

# I-spline Smoothing for Calibrating Predictive Models

Yuan Wu\*, Xiaoqian Jiang\*, Jihoon Kim, Lucila Ohno-Machado<sup>1</sup>

Division of Biomedical Informatics  
University of California at San Diego  
La Jolla, California 92093

## Abstract

*We proposed the I-spline Smoothing approach for calibrating predictive models by solving a nonlinear monotone regression problem. We took advantage of I-spline properties to obtain globally optimal solutions while keeping the computational cost low. Numerical studies based on three data sets showed the empirical evidences of I-spline Smoothing in improving calibration (i.e., 1.6x, 1.4x, and 1.4x on the three datasets compared to the average of competitors-Binning, Platt Scaling, Isotonic Regression, Monotone Spline Smoothing, Smooth Isotonic Regression) without deterioration of discrimination.*

## Introduction

Learning models focused on maximizing *discrimination* (i.e., the ability to separate positive cases from negative cases) often ignore *calibration*, which relates to the *correctness* of predicted values. However, the latter aspect is important to medical decision-making, since clinicians may use predictive model estimates as surrogates to individualized risk scores<sup>2,3</sup>. However, if the predictive model is not calibrated (e.g., raw outputs of Support Vector Machine are used to represent risk), the decisions may be wrong. As molecular markers from genomics and proteomics are increasingly considered in predictive models and become available to consumers<sup>4,5</sup>, calibration is even more crucial to enable reliable risk assessment diagnosis, and prognosis based on individual genomics and proteomics<sup>6-8</sup>.

Such challenge is getting even more critical when medicine is becoming more and more “personalized”, for which predicted scores need to faithfully reflect the probability of outcomes of individual patients for best performance. Unfortunately, many popular predictive models (i.e., Decision Trees, and Naive Bayes classifiers) do not optimize calibration<sup>9</sup>. To improve on a predictive model calibration without deteriorating its discrimination, we need to develop novel and practical approaches.

## Related work

There are a number of attempts towards improving the calibration of predictive models. To understand the pros and cons of each of them, we briefly review state-of-the-art methods. A most intuitive idea is called *binning*<sup>10</sup>, which sorts and groups predicted scores into bins, and replaces the predicted scores as the fraction of positive cases within each bin, so as to reduce the discrepancy between predictions and the unknown true probabilities. This intuitive approach, although capable of improving calibration, may decrease discrimination due to the loss of rankings within each bin. Platt suggested a rescaling model<sup>11</sup> that uses an additional logistic regression model to refit predictions (i.e., predictor variables) against class labels (i.e., the target variable). This method can convert arbitrary predicted scores (e.g., outputs of a Support Vector Machine model) into estimated probabilities. However, the approach is parametric, and it has limited ability to calibrate predictive models. Zadrozny and Elkan proposed another approach<sup>12</sup> named Isotonic Regression (IR), a model to minimize the mean squared errors while respecting monotonic constraints, i.e., keeping the order of the predictions and hence no altering the ROC curve. These authors showed that IR can achieve superior performance over some baseline methods. A problem with this approach is that the lack of smoothness might decrease its generalization performance<sup>13</sup>.

Yet another approach using a monotone spline smoothing technique proposed by Wang and Li<sup>14</sup> offers both smoothness and non-parametricity. The method is theoretically sound but it is complicated to implement in practice (high dimension of the problem due to large number of spline knots, complicated constraints for the monotonicity

---

<sup>1</sup> \* These authors contributed equally to this manuscript.

of the estimation, etc.), and as acknowledged by the authors: more efficient ways for choosing the penalty term parameters are necessary. A most recent attempt by Jiang et al<sup>13</sup> uses a two-step approach to obtain a smoothed isotonic regression: (1) fitting an isotonic regression model to obtain the knot points; (2) use these knot points to re-fit a Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) model. However, this heuristic approach has no theoretical guarantees of optimality. The following table summarizes the characteristics of a number of popular calibration approaches. The last row of the table summarizes the properties of I-spline Smoothing, a new calibration method introduced in this paper.

