmates released from prison (e.g., Fazel et al., 2016). In selecting a predictive model, it is important that calibration be high for a new data set (as can be evaluated by internal cross-validation, e.g.,) and not just for the training data set (Schmid & Griffith, 2005).

Implications for Clinical Psychology

Calibration is particularly relevant for predicting recidivism (e.g., Fazel et al., 2016). It has been shown that predictive tools can overestimate the probability of recidivism even though they have good discrimination (e.g., Duwe & Freske, 2012). For a tool designed to assess the risk of recidivism to be well-calibrated, the predicted probabilities must closely match actual recidivism rates. This clearly has real-world consequences for those facing parole decisions. Calibration is also important when expert diagnosticians assign confidence values to diagnoses for which no definitive (100% accurate) tests are available. For example, for all diagnoses in which confidence is assigned in the .70 to .79 ranges, the diagnosis should in fact be present roughly 70% to 79% of the time. It is well documented that individuals tend to be overconfident on difficult and moderately difficult tasks and underconfident on easy tasks. Furthermore, calibration tends to be better when the base rate of the event being predicted is close to 50% and worse for rare events (Lichtenstein et al.,

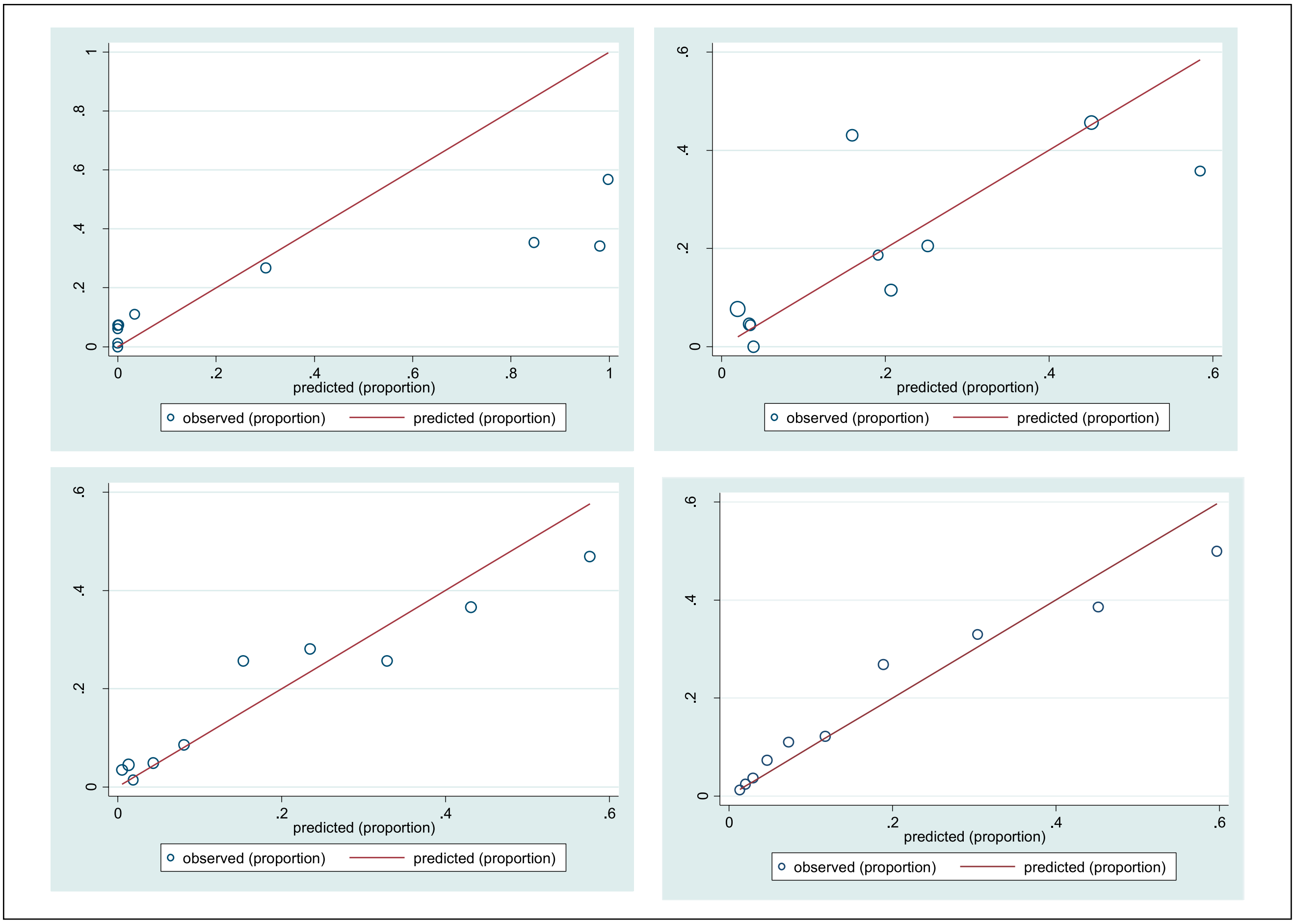

Figure 4. Calibration plots for four models with fivefold internal cross-validation: Naïve Bayes (top, left), CART (top, right), random forest (bottom, left), and logistic regression (bottom, right).
Note. CART = classification and regression tree. Perfect calibration is represented by the diagonal line.

