Appendix A

Let us look at a generalized linear model (GLM), $g(\eta) = w_0 + \sum_{i=1}^{n} w_i x_i^*$, where $W =< w_0, \ldots, w_n >$ is the weight parameter, $\eta = p(Y^* = 1|X^*)$ indicates the estimated probability of a pattern $X^* =< x_0^*, \ldots, x_n^* >$, and $g(\cdot)$ is a link function such that $p(Y^* = 1|X^*) = g^{-1}(W^T X^*)$.

Hosmer and Lemeshow describe different approaches to calculate the confidence intervals of predictions with and without the covariance matrix of the Logistic Regression coefficients[46]. Here we focus on the scenario when we have the estimated coefficients as well as an estimate of the covariance matrix. Without knowledge of these items, it is currently not possible to estimate individualized confidence interval of test cases, as Hosmer and Lemeshow[46] have previous reported. Given both coefficients and covariance matrix, we can estimate confidence intervals of predictions through a two-step procedure. First, we estimate the variance of $g(\eta)$. Then, we use the Delta method[51] to estimate the true variance of η. The idea is to treat w_i as a random variable while fixing x_i^* as a constant factor. The variance of a linked prediction $g(\eta)$ can be estimated by:

$$var(g(\eta)) = var(\sum_{i=0}^{n} w_i x_i^*) = (\Sigma^{1/2})'(X^*)^2(\Sigma^{1/2})), \quad (1)$$

where Σ corresponds to the covariance matrix of the parameters, W, and $\Sigma = (\Sigma^{1/2})'(\Sigma^{1/2})$. Because $g(\eta) \sim N(\cdot, (\Sigma^{1/2})' X^2 (\Sigma^{1/2}))$ and $\eta = g^{-1}(W^T X^*) = h(W^T W^*)$, we can derive $\eta \sim N(\cdot, var(W^T X^*)(h'(W^T X^*))^2)$ using the Delta method[51] so that

$$var(\eta) = ((\Sigma^{1/2})'(X^*)^2(\Sigma^{1/2})) * (h'(W^T X^*))^2 \quad (2)$$

When the link function $g(\eta) = ln(\frac{\eta}{1-\eta})$ (log-loss function), Equation (4) can be converted into

$$var(\eta) = ((\Sigma^{1/2})'(X^*)^2(\Sigma^{1/2})) * (\eta^2(1-\eta)^2) \quad (3)$$

where $h'(W^T X^*) = \frac{e^{W^T X^*}}{(1+e^{W^T X^*})^2}$. Therefore, the 95% CI of $\eta = p(Y^* = 1|X^*)$ is $\eta \pm 1.96 * \sqrt{var(\eta)}$. Since $\eta^2(1-\eta)^2$ is maximized when the predicted score is $\eta = 0.5$, the confidence intervals are wide since they are close to the decision boundary. On the other hand, Σ, the covariance matrix of weight parameters will have large values for feature dimensions that are sparsely populated (i.e., few training samples), and small values for feature dimensions that are densely populated. Therefore, if a pattern X^* is observed in a sparse region of the feature space, its associated elements in Σ will be large, and therefore the term $(\Sigma^{1/2})'(X^*)^2(\Sigma^{1/2})$ will be large. Similarly, this implies $(\Sigma^{1/2})'(X^*)^2(\Sigma^{1/2})$ is small if X^* is observed in a dense region.

Appendix B

To check the method's applicability to select models learned with different features, we did additional experiments by applying APAPT to a variation of myocardial infarction data in Section 3.2.2. Specifically, we used all 48 features to construct a local model for the Edinburgh training cohort, and we used 9 features (i.e., ST elevation, New Q waves, Hypoperfusion, ST depression, Vomiting, LVF, T wave inversion, Pain in right arm, and Nausea) suggested by Kennedy[47] to construct another local model for the Sheffield training cohort. We used the same study design and compared other strategies to ADAPT as described in Table 4 in terms of AUCs and *p*-values (HL decile-based test). The results, as indicated in Figure B1, show that for Sheffield test data, ADAPT has lower AUCs compared to S2S (p=0.03), higher AUCs compared to RANDOM (p<0.01), and comparable AUCs compared to E2S (p=0.43). For the Edinburgh test data, ADAPT has lower AUCs compared to E2E (p<0.01), higher AUCs compared to RANDOM (p<0.01), and higher AUCs compared to S2E (p<0.01). Regarding calibration, when evaluated on Sheffield test data, the difference between ADAPT and E2S (p=0.35) and the difference between ADAPT and RANDOM (p=0.98) are not significant. When evaluated on Edinburgh test data, ADAPT shows better calibration than S2E (p=0.03) and RANDOM (p=0.03). For both experiments, ADAPT is inferior only to the best performing strategies (i.e., E2E and S2S), which are based on patient cohort memberships.

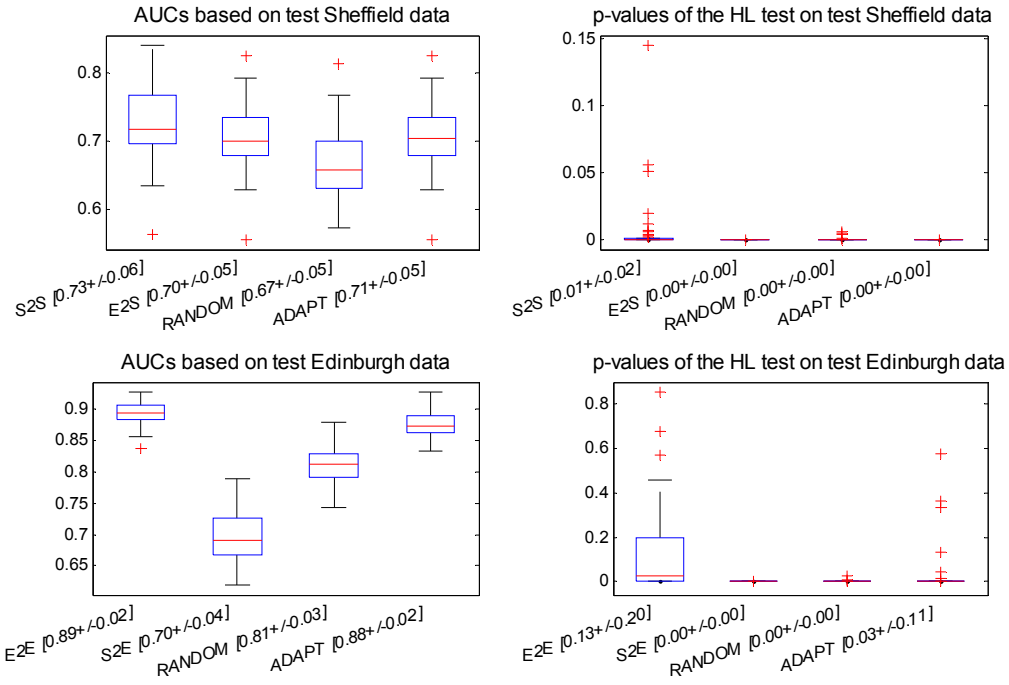

Figure B1: Comparison of effectiveness of different strategies (i.e., BEST, CROSS, RANDOM, and ADAPT) in discrimination (Areas Under the ROC Curve - AUC) and calibration (*p*-value for Hosmer-Lemeshow decile-based test) for selecting "appropriate" models, which are trained with different features using myocardial infarction data. Note that x+/-y in the labels of x-axis indicates that the mean equals x, and the standard deviation equals y.