**Table 1:** Summarization of popular calibration approaches.

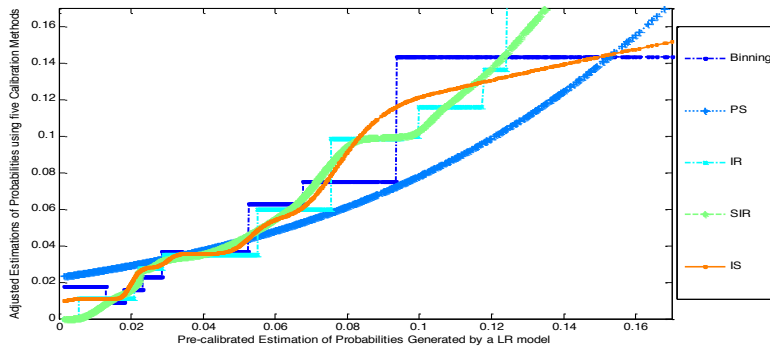
	Monotonic	Non-parametric	Non-exponential complexity	Continuous
Binning <sup>10</sup>		x	x	
Platt scaling <sup>11</sup>	x		x	
Isotonic Regression <sup>12</sup>	x	x	x	
Smooth Isotonic Regression <sup>13</sup>	x	x	x	x
Monotone Spline Smoothing <sup>14</sup>	x	x		x
I-spline Smoothing	x	x	x	x

All above mentioned techniques<sup>11-14</sup>, except for binning, aim at solving a monotonic regression problem

$$\sum_{k=1}^n (c_k - f(p_k))^2$$

$$s.t. \quad f(p_k) \leq f(p_{k+1}), \quad 0 \leq f(p_k) \leq 1, \quad \forall k,$$

where  $p_k$  is the pre-calibrated estimate from the model, and  $c_k$  is the observed binary outcome. Therefore, another way to understand differences and challenges among various methods is to look at their specific assumptions of functions. For example, Platt assumes  $f(\cdot)$  to be an inverse logit function<sup>11</sup>, Isotonic Regression assumes a free-form  $f(\cdot)$  that only has values at  $p_k$ , and Monotone Spline Smoothing assumes  $f(\cdot)$  to be a natural cubic spline<sup>14</sup>. Due to these assumptions, each method has its own challenge, as we discussed earlier. This paper intends to introduce a new approach for which we assume the function  $f(\cdot)$  to be a member of the cubic spline family with an I-spline basis. Thanks to compelling properties, the computation of I-spline Smoothing can be made much easier when compared to optimizing natural cubic splines using Monotone Spline Smoothing<sup>14</sup>. The following figure illustrates the adjusted estimations of probabilities using five different calibration approaches and the predictions of a LR model on a linearly separable data set.



**Figure 1:** Illustration of calibration functions of four different approaches, including Binning, Platt Scaling (PS), Isotonic regression (IR), Smooth Isotonic Regression (SIR) and our proposed method I-spline Smoothing (IS).

## Methodology

I-splines are monotone splines that have the most obvious applications in monotone nonlinear regression problem, as discussed by Ramsay<sup>15</sup>. Though attractive in their simple expression and theoretical properties, very few articles described real applications of I-spline techniques. Lu et al<sup>16</sup> and Wu<sup>17</sup> used I-splines in solving Maximum Likelihood Estimation (MLE) problems. In this paper, we identified a new application of I-splines to solve calibration problems. Let us start with an introduction of I-spline basic concepts. The  $l$ -th order I-splines based on a knot sequence are defined by Ramsay<sup>15</sup> as

$$I_i^l(s) = 1, \quad (1a)$$

$$I_i^l(s) = \int_L^s M_i^l(t) dt, 1 < i \leq q, \quad (1b)$$

with  $L \leq s \leq U$ , where  $L$  and  $U$  are the left and the right end knots of the knot sequence, respectively. The number of I-splines is  $q = n^{(knot)} + l + 1$ , where  $n^{(knot)}$  is the number of interior knots (i.e., the knots that are not end knots in the knot sequence). Note that  $i$  corresponds to the index of I-splines. Wu<sup>17</sup> identified an interesting relationship: each  $M_i^l(t)$  is related to a B-spline<sup>18</sup>  $N_i^l(t)$  such that  $M_i^l(t) = (l \cdot N_i^l(t)) / (u_{i+1} - u_i)$ , and therefore B-splines could be used to construct I-splines