1982). Calibration is also critical to consider when developing models to predict the onset of disorder in the future, the probability of remission, or the probability or relapse. Finally, calibration is a key construct to consider when developing any psychological test for which the results are presented as probability values and interpreted as such.

Summary and Conclusion

In summary, although calibration has received less attention from clinical psychologists over the years, relative to discrimination, it is a crucial aspect of accuracy to consider for all prognostic models, clinical judgments, and psychological testing. It will become increasingly important in the future as advances in computing lead to an increasing reliance on probabilistic assessments and risk calculators. Good calibration is crucial for making personalized clinical judgments, diagnostic assessments, and to psychological testing in general. Better attention to calibration, in addition to discrimination, should promote more accurate predictive models, more effective individualized care in clinical psychology, and improved psychological tests.

Acknowledgments

The authors would like to thank Teresa Treat, Richard Viken, and Danella Hafeman for their feedback on an early draft on the manuscript, and Jordan L. Harris for assistance with formatting.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported in part by NIH R01 MH066647 (PI: E. Youngstrom).

Notes

1. The term "response bias" has a different meaning in the area of validity scale research where it is used to describe biased patterns of responses to self-report or test items such as overreporting, underreporting, fixed reporting, and random reporting.
2. Another important side note is that the term "calibration" has been defined differently in related fields. In the area of human judgment and metacognitive monitoring, for example, calibration has been defined as the fit between a rater's judgment of performance on a task and actual performance (e.g., Bol & Hacker, 2012). Throughout this article, we use the term "calibration" only in the specific sense defined in the preceding paragraph, namely, the degree to which a probabilistic prediction of an event reflects the true underlying probability of that event.

ORCID iD

Isaac T. Petersen https://orcid.org/0000-0003-3072-6673

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: American Psychiatric Association.

Babyak, M. A. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine*, *66*, 411-421.

Baird, C., & Wagner, D. (2000). The relative validity of actuarial- and consensus-based risk assessment systems. *Children and Youth Services Review*, *22*, 839-871. doi:10.1016/S0190-7409(00)00122-5

Bleeker, S. E., Moll, H. A., Steyerberg, E. W., Donders, A. R. T., Derksen-Lubsen, G., Grobbee, D. E., & Moons, K. G. M. (2003). External validation is necessary in prediction research: A clinical example. *Journal of Clinical Epidemiology*, *56*, 826-832. doi:10.1016/S0895-4356(03)00207-5

Bol, L., & Hacker, D. J. (2012). Calibration research: Where do we go from here? *Frontiers in Psychology*, *3*, 229. doi:10.3389/fpsyg.2012.00229

Bolger, F., & Önkal-Atay, D. (2004). The effects of feedback on judgmental interval predictions. *International Journal of Forecasting*, *20*(1), 29-39. doi:10.1016/S0169-2070(03)00009-8

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32. doi:10.1023/A:1010933404324

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. Boca Raton, FL: CRC press.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1-3.

Cook, N. R. (2007). Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*, *115*, 928-935. doi:10.1161/circulationaha.106.672402

Copas, J. B. (1983a). Plotting *p* against *x*. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *32*, 25-31.

Copas, J. B. (1983b). Regression, prediction and shrinkage. *Journal of the Royal Statistical Society: Series B (Methodological)*, *45*, 311-354.

Duwe, G., & Freske, P. J. (2012). Using logistic regression modeling to predict sexual recidivism: The Minnesota Sex Offender Screening Tool-3 (MnSOST-3). *Sexual Abuse: A Journal of Research and Treatment*, *24*, 350-377. doi:10.1177/1079063211429470

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 249-267). Cambridge, England: Cambridge University Press.

Fazel, S., Chang, Z., Fanshawe, T., Långström, N., Lichtenstein, P., Larsson, H., & Mallett, S. (2016). Prediction of violent reoffending on release from prison: Derivation and external validation of a scalable tool. *Lancet Psychiatry*, *3*, 535-543. doi:10.1016/S2215-0366(16)00103-6

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*, 1189-1232.

