Appendix

Description of statistical technique

Let Y_j, $j = 1, \ldots, J$, and X_i, $i = 1, \ldots, I$, be the jth endpoint and ith variable, respectively. Let $k = 1, \ldots, N$ be patient indexes. The observed data for the kth patient is represented as $(Y_{jk}, X_{1k}, \ldots, X_{22k})$. In our case, $J = 5$, as there are a total of 5 endpoints, $I = 22$, corresponding to all 22 variables and $N = 269$, the total number of patients in the cohort. The total number of patients may be smaller for different endpoints, when an entry for some patients of the endpoint is missing. Our goal is to find variables for all five endpoints, simultaneously, and to predict whether a patient should be classified as a case or control.

As there are five endpoints, we must fit five regression models to assess the impact of each variable on each endpoint. The endpoints Y_j, $j = 1, \ldots, 5$, are all binary, i.e., patients are either classified as cases or controls. The standard approach is to fit a logistic regression model. It has the form:

$$\log \frac{\Pr(Y_{jk} = 1 | X_{1jk}, \ldots, X_{Ijk})}{\Pr(Y_{jk} = 0 | X_{1jk}, \ldots, X_{Ijk})} = \theta_{0j} + \sum_{i=1}^{I} \theta_{ij} X_{ijk}, \tag{1}$$

where θ_{0j} is the intercept of the jth endpoint and θ_{ij} is the effect size of the ith variable of the jth endpoint. The term $\Pr(Y_{jk} = 1 | X_{ijk}, \ldots, X_{Ijk})$ is the probability that the kth patient, with respect to the jth endpoint and given patient characteristics, is a case. However, instead of working with effect sizes θ_{ij}, it is often better to work with a standardized version, which is obtained by simply dividing the effect sizes by their respective standard deviations (since the standard deviation is of a parameter, it is referred as a standard error, se), i.e.,

$$\beta_{ij} = \theta_{ij} / \mathrm{se}(\theta_{ij}).$$

With standardization, all coefficients are in the same magnitude and comparisons are more meaningful.

Thus, if the expression in equation (1) is greater than zero, $\Pr(Y_{jk} = 1 | X_{1jk}, \ldots, X_{Ijk})$ is greater than $\Pr(Y_{jk} = 0 | X_{1jk}, \ldots, X_{Ijk})$. In other words, when the model in equation (1) is used as a classifier, $\theta_{0j} + \sum_{i=1}^{I} \theta_{ij} X_{ijk} > 0$, or in terms of the standardized coefficients, $\beta_{0j} + \sum_{i=1}^{I} \beta_{ij} X_{ijk} > 0$ (known as the Fisher discriminant function), suggests that the kth patient should be classified as a case, otherwise it should be classified as a control.

The parameters of interest θ_{ij} (and thus β_{ij}), for $i = 1, \ldots, 22$ and $j = 1, \ldots 5$, are often estimated by maximizing the loglikelihood associated to equation (1). This is the standard procedure and inference is made based on the maximum likelihood estimates (MLEs), henceforth denoted as $\hat{\boldsymbol{\theta}}^{\mathrm{MLE}}$ (and their standardized versions as $\hat{\boldsymbol{\beta}}^{\mathrm{MLE}}$).

Individual MLEs of a given variable can be further improved by using a combination of other MLEs obtained for the same variable, but from different endpoints. That is, e.g., the estimate $\hat{\beta}_{2,3}^{\mathrm{MLE}}$, which correspond to the effect size of the second variable (X_2) for predicting the third endpoint (Y_3), can be improved by taking the other estimates $\hat{\beta}_{2,1}^{\mathrm{MLE}}, \hat{\beta}_{2,3}^{\mathrm{MLE}}, \hat{\beta}_{2,4}^{\mathrm{MLE}}$ and $\hat{\beta}_{2,5}^{\mathrm{MLE}}$ into account. This seems controversial at first and it is known as Stein's paradox [1]. James and Stein, [2] showed that, when there are at least three parameters that are simultaneously estimated, a better approach, in the sense that it has lower mean squared error (MSE), may be constructed. This estimator will be called the James-Stein estimator (JSE). It was later shown that the JSE may also lead to better predictions than those using MLEs, [3].

As the effect size of each variable was estimated five times (corresponding to the five endpoints),

the James-Stein estimator $\hat{\beta}_{ij}^{\text{JSE}}$ is directly applicable and may be constructed from the MLE as

$$\hat{\beta}_{ij}^{\text{JSE}} = \lambda_i \hat{\beta}_{ij}^{\text{MLE}}, \text{ with } \lambda_i = \left(1 - \frac{J-2}{S_i}\right) \text{ and } S_i = \sum_{j}^{J} (\hat{\beta}_{ij}^{\text{MLE}})^2. \tag{2}$$

Where J corresponds to the number of endpoints (5, in our case). The parameter λ_i is the shrinkage parameter corresponding to the ith variable, across all five endpoints. It is believed that, the average effect size is zero, since it is likely that most variables have no effect on the endpoint. Notice that the JSE estimator shrinks the MLEs towards zero (by multiplying the MLE with λ_i), the average effect size, when the sum of squares S_i is approximately equal to $J-2$. When S_i is large ($S_i \gg J-2$), there is almost no shrinkage of the MLEs towards zero suggesting that the MLEs cannot benefit from their neighbours. Finally, when $S_i < J-2$ the shrinkage parameter λ_i becomes negative, in that case $\lambda^+ = \max(\lambda_i, 0)$ is used instead of λ_i. Using λ_i^+ when λ_i is negative, sets effect sizes exactly to zero.

The correlations between the endpoints were investigated and no significant correlations were found. Similarly, no significant correlation was seen among the variables. As a consequence, the sum of squares S_i can be used as a test statistic, which has a chi-squared distribution with J degrees of freedom. With $J = 5$ and at a 5% significance level, the critical value is 11.07. Sum of squares greater than 11.07 are significantly different from zero at a 5% significance level and their corresponding p-values can be calculated.

References

[1] B. Efron and C. N. Morris. *Stein's paradox in statistics.* WH Freeman, 1977.

[2] W. James and C. Stein. Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 361–379, 1961.

[3] R. I. Jennrich and S. D. Oman. How much does stein estimation help in multiple linear regression? *Technometrics*, 28(2):113–121, 1986.