$$I_i^l(s) = \sum_{m=i}^q N_m^{l+1}(s). \quad (2)$$

It is easy to compute I-splines using formula (2), since B-splines can be efficiently computed, and are already available in statistical packages. De Boor<sup>19</sup> showed that  $\int_L^U M_i^l(t) dt = 1$ , and that each  $M_i^l(t)$  is nonnegative, which implies that I-splines in (1a) and (1b) are monotone and have function values between 0 and 1. The non-parametricity, monotonicity, and the range constraint between  $[0,1]$  altogether make I-splines good candidates to model distribution functions.

Given pre-calibrated prediction probabilities  $P = \{p_1, \dots, p_n\}$  and class labels  $C = \{c_1, \dots, c_n\}$ , we now show how to use I-spline based smoothing techniques to calibrate predictive models by solving a nonlinear monotone least square regression problem.

Define  $\Omega = \{f : f(t) = \sum_{i=1}^q \alpha_i I_i^l(t), \alpha_i \geq 0, \sum_{i=1}^q \alpha_i \leq 1\}$  as the space of I-spline functions. The monotone least square regression finds  $f^* \in \Omega$  that minimizes:

$$\sum_{k=1}^n (c_k - f(p_k))^2. \quad (3)$$

As mentioned before, I-splines are monotone and their values are between 0 and 1, the constraints for  $\{\alpha_i\}_1^q$  in  $\Omega$  guarantee each  $f \in \Omega$  is monotone with function values lay between 0 and 1. Given a knot sequence, I-splines are fixed, hence this monotone regression problem is actually minimizing (3) with respect to I-spline coefficients  $\{\alpha_i\}_1^q$  with constraints  $\alpha_i \geq 0$  for  $i = 1, \dots, q$  and  $\sum_{i=1}^q \alpha_i \leq 1$ . We can rewrite this problem as a maximization problem:

$$\begin{aligned}
& \arg \max_{\{\alpha_i\}_i^q} \left[ -\sum_{k=1}^n (c_k - \sum_{i=1}^q \alpha_i I_i^l(p_k))^2 \right] \\
& \text{s.t. } \alpha_i \geq 0, \forall i, \quad \sum_{i=1}^q \alpha_i \leq 1
\end{aligned} \tag{4}$$

The next question is how to pick interior knots, which is always critical to any spline-based technique. Intuitively, there are two general rules:

- (i) More interior knots should be added to allow more flexibility;
- (ii) More interior knots should be added where samples are frequently observed.

The second rule is easy to follow, after deciding the number of interior knots we could position according to the sample percentiles. But how to decide the number of interior knots is really decided on a case-by-case basis. Ramsay<sup>15</sup> mentioned very few interior knots are necessary, say, 1 or 2, for I-spline based regression problems. However, both Lu et al.<sup>16</sup> and Wu<sup>17</sup> chose the cube root of sample size as the number of interior knots for the MLE based spline estimations, and their experiments supported this choice. Given our sample size  $n$ , we used  $\max\{1, (n^{1/3} - 4)\}$  as the number of interior knots, which works best for the proposed estimation in this paper when cubic splines are applied.

The computing for the maximization problem (4) can be done by a generalized gradient projection algorithm<sup>1</sup>. First we rewrite the constraints in (4) as  $\bar{X}\alpha \leq \bar{y}$ , where  $\bar{X} = (x_1, x_2, \dots, x_{q+1})^T$  with  $x_1 = (-1, 0, \dots, 0)^T$ ,  $x_2 = (0, -1, 0, \dots, 0)^T$ ,  $\dots$ ,  $x_q = (0, \dots, 0, -1)^T$ ,  $x_{q+1} = (1, \dots, 1)^T$ ;  $\alpha = (\alpha_1, \dots, \alpha_q)^T$ ; and  $\bar{y} = (0, \dots, 0, 1)^T$ . If some I-spline coefficients equal 0 or all coefficients sum up to 1, then we say their according constraints are active and let  $\bar{X}\alpha = \bar{y}$  represent all active constraints, where rows of  $\bar{X}$  and  $\bar{y}$  are from a subset of rows of  $\bar{X}$  and  $\bar{y}$ .  $\bar{X}$  is used to facilitate the computation.