Griffin, D., & Brenner, L. (2004). Perspectives on probability judgment calibration. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 177-199). Malden, MA: Blackwell.

Hafeman, D. M., Merranko, J., Goldstein, T. R., Axelson, D., Goldstein, B. I., Monk, K., . . . Birmaher, B. (2017). Assessment of a person-level risk calculator to predict new onset bipolar spectrum disorder in youth at familial risk. *JAMA Psychiatry*, *74*, 841-847. doi:10.1001/jamapsychiatry.2017.1763

Harrell, F. E., Jr. (2015). *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis* (2nd ed.). New York, NY: Springer.

Hoch, S. J. (1985). Counterfactual reasoning and accuracy in predicting personal events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 719-731. doi:10.1037/0278-7393.11.1-4.719

Hosmer, D. W., & Lemeshow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics: Theory and Methods*, *9*, 1043-1069.

Jiang, X., Osl, M., Kim, J., & Ohno-Machado, L. (2012). Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association*, *19*, 263-274. doi:10.1136/amiajnl-2011-000291

Johnson, J. E. V., & Bruce, A. C. (2001). Calibration of subjective probability judgments in a naturalistic setting. *Organizational Behavior and Human Decision Processes*, *85*, 265-290. doi:10.1006/obhd.2000.2949

Kaufman, J., Birmaher, B., Brent, D., Rao, U. M. A., Flynn, C., Moreci, P., . . . Ryan, N. (1997). Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime Version (K-SADS-PL): Initial reliability and validity data. *Journal of the American Academy of Child & Adolescent Psychiatry*, *36*, 980-988. doi:10.1097/00004583-199707000-00021

Keren, G. (1987). Facing uncertainty in the game of bridge: A calibration study. *Organizational Behavior and Human Decision Processes*, *39*, 98-114. doi:10.1016/0749-5978(87)90047-1

Koehler, D. J., Brenner, L., & Griffin, D. (2002). The calibration of expert judgment: Heuristics and biases beyond the laboratory. In T. Gilovich, D. Griffin & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 687-715). Cambridge, England: Cambridge University Press.

König, I. R., Malley, J. D., Weimar, C., Diener, H. C., & Ziegler, A. (2007). Practical experiences on the necessity of external validation. *Statistics in Medicine*, *26*, 5499-5511. doi:10.1002/sim.3069

Kononenko, I. (1990). Comparison of inductive and naive Bayesian learning approaches to automatic knowledge acquisition. In J. Breuker, N. Guarino, J. N. Kok, R. López de Mántaras, J. Liu, R. Mizoguchi, . . . N. Zhong (Series Eds.), *Frontiers in artificial intelligence and applications: Vol 8: Current trends in knowledge acquisition* (pp. 190-197). Amsterdam, Netherlands: IOS.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 107-118. doi:10.1037/0278-7393.6.2.107

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306-334). Cambridge, England: Cambridge University Press.

Lindhiem, O., Kolko, D. J., & Yu, L. (2013). Quantifying diagnostic uncertainty using item response theory: The Posterior Probability of Diagnosis Index. *Psychological Assessment*, *25*, 456-466. doi:10.1037/a0031392

Lindhiem, O., Yu, L., Grasso, D. J., Kolko, D. J., & Youngstrom, E. A. (2015). Adapting the Posterior Probability of Diagnosis (PPOD) Index to enhance evidence-based screening: An application to ADHD in primary care. *Assessment*, *22*, 198-207. doi:10.1177/1073191114540748

Macmillan, N. A., & Creelman, C. D. (1990). Response bias: Characteristics of detection theory, threshold theory, and "nonparametric" indexes. *Psychological Bulletin*, *107*, 401-413.

McCullagh, P., & Nelder, J. A. (1989). Generalized linear models, (2nd ed. London, England: Chapman & Hall.

Moons, K. G., Kengne, A. P., Woodward, M., Royston, P., Vergouwe, Y., Altman, D. G., & Grobbee, D. E. (2012). Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio) marker. *Heart*, *98*, 683-690. doi:10.1136/heartjnl-2011-301246

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*, 502-517. doi:10.1037/0033-295X.115.2.502

Murphy, A. H., & Winkler, R. L. (1984). Probability forecasting in meterology. *Journal of the American Statistical Association*, *79*, 489-500. doi:10.2307/2288395

Oskamp, S. (1965). Overconfidence in case-study judgments. *Journal of Consulting Psychology*, *29*, 261-265. doi:10.1037/h0022125

Perlis, R. H. (2013). A clinical risk stratification tool for predicting treatment resistance in major depressive disorder. *Biological Psychiatry*, *74*(1), 7-14. doi:10.1016/j.biopsych.2012.12.007

Redelmeier, D. A., Bloch, D. A., & Hickam, D. H. (1991). Assessing predictive accuracy: How to compare Brier scores. *Journal of Clinical Epidemiology*, *44*, 1141-1146. doi:10.1016/0895-4356(91)90146-Z

Rufibach, K. (2010). Use of Brier score to assess binary predictions. *Journal of Clinical Epidemiology*, *63*, 938-939. doi:10.1016/j.jclinepi.2009.11.009

Russo, J. E., & Schoemaker, P. J. H. (1992). Managing overconfidence. *Sloan Management Review*, *33*(2), 7-17.

