Linear Regression on the BIOSCAD data set

Introduction

Overview:
The aim of the project was to create a model that, based off the biomarkers in the skin, would predict the SCORAD of a patient. In addition, biomarkers which are shown to have a large effect on SCORAD should be identified. This was done based off the data from the AMC BIOSCAD study.

Method

Data sets:
The model's data is based off the data from the AMC BIOSCAD study. It is first pre-processed to make it suitable for training – see pre-processing report (../Preprocessing/preprocessing-report.pdf) for details. The pre-processed data (see pre-processing document) is split in to three variants. All three variants are based on the combined non-lesional data for both AD and control patients. This gave a total of 100 data points. Data from lesional skin was not used as, after pre-processing, there were only 17 data points remaining.

- In the first variant, the continuous variables are logged before normalizing.
- In the second, the continuous variables are not logged before normalizing.
- The third variant is based off the first (logged continuous data) but only two of the continuous data attributes are used: IL-1a and IL1β. This is because, in the original data set, the remaining attributes have significant numbers of values marked as being below the detection range. The categorical attributes are still used.

Models:
For data sets 1 and 2, elastic net regularization with varying values of both alpha and lambda was used. The data sets were split up to use 60% for training, 20% for cross-validation, and 20% for testing. The models were also trained on 100 different splits in the data to see how the best values of alpha and lambda varied. Using the validation data, we chose the best alpha and lambda values. We then combined these using a weighted mean based on the performance against the testing data to find the overall best values.

For data set 3, generalized linear regression was used. The data set was split in to 80% training and 20% testing. Once again, 100 different splits of the data set were used to see how model performance varied.

Each model outlined above was created twice, once using objective SCORAD as the dependent variable, and once using total SCORAD.

Performance evaluation:
To calculate performance of a given model the root mean square error (RMSE) between the predicted result and the actual result was calculated. We also calculated the accuracy, defined as the frequency of 'successful predictions'. A successful prediction happened when the absolute prediction error was less than 9 points for objective SCORAD and 10 points for total SCORAD. These values were based on the minimum clinically important difference (1).

In order to see how the performance could vary depending on how the testing and training data was split, we recorded the performance of each cross-validation iteration and plotted these in a box plot.

Feature selection:
To see which biomarkers are important, we plotted the coefficients of the final, optimised, model using a bar graph. As all of the attributes are normalised before training, these coefficients can be used to directly compare the importance of each attribute to the output SCORAD.

Residual analysis:
We also created a plot of the residuals after training and testing on the whole data set to allow us to check the assumptions of our statistical models.

Average predictor:
We also created a model that simply predicted the average value of either total or objective SCORAD. This was used as a null model. To compare the models, we computed both the kappa coefficient and the p-value for a one sided Wilcoxon rank-sum test where the alternative hypothesis was that the model accuracy was higher than the average predictor accuracy.

Results

Alpha and Lambda parameter values:

Alpha values close to 1 favour lasso regression, values close to 0 favour ridge. The best alpha values for each model are shown below:

	Logged (Elastic net)	Unlogged (Elastic net)
Total SCORAD	0.376	0.624
Objective SCORAD	0.393	0.606

Table 1

The best lambda values for each model are shown below:

	Logged (Elastic net)	Unlogged (Elastic net)
Total SCORAD	1.62	2.34
Objective SCORAD	1.73	2.12

Table 2

Model performance:

The RMSE for each model is shown below:

	Logged (Elastic net)	Unlogged (Elastic net)	Reduced (General linear regression)
Total SCORAD	13.65	27.60	14.74
Objective SCORAD	10.78	15.75	10.75

Table 3

The success rate of each model is shown below:

	Logged (Elastic net)	Unlogged (Elastic net)	Reduced (General linear regression)
Total SCORAD	52.7%	49.0%	48.6%
Objective SCORAD	68.8%	64.9%	60.8%

Table 4

Average predictors:

The average prediction models (null models) had the following performance:

	RMSE	Successful predictions
Total SCORAD	15.06	35.8%
Objective SCORAD	10.76	58.8%

Table 5

In order to compare the predictions to the null model, we computed the kappa coefficient relative to average predictors for null and total SCORAD accordingly. This is shown below:

	Logged (Elastic net)	Unlogged (Elastic net)	Reduced (General linear regression)
Total SCORAD	0.263	0.206	0.017
Objective SCORAD	0.243	0.148	0.063

Table 6

The p-values for a one sided Wilcoxon rank-sum test where the alternative hypothesis was that the model accuracy was higher than the average predictor accuracy are shown below:

	Logged (Elastic net)	Unlogged (Elastic net)	Reduced (General linear regression)
Total SCORAD	$1.77 * 10^{-21}$	$5.49 * 10^{-7}$	$5.47 * 10^{-4}$
Objective SCORAD	$8.77 * 10^{-2}$	$3.88 * 10^{-2}$	$2.34 * 10^{-2}$

Table 7

The p-values for a one sided t- test where the alternative hypothesis was that the model RMSE was lower than the average predictor RMSE are shown below:

	Logged (Elastic net)	Unlogged (Elastic net)	Reduced (General linear regression)
Total SCORAD	$1.91 * 10^{-20}$	1	$3.85 * 10^{-2}$
Objective SCORAD	$2.23 * 10^{-7}$	1	$4.44 * 10^{-3}$

Table 8

oSCORAD logged data results:

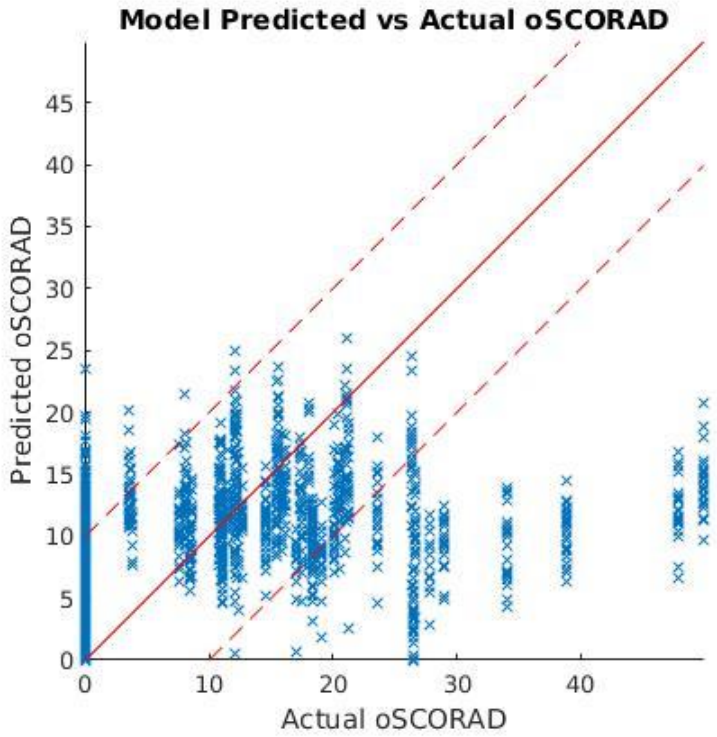

Figure 1

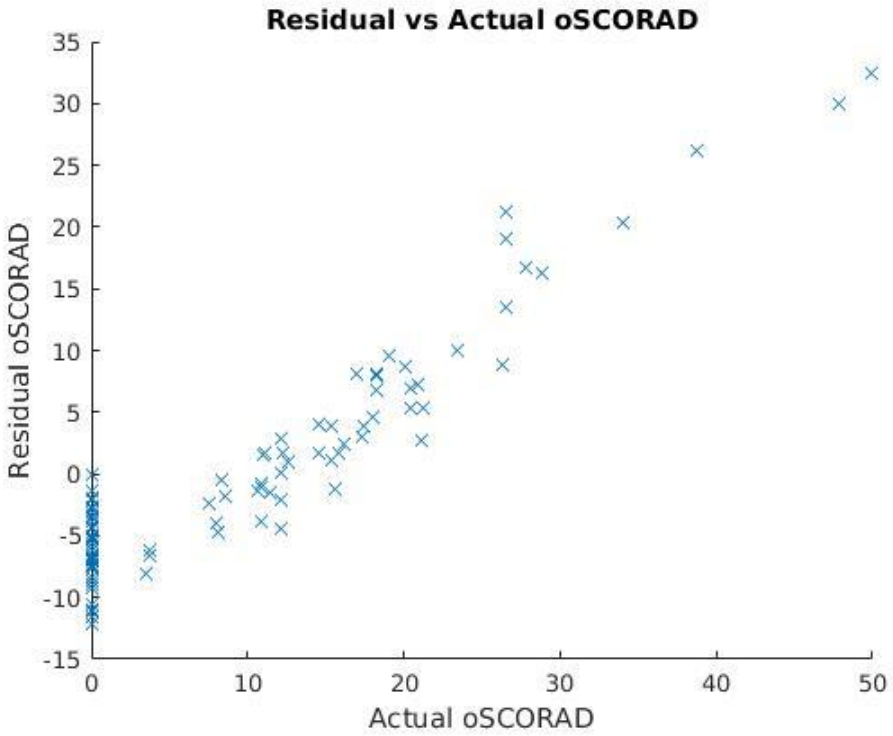

Figure 2

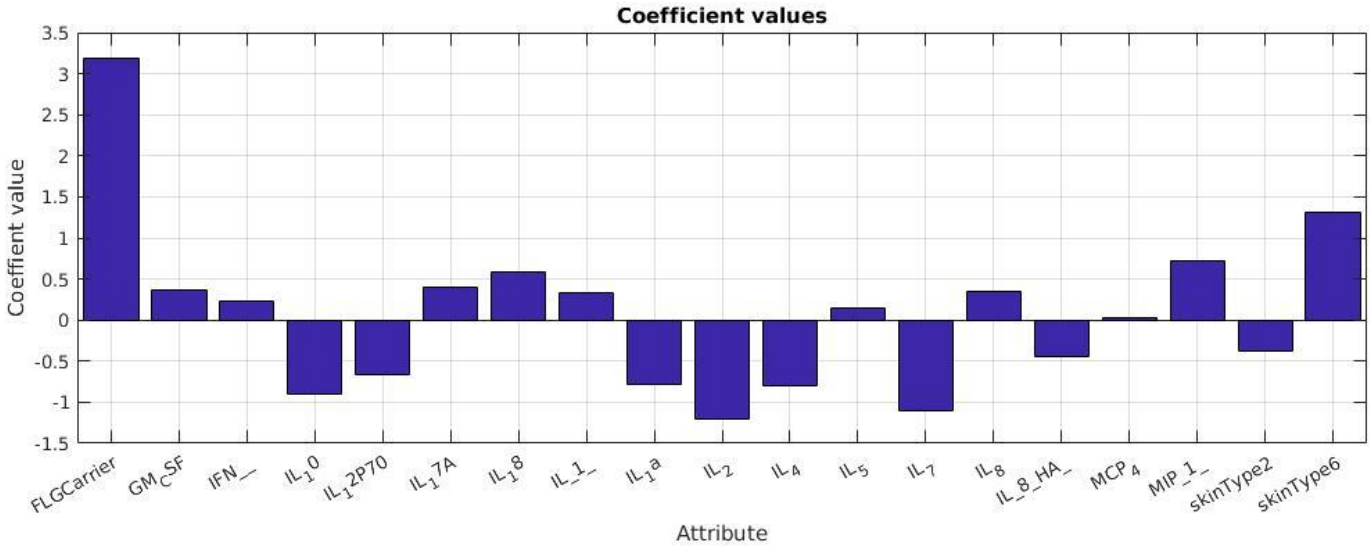

Figure 3

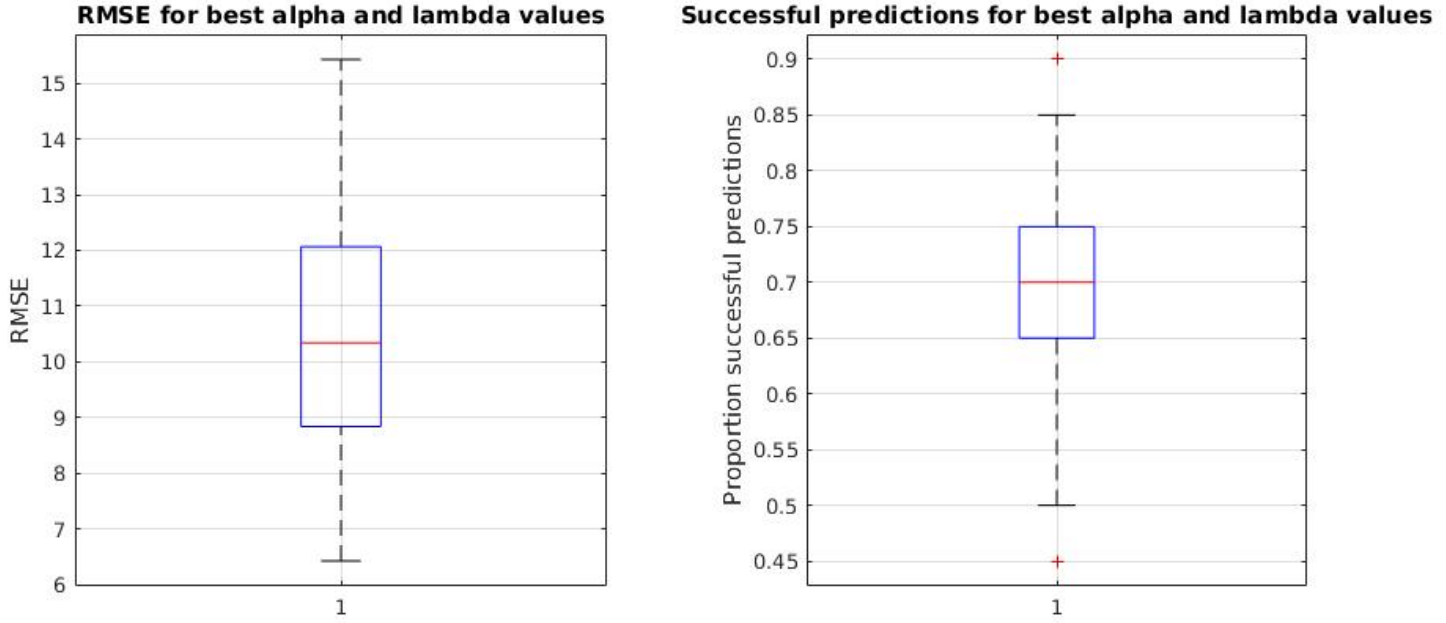

Figure 4

totSCORAD logged data results:

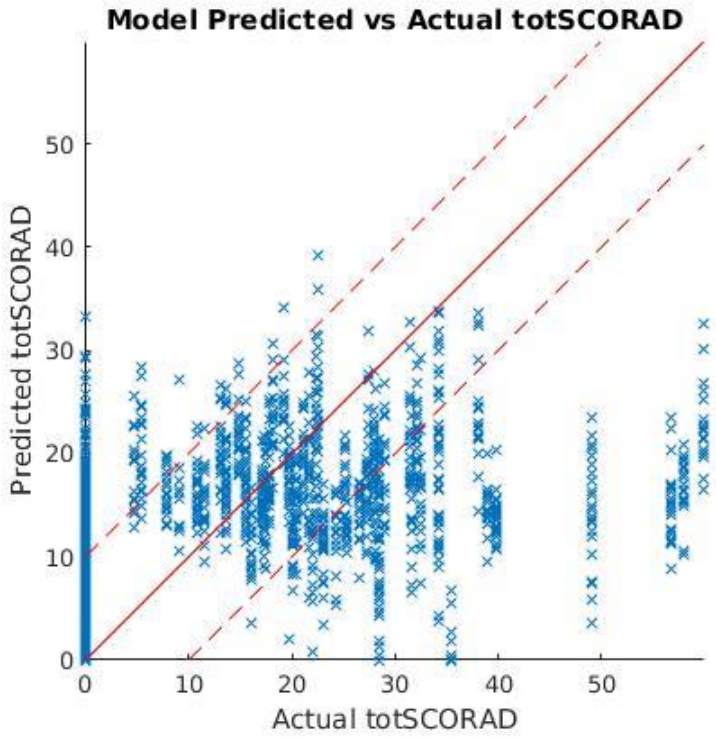

Figure 5

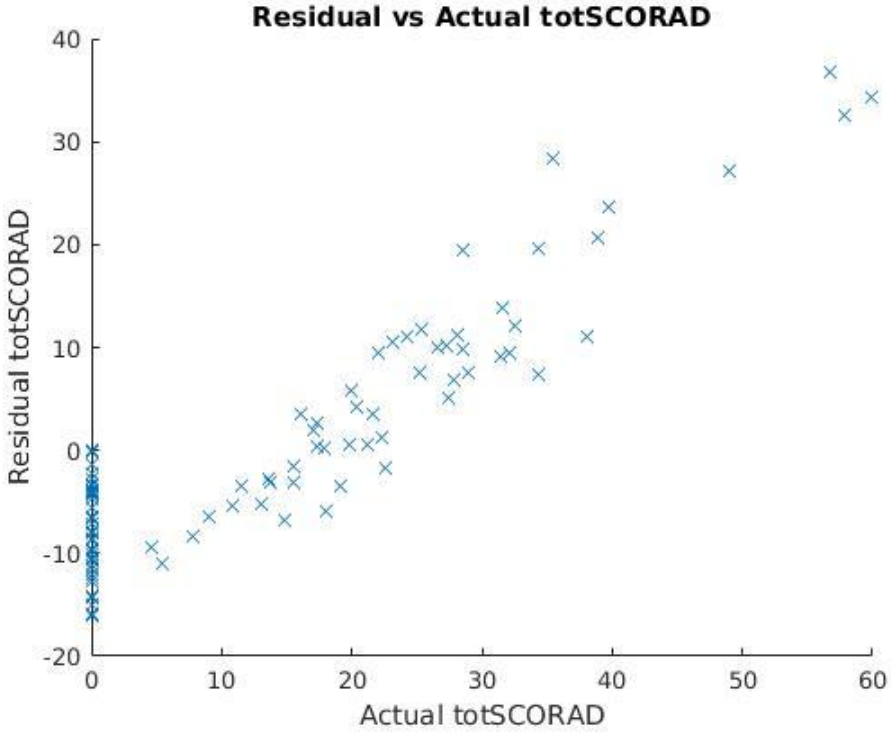

Figure 6