Initially we put integers representing active constraints in vector  $\Lambda$  (including indexes of I-spline coefficients for those equal to 0 and  $(q+1)$  when all coefficients sum up to 1). The vector  $\Lambda$  with  $r$  scalars corresponds to an  $r \times q$  matrix  $\bar{X}$ . For example, if  $\Lambda = (2, 1, (q+1))$ , then  $\bar{X} = (x_2, x_1, x_{q+1})^T$ .

We denote the target function  $[-\sum_{k=1}^n (c_k - \sum_{i=1}^q \alpha_i I_i^l(p_k))^2]$  in (4) as  $F(\alpha)$  with. Let  $\partial F(\alpha)$  and  $H(\alpha)$  be gradient and Hessian matrix of  $F(\alpha)$  with respect to  $\alpha$ , respectively. Let  $W = -H(\alpha) + \gamma I$ , where  $I$  is an identity matrix, and  $\gamma$  is set to be large enough to make  $W$  positive definite. With that introduced, the generalized gradient projection algorithm is implemented as Algorithm 1.

## Experiments

To compare different calibration methods, we used two indices, the Area Under the ROC Curve (AUC)<sup>20</sup> and the decile-based Hosmer-Lemeshow goodness-of-fit test (HL-test)<sup>21</sup>, to assess model's discrimination and calibration, respectively. We compared the original logistic regression model and four calibration approaches: Platt Scaling (PS), Isotonic Regression (IR), Smooth Isotonic Regression (SIR), and I-spline Smoothing (IS). Because I-spline Smoothing is strictly monotonic, we expect it would not decrease the AUC of an input model. The experiment is to verify our expectation, and evaluate if IS has the potential to improve calibration.

## Data

We used three real-world data sets to evaluate the performance of proposed I-spline Smoothing calibration method. Table 2 summarizes the data in terms of their feature dimension, sample size, training to test set ratio, and short descriptions of the data.