Schmid, C. H., & Griffith, J. L. (2005). Multivariate classification rules: Calibration and discrimination. In P. Armitage & T. Colton (Eds.), *Encyclopedia of biostatistics* (Vol. 5, pp. 3491-3497). Chichester, England: John Wiley. doi:10.1002/0470011815.b2a13049

Skala, D. (2008). Overconfidence in psychology and finance: An interdisciplinary literature review. *Bank i Kredyt*, *4*, 33-50.

Smith, D., & Dumont, F. (1997). Eliminating overconfidence in psychodiagnosis: Strategies for training and practice. *Clinical Psychology: Science and Practice*, *4*, 335-345. doi:10.1111/j.1468-2850.1997.tb00125.x

Spiegelhalter, D. J. (1986). Probabilistic prediction in patient management and clinical trials. *Statistics in Medicine*, *5*, 421-433. doi:10.1002/sim.4780050506

Steyerberg, E. W. (2008). *Clinical prediction models: A practical approach to development, validation, and updating*. New York, NY: Springer.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., . . . Kattan, M. W. (2010). Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology*, *21*, 128-138. doi:10.1097/EDE.0b013e3181c30fb2

Stone, E. R., & Opel, R. B. (2000). Training to improve calibration and discrimination: The effects of performance and environmental feedback. *Organizational Behavior and Human Decision Processes*, *83*, 282-309. doi:10.1006/obhd.2000.2910

Tetlock, P. E., Mellers, B. A., Rohrbaugh, N., & Chen, E. (2014). Forecasting tournaments: Tools for increasing transparency and improving the quality of debate. *Current Directions in Psychological Science*, *23*, 290-295. doi:10.1177/0963721414534257

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*, 267-288.

Treat, T. A., & Viken, R. J. (2012). Measuring test performance with signal detection theory. In H. Cooper (Ed.), *APA Handbook of Research Methods in Psychology: Vol 1: Foundations, planning, measures, and psychometrics* (pp. 723-744). Washington, DC: American Psychological Association. doi:10.1037/13619-038

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124-1131. doi:10.1126/science.185.4157.1124

Van Calster, B., Nieboer, D., Vergouwe, Y., De Cock, B., Pencina, M. J., & Steyerberg, E. W. (2016). A calibration hierarchy for risk models was defined: From utopia to empirical data. *Journal of Clinical Epidemiology*, *74*, 167-176. doi:10.1016/j.jclinepi.2015.12.005

Wagner, K. D., Hirschfeld, R., Findling, R. L., Emslie, G. J., Gracious, B., & Reed, M. (2006). Validation of the Mood Disorder Questionnaire for bipolar disorders in adolescents. *Journal of Clinical Psychiatry*, *67*, 827-830. doi:10.4088/JCP.v67n0518

Winkler, R. L., & Poses, R. M. (1993). Evaluating and combining physicians' probabilities of survival in an intensive care unit. *Management Science*, *39*, 1526-1543. doi:10.1287/mnsc.39.12.1526

Wolraich, M. L., Hannah, J. N., Baumgaertel, A., & Feurer, I. D. (1998). Examination of DSM-IV critieria for attention deficit/hyperactivity disorder in a county-wide sample. *Journal of Developmental & Behavioral Pediatrics*, *19*, 162-168.

Youngstrom, E. A., Halverson, T. F., Youngstrom, J. K., Lindhiem, O., & Findling, R. L. (2017). Evidence-based assessment from simple clinical judgments to statistical learning: Evaluating a range of options using pediatric bipolar disorder as a diagnostic challenge. *Clinical Psychological Science*. Advance online publication. doi:10.1177/2167702617741845

Youngstrom, E. A., Meyers, O. I., Demeter, C., Kogos Youngstrom, J., Morello, L., Piiparinen, R., . . . Calabrese, J. R. (2005). Comparing diagnostic checklists for pediatric bipolar disorder in academic and community mental health settings. *Bipolar Disorders*, *7*, 507-517. doi:10.1111/j.1399-5618.2005.00269.x