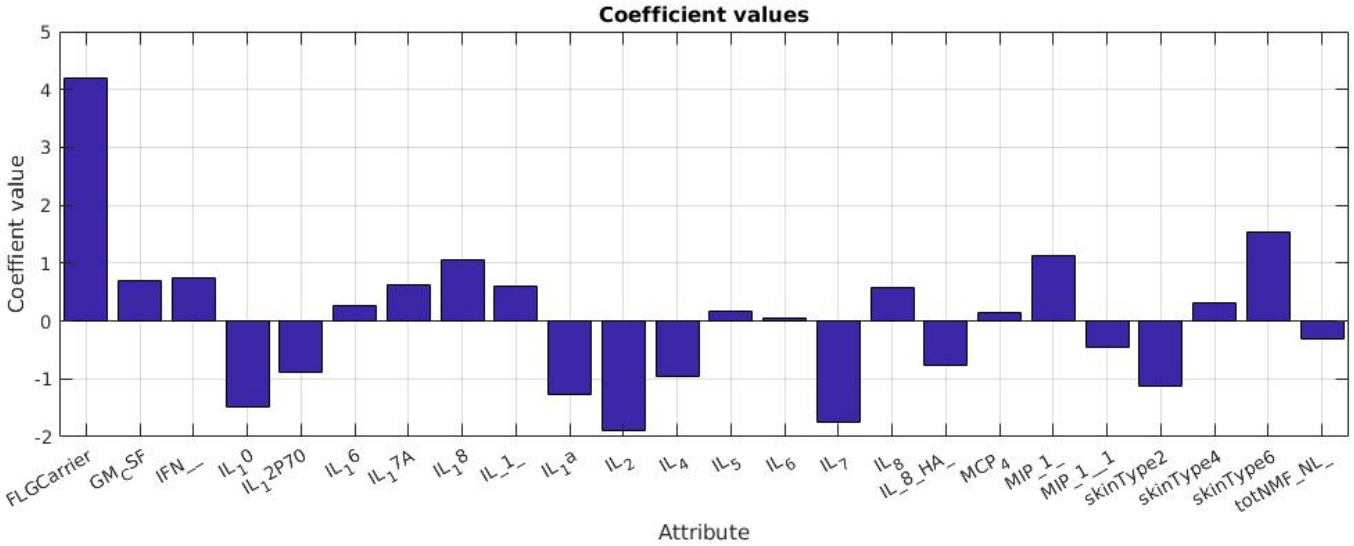

Figure 7

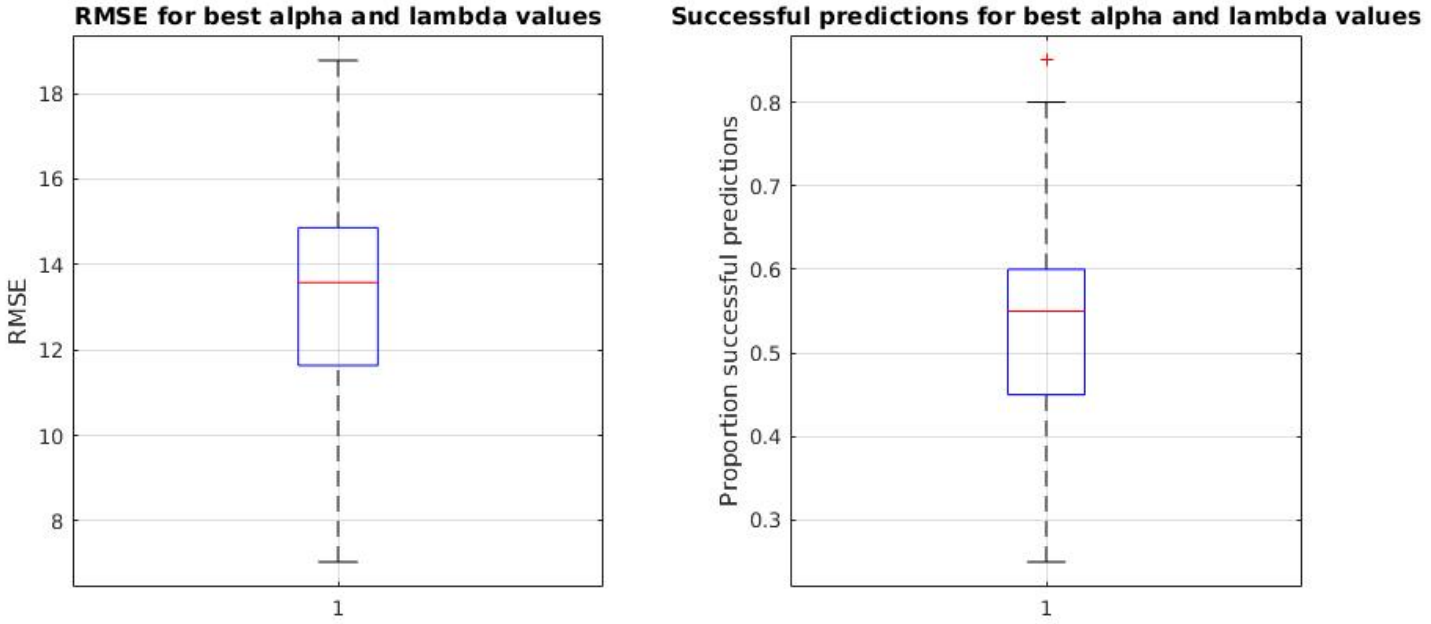

Figure 8

oSCORAD unlogged data results:

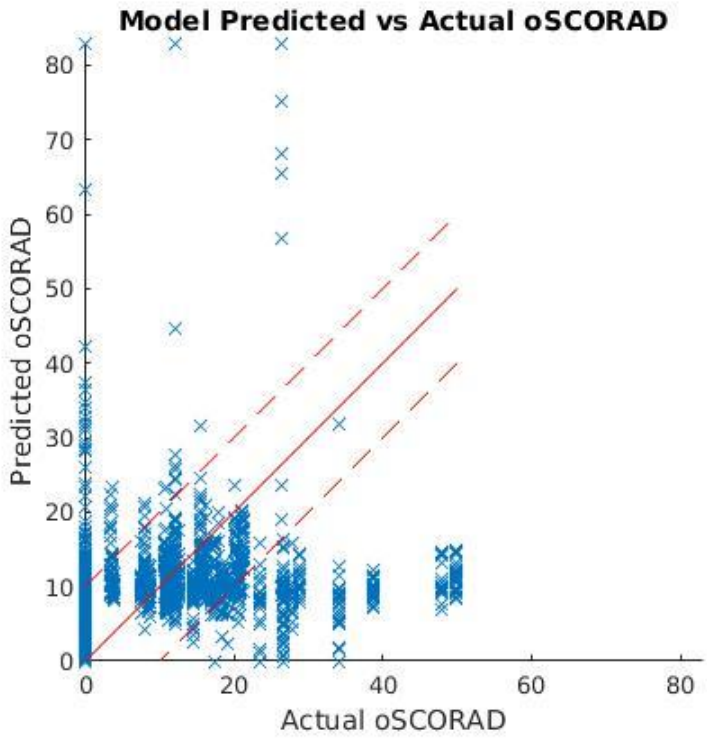

Figure 9

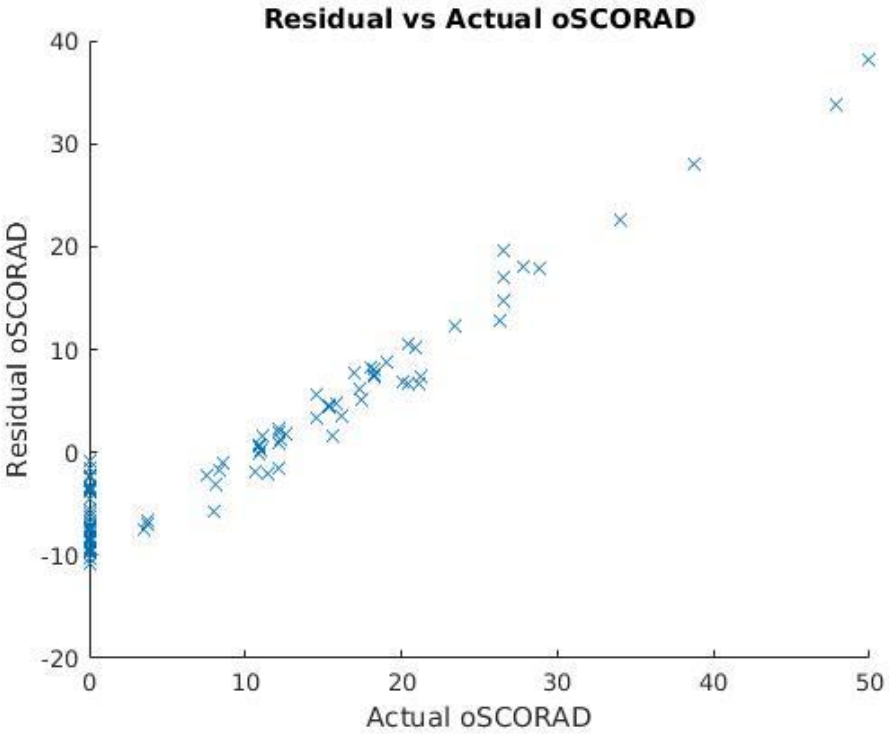

Figure 10

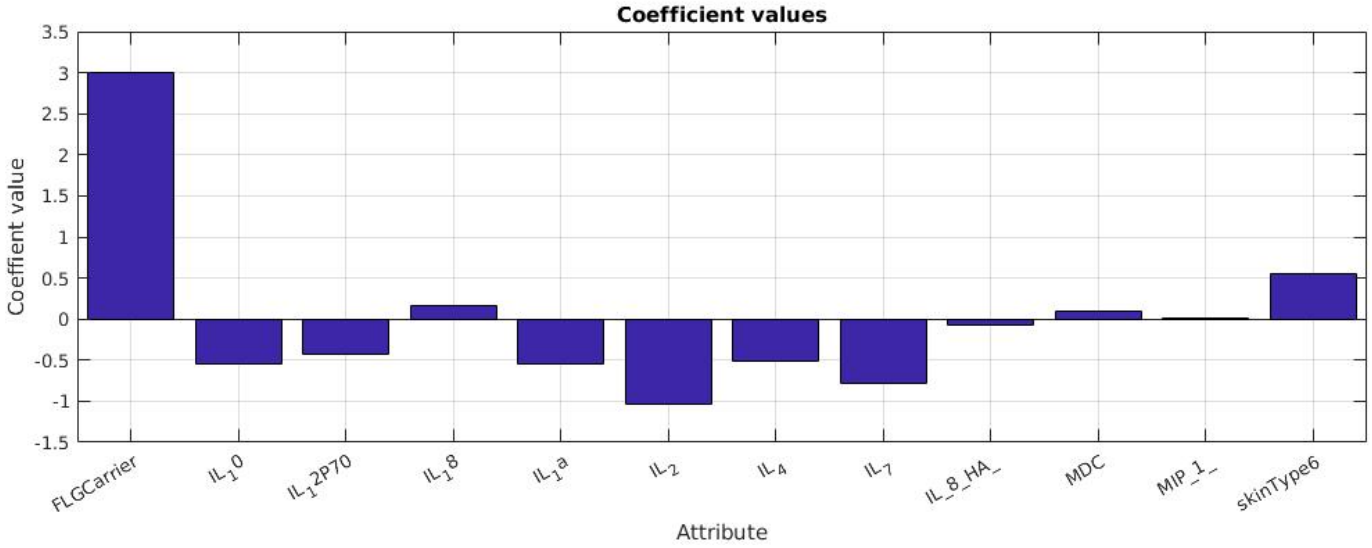

Figure 11

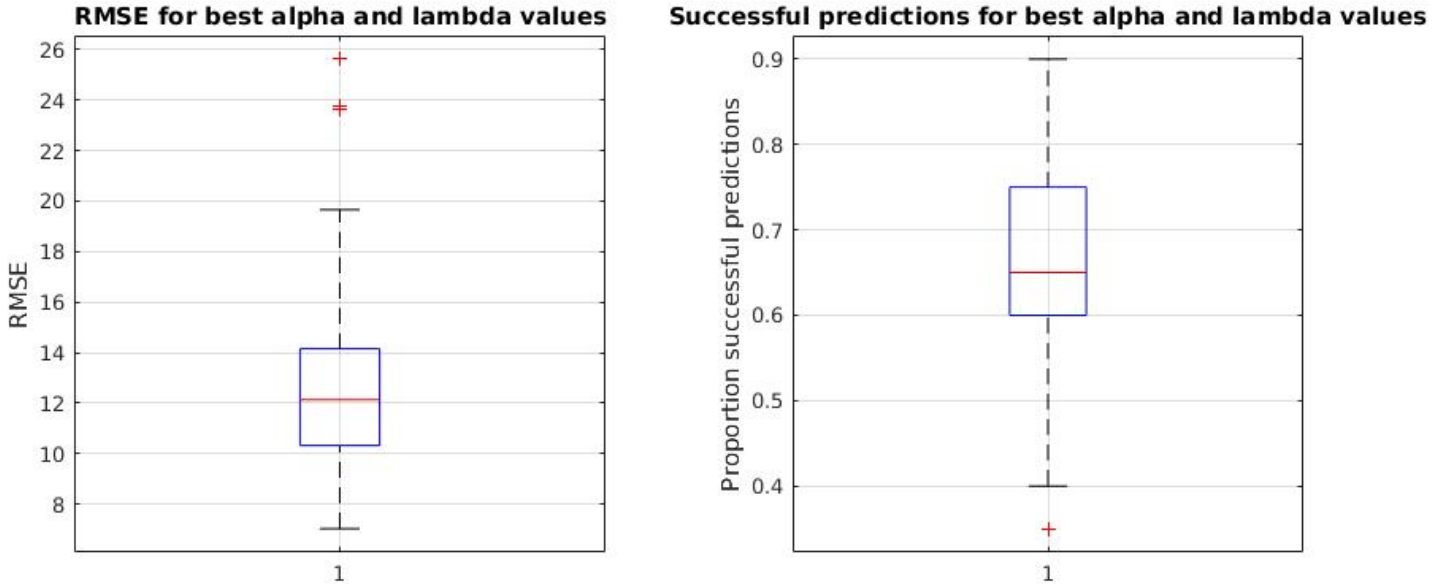

Figure 12

totSCORAD unlogged data results:

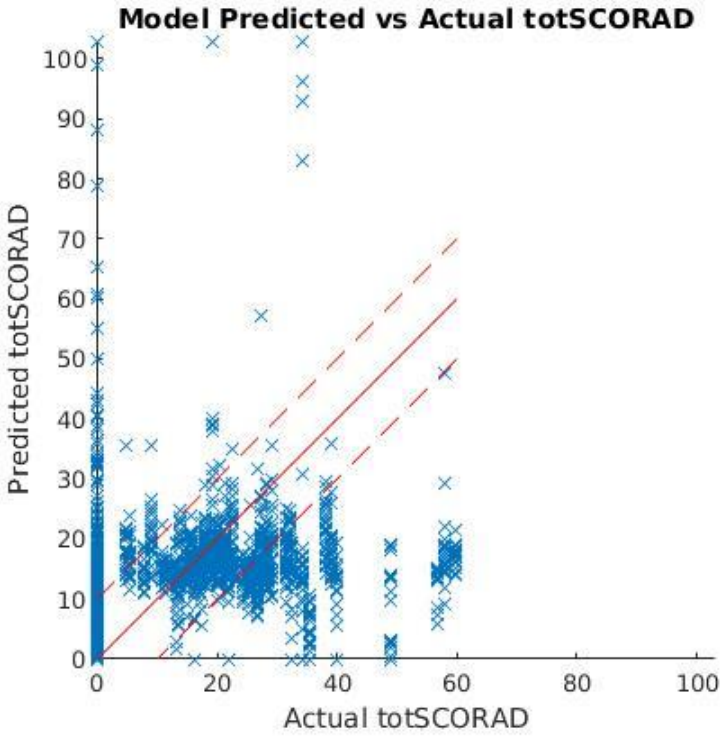

Figure 13

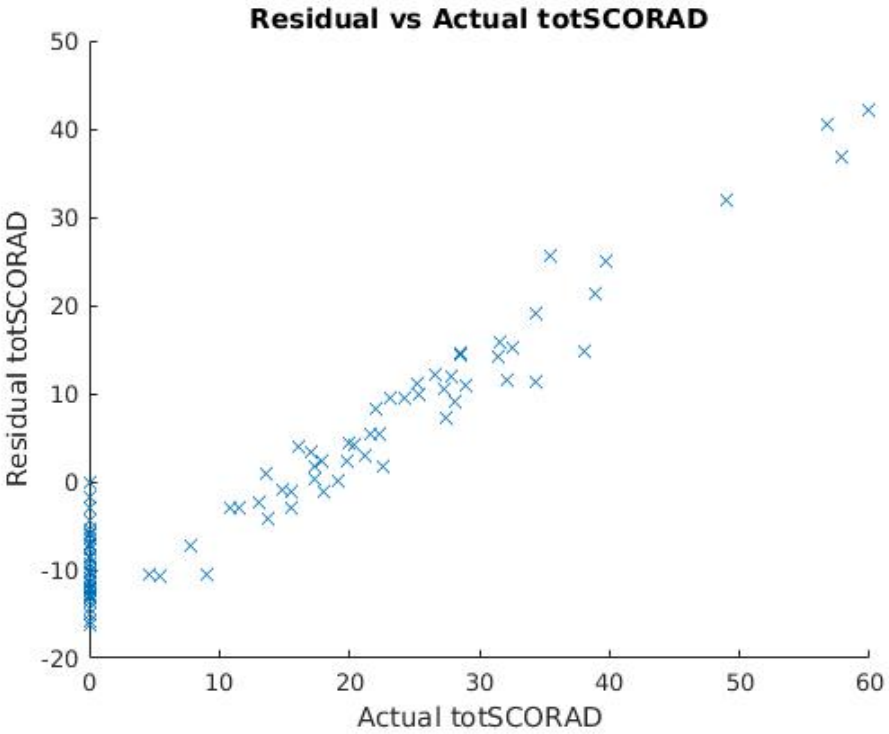

Figure 14

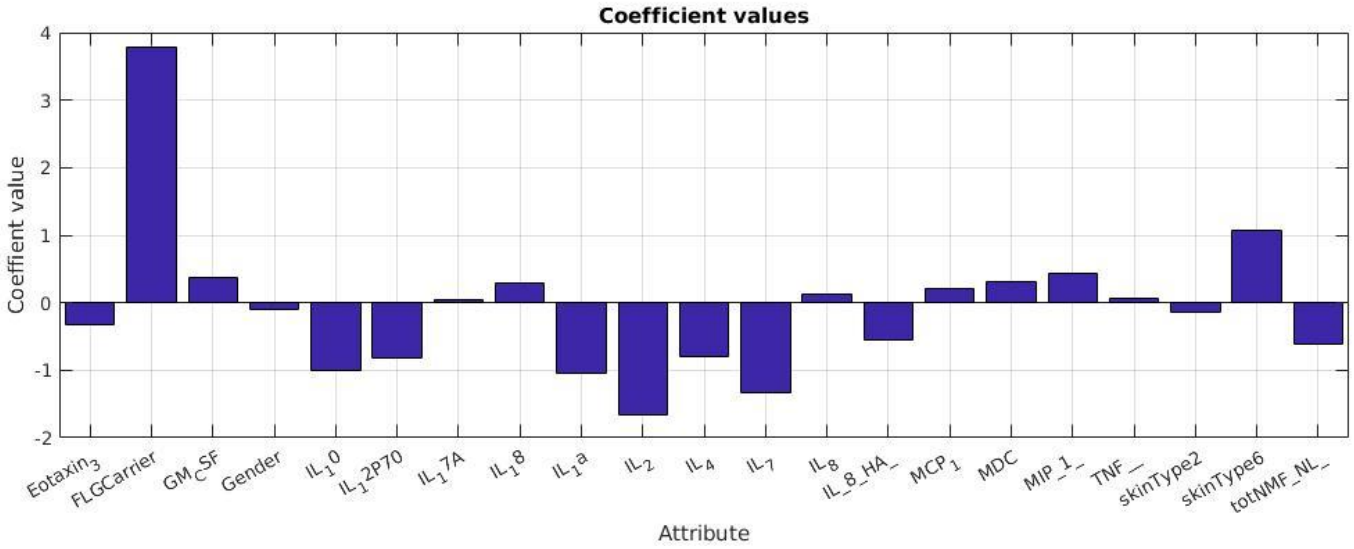

Figure 15

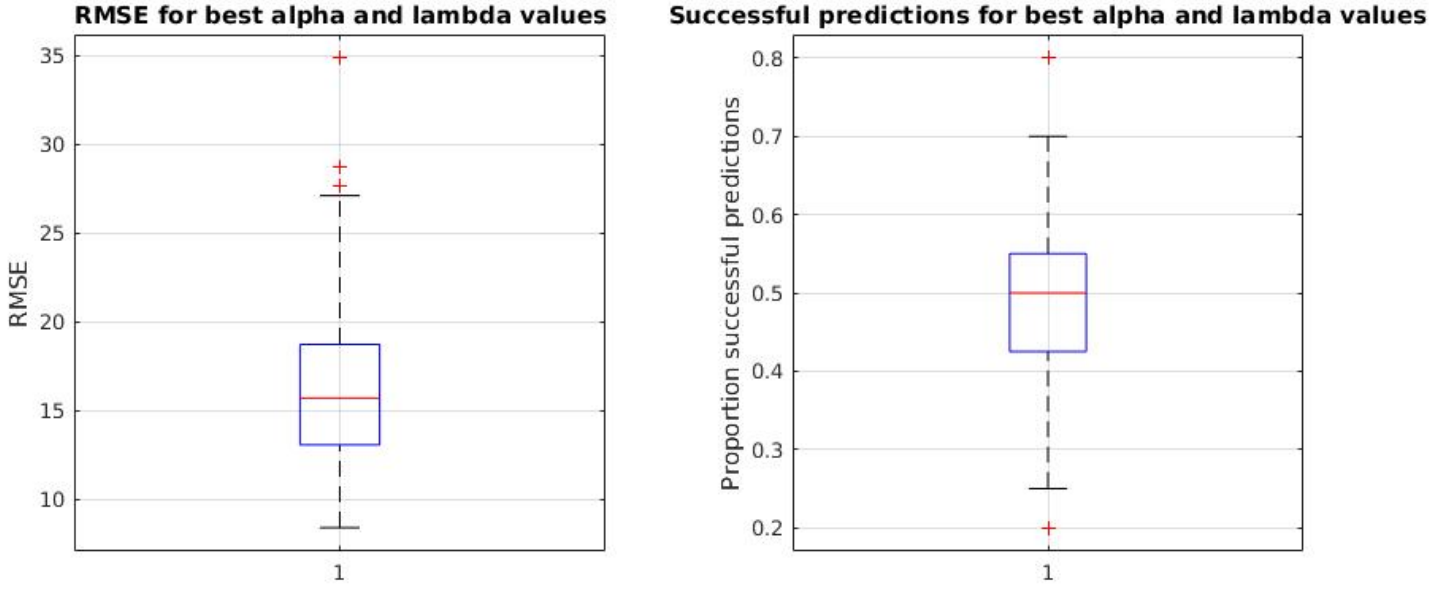

Figure 16

oSCORAD reduced subset data results:

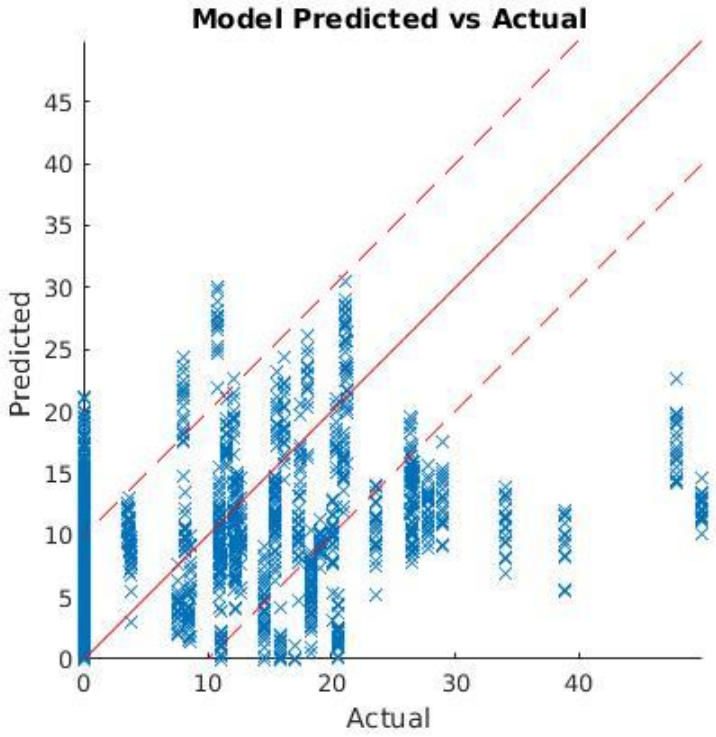

Figure 17

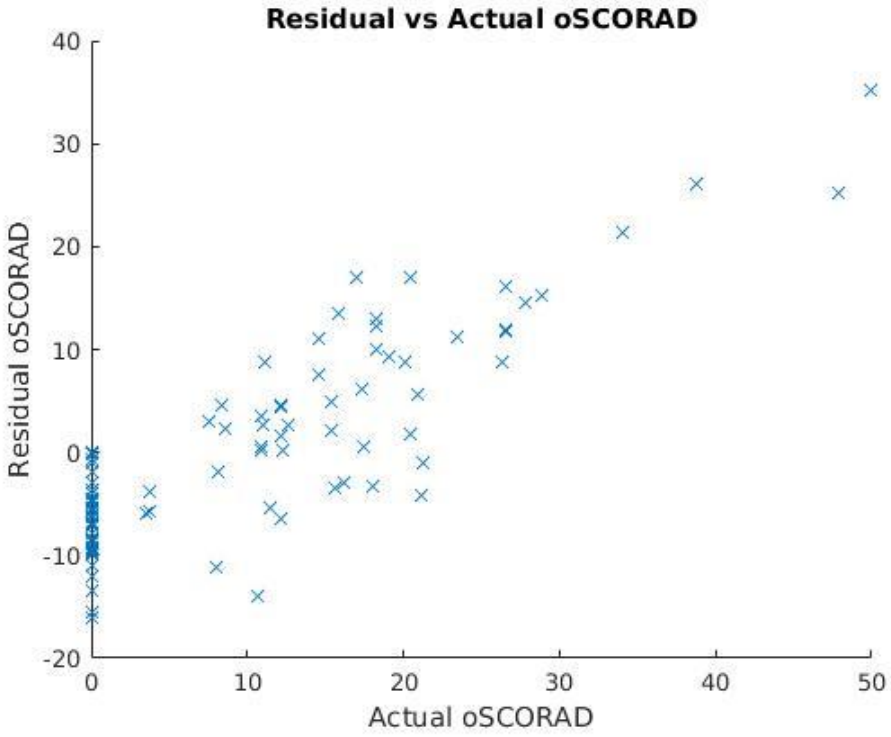

Figure 18

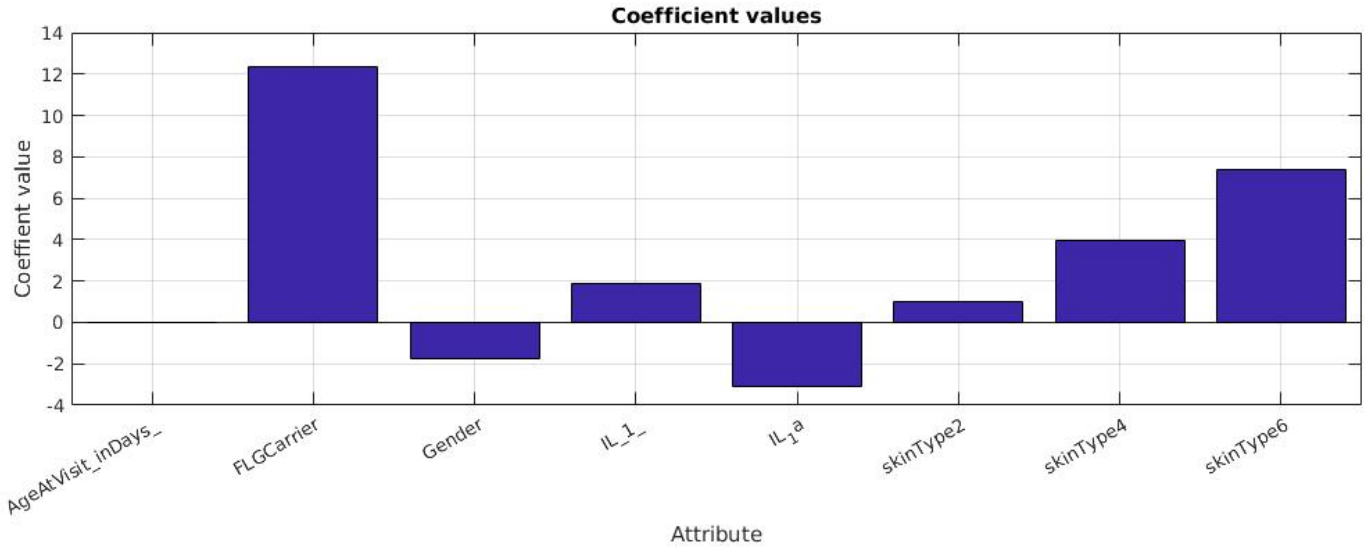

Figure 19

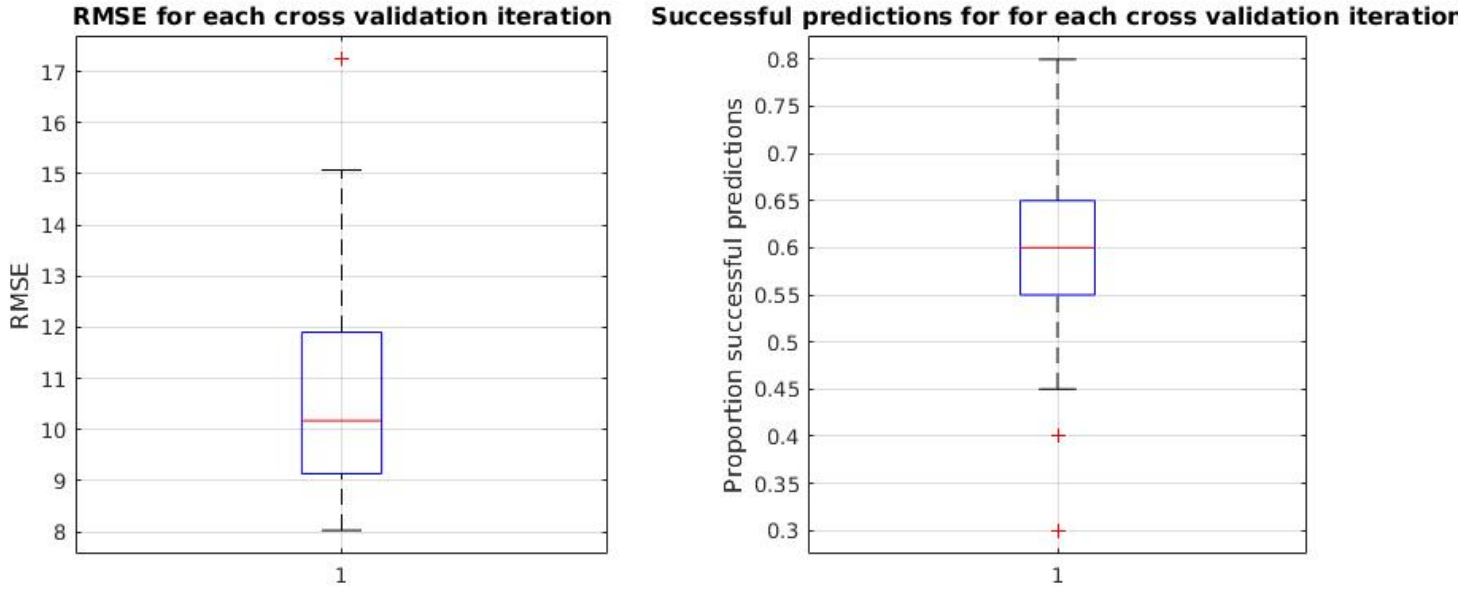

Figure 20

totSCORAD reduced subset data results:

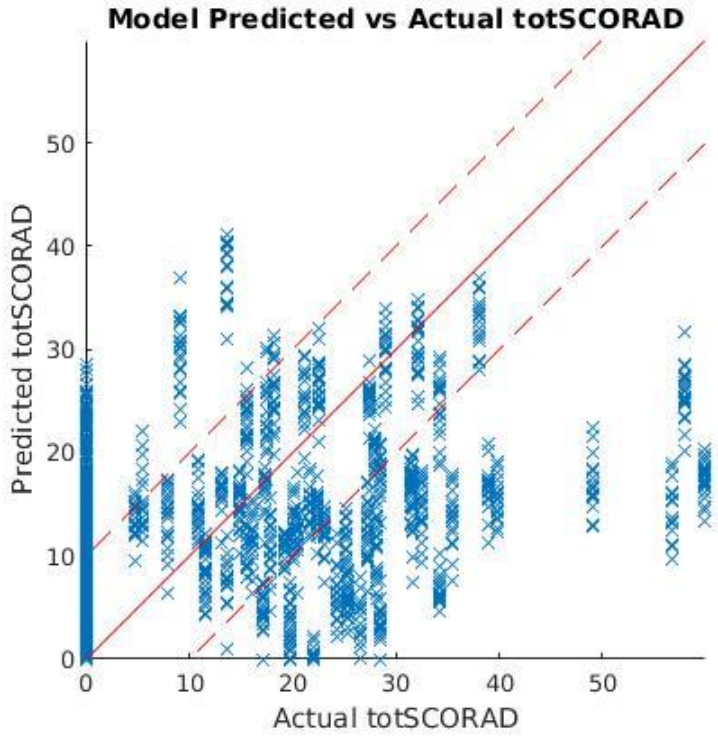

Figure 22

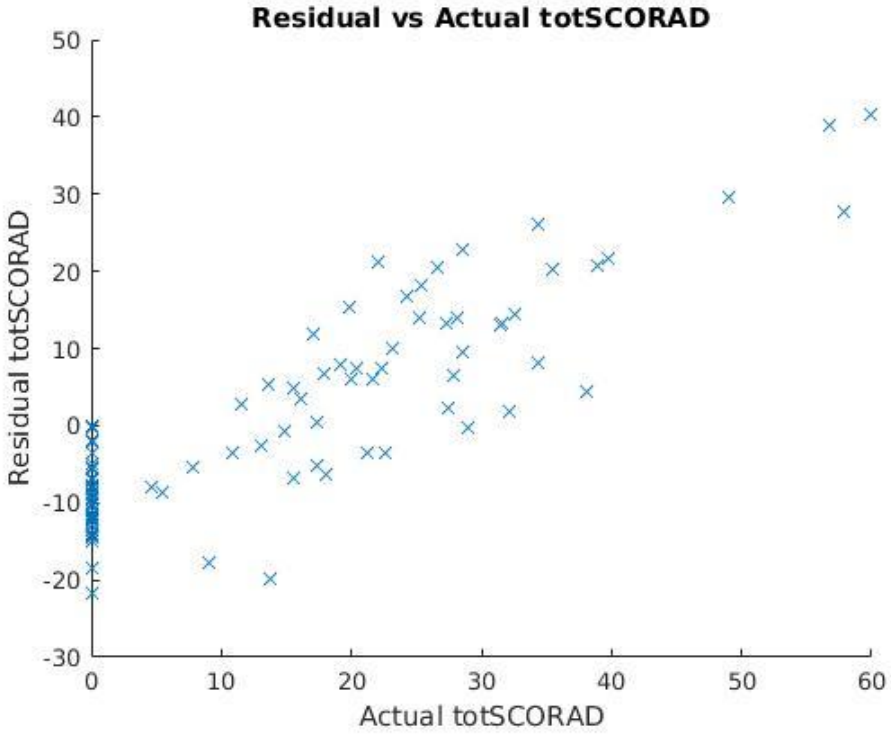

Figure 22

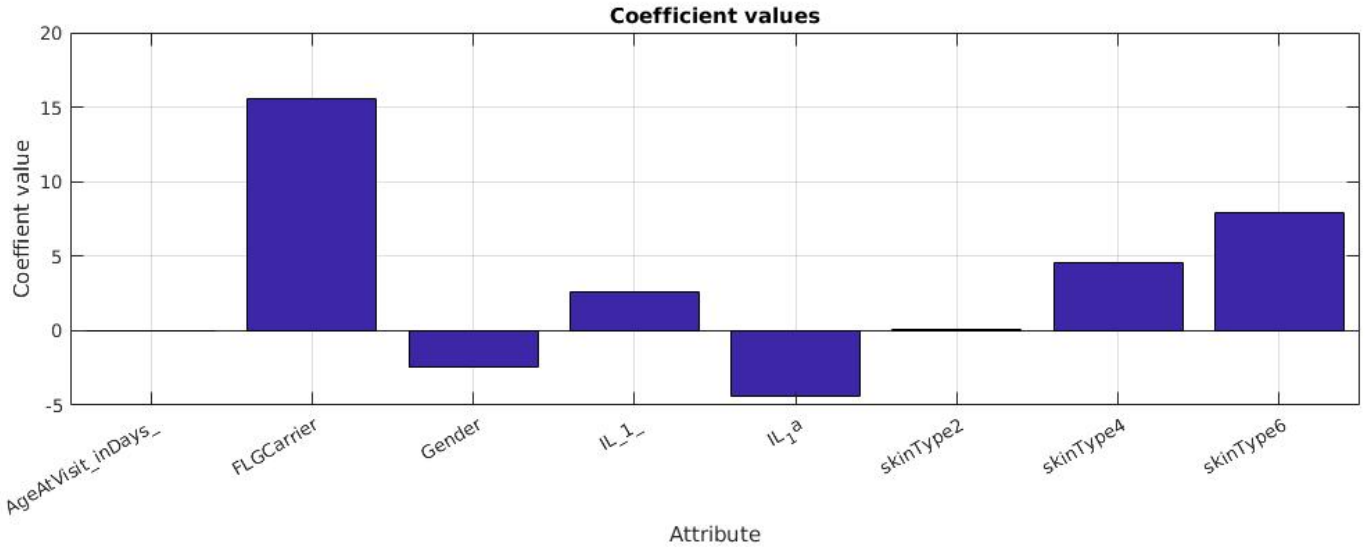

Figure 33