**Algorithm 1:** Generalized gradient projection algorithm to solve I-spline Smoothing.

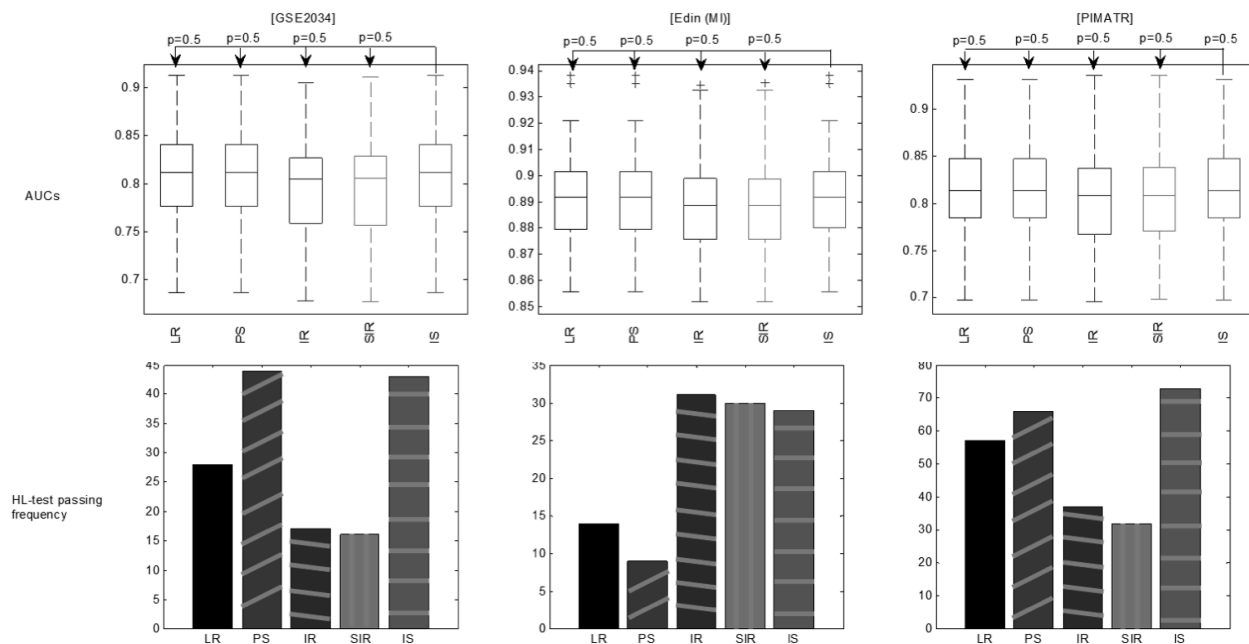
<p><b>Step 1</b> (Computing the feasible search direction) Compute</p> $\bar{d} = (d_1, d_2, \dots, d_q) = \{I - W^{-1} \bar{X}^T (\bar{X} W^{-1} \bar{X}^T)^{-1} \bar{X}\} W^{-1} \partial F(\alpha).$
<p><b>Step 2</b> (Forcing the updated <math>\alpha</math> to fulfill the constraints) Compute</p> $\gamma = \min \left\{ \min_{i: d_i < 0} \left\{ -\frac{\alpha_i}{d_i} \right\}, \frac{1 - \sum_{i=1}^q \alpha_i}{\sum_{i=1}^q d_i} \right\}, \quad \text{if } \sum_{i=1}^q d_i > 0,$ $\gamma = \min \left\{ \min_{i: d_i < 0} \left\{ -\frac{\alpha_i}{d_i} \right\} \right\}, \quad \text{otherwise.}$
<p>The execution guarantees that <math>\alpha_i + \gamma d_i \geq 0</math> for <math>i = 1, 2, \dots, q</math>, and <math>\sum_{i=1}^q (\alpha_i + \gamma d_i) \leq 1</math>.</p>
<p><b>Step 3</b> (Updating the solution by Step-Halving line search<sup>1</sup>) Find the smallest integer <math>\omega</math> starting from 0 such that</p> $F(\alpha + (1/2)^\omega \gamma \bar{d}) \geq F(\alpha).$ <p>Replace <math>\alpha</math> by <math>\hat{\alpha} = \alpha + \min\{(1/2)^\omega \gamma, 0.5\} \cdot \bar{d}</math>.</p>
<p><b>Step 4</b> (Updating <math>\Lambda</math>, <math>\bar{X}</math>) If <math>k = 0</math> and <math>\gamma \leq 0.5</math>, modify <math>\Lambda</math> by adding indexes of new I-spline coefficients when these new coefficients become 0, or adding <math>(q+1)</math> when <math>\sum_{i=1}^q \alpha_i</math> becomes 1, and modify <math>\bar{X}</math> accordingly,</p>
<p><b>Step 5</b> (Checking the stopping criterion) If <math>\ \bar{d}\  \geq \varepsilon</math>, for small <math>\varepsilon</math>, go to Step 1, otherwise compute</p> $\lambda = (\bar{X} W^{-1} \bar{X}^T) \bar{X} W^{-1} \partial F(\alpha).$ <p>(i) If the <math>j</math>-th component <math>\lambda_j \geq 0</math> for all <math>j</math>, set <math>\hat{\alpha} = \alpha</math> and stop.</p> <p>(ii) If there is at least one <math>j</math> such that <math>\lambda_j &lt; 0</math>, let <math>j^* = \arg \min_{j: \lambda_j &lt; 0} \{\lambda_j\}</math>, then remove the <math>j^*</math>-th component from <math>\Lambda</math> and remove the <math>j^*</math>-th row from <math>\bar{X}</math>, and go to Step 1.</p>

**Table 2:** Summarization of the data used to conduct the experiments.

	Feature dimension	Sample size	Training / Test ratio	Note
GSE2034 <sup>22</sup>	15	209	6 / 4	Breast cancer data sets from NCBI Gene Expression Omnibus (GEO) used to construct a decision support system for predicting reoccurrences of breast cancer using extracted gene expression features. We followed Osl et al <sup>23</sup> to select features.
Edin (MI) <sup>24</sup>	48	1,253	6 / 4	This data contain clinical and electrocardiographic information about 500 patients with and without myocardial infarction (MI) admitted with chest pain into an emergency department in Sheffield, England. The study was to determine which, and how many, data items are required to construct a decision support system for early diagnosis of acute myocardial infarction <sup>24</sup> .
PIMATR <sup>25</sup>	8	768	6 / 4	Pima Indians Diabetes data set from National Institute of Diabetes and Digestive and Kidney Diseases. The population lives near Phoenix, Arizona, USA, and all patients are females at least 21 years old of Pima Indian heritage <sup>25</sup> . The diagnostic, binary-valued variable investigated is whether the patient shows signs of diabetes according to World Health Organization criteria (i.e., if the 2 hour post-load plasma glucose was at least 200 mg/dl at any survey examination or if found during routine medical care) <sup>25</sup> .

## Results

Following aforementioned ratios, we randomly divided the data into training and test sets, 100 times, to evaluate the performances of methods. Figure 1 below shows the box plots of AUCs in the first row, where each color represents a method and every subplot corresponds to one data set, as denoted in the caption. In the second row of Figure 2, we illustrate the rate of ‘passing’ the HL-test at the significance level of 0.05 for each method in all three data sets.



**Figure 2:** Illustration of AUCs and HL-tests of all five methods in comparison using three different data.

The figure visually demonstrates that all five models were comparable in terms of AUCs, while I-spline Smoothing stood out in terms of calibration performance. Table 3 lists actual values of these comparisons. For GSE2034 data, I-spline Smoothing ranked second in calibration since its HL-test passing rate was 43%, compared to LR (28%), PS (17%), IR (44%), and SIR (16%). The AUCs of I-spline Smoothing were not significantly smaller than any of the other methods. Note that we used one-tailed paired t-test to compare different AUCs.

The results on Edin (MI) data showed similar patterns for calibration. I-spline Smoothing had the third highest rate of passing the HL-test (29%), which is close to SIR (30%) and IR (30%), and better than LR (14%) and PS (9%). Regarding discrimination, the AUCs of I-spline Smoothing were not significantly smaller than any other models. Finally, the experiments using PIMATR data, I-spline Smoothing outperformed all the other methods in calibration with a HL-test passing rate of 73%, followed by PS (66%), LR (57%), IR (37%), and SIR (32%). The AUCs of I-spline Smoothing are not smaller than any of those of other models.

**Table 3:** Performance of different models using different data.

	Logistic Regression (LR)	Platt Scaling (PS)	Isotonic Regression (IR)	Smooth Isotonic Regression (SIR)	I-spline Smoothing (IS)
GSE2034 (AUC±std) / (HL-test pass rate)	(0.81±0.04) / (0.28)	(0.81±0.04) / (0.44)	(0.80±0.05) / (0.17)	(0.80±0.05) / (0.16)	(0.81±0.04) / (0.43)
Edin (MI) (AUC±std) / (HL-test pass rate)	(0.89±0.02) / (0.14)	(0.89±0.02) / (0.09)	(0.89±0.02) / (0.31)	(0.89±0.02) / (0.30)	(0.89±0.02) / (0.29)
PIMATR (AUC±std) / (HL-test pass rate)	(0.82±0.05) / (0.57)	(0.82±0.05) / (0.66)	(0.80±0.05) / (0.37)	(0.81±0.05) / (0.32)	(0.82±0.05) / (0.73)

## **Discussion and Conclusion**

In this paper, we introduced a novel method called I-spline Smoothing (IS) for calibrating predictive models as an alternative to existing approaches. The advantages of IS lie in the following aspects: (1) IS is a non-parametrically monotonic transformation, that provides more flexibility in calibrating predictive models, when compared to parametric approaches like Platt Scaling. (2) IS is globally optimized, as opposed to Smooth Isotonic Regression which is a heuristic approach. (3) IS is easy to implement, compared to Monotone Spline Smoothing. The results using three real-world data sets showed advantages of IS in both discrimination and calibration, empirically. In these experiments, IS demonstrates superior calibration without significant deterioration of discrimination. Although these experiments were conducted at small scale, they suggest that future research on IS is warranted and may improve calibration.

A limitation of this technique is that we need to set the number of interior knots heuristically. Even though it worked well in our experiments, a theoretical result describing systematic ways to choose the number of interior knots is needed.

## **Acknowledgement**

The authors were funded in part by the National Library of Medicine (R01LM009520) and NHLBI (U54 HL10846). We thank Dr. Fraser and Dr. El-Kareh for making the data sets available for this study.

## Reference

1. Jamshidian M. On algorithms for restricted maximum likelihood estimation. *Computational Statistics and Data Analysis* 2004;45: 137-57.
2. Kamath PS, Kim W. The model for end stage liver disease (MELD). *Hepatology* 2007;45(3): 797-805.
3. Racowsky C, Ohno-Machado L, Kim J, Biggers JD. Is there an advantage in scoring early embryos on more than one day? *Human Reproduction* 2009;24(9): 2104.
4. Altman RB, Miller KS. 2010 Translational bioinformatics year in review. *Journal of the American Medical Informatics Association* 2010;18(4): 358-66.
5. Sarkar IN, Butte AJ, Lussier YA, Tarczy-Hornoch P, Ohno-Machado L. Translational bioinformatics: linking knowledge across biological and clinical realms. *Journal of the American Medical Informatics Association*. 2011;18(4): 354-57.
6. Jiang X, Osl M, Kim J, Ohno-Machado L. Calibrating Predictive Model Estimates to Support Personalized Medicine. *Journal of the American Medical Informatics Association* (Epub ahead of print) 2011.
7. Osl M, Ohno-Machado L, Baumgartner C, Tilg B, Dreiseitl S. Improving calibration of logistic regression models by local estimates. *AMIA Annual Symposium Proceedings*, 2008:535-39.
8. Wei W, Visweswaran S, Cooper GF. The application of naive Bayes model averaging to predict Alzheimer's disease from genome-wide data. *Journal of the American Medical Informatics Association*. 2011;18(4): 370-75.
9. Zadrozny B, Elkan C. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. *International Conference on Machine Learning*, 2001:609–16.
10. Zadrozny B, Elkan C. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. *The Eighteenth International Conference on Machine Learning*, 2001:609–16.
11. Platt JC. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*. 1999: 61-74.
12. Zadrozny B, Elkan C. Transforming classifier scores into accurate multiclass probability estimates. *The Eighth International Conference on Knowledge Discovery and Data Mining*, 2002:694–99.
13. Jiang X, Osl M, Kim J, Ohno-Machado L. Smooth Isotonic Regression: a new method to calibrate predictive models. *AMIA Summit on Clinical Research Informatics (CRI'11)*, 2011.
14. Wang X, Li F. Isotonic Smoothing Spline Regression. *Journal of Computational and Graphical Statistics*. 2008;17(1): 21-37.
15. Ramsay JO. Monotone regression splines in action. *Statistical Science* 1988;3: 425-41.
16. Lu M, Zhang Y, Huang J. Estimation of the mean function with panel count data using monotone polynomial splines. *Biometrika* 2007;94(705-718).
17. Wu Y. The partially monotone tensor spline estimation of joint distribution function with bivariate current status data. Department of Mathematics. Iowa City: University of Iowa, 2010.
18. Schumaker L. *Spline Functions: Basic Theory*. New York: Wiley, 1981.
19. De Boor C. *A Practical Guide to Splines*. New York: Springer-Verlag, 2001.
20. Hanley J, McNeil B. The meaning and use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology*. 1982;143: 29-36.
21. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*. 1997;16(9): 965-80.
22. Wang Y, Klijn J, Zhang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* 2005;365(9460): 671-79.
23. Osl M, Dreiseitl S, Kim J, Patel K, Baumgartner C, Ohno-Machado L. Effect of data combination on predictive modeling: a study using gene expression data. *AMIA Annual Symposium Proceedings*, 2010:567-71.
24. Kennedy RL, Burton AM, Fraser HS, McStay LN, Harrison RF. Early diagnosis of acute myocardial infarction using clinical and electrocardiographic data at presentation: derivation and evaluation of logistic regression models. *European Heart Journal*. 1996;17(8): 1181–91.
25. Asuncion A, Newman DJ. UCI machine learning repository. 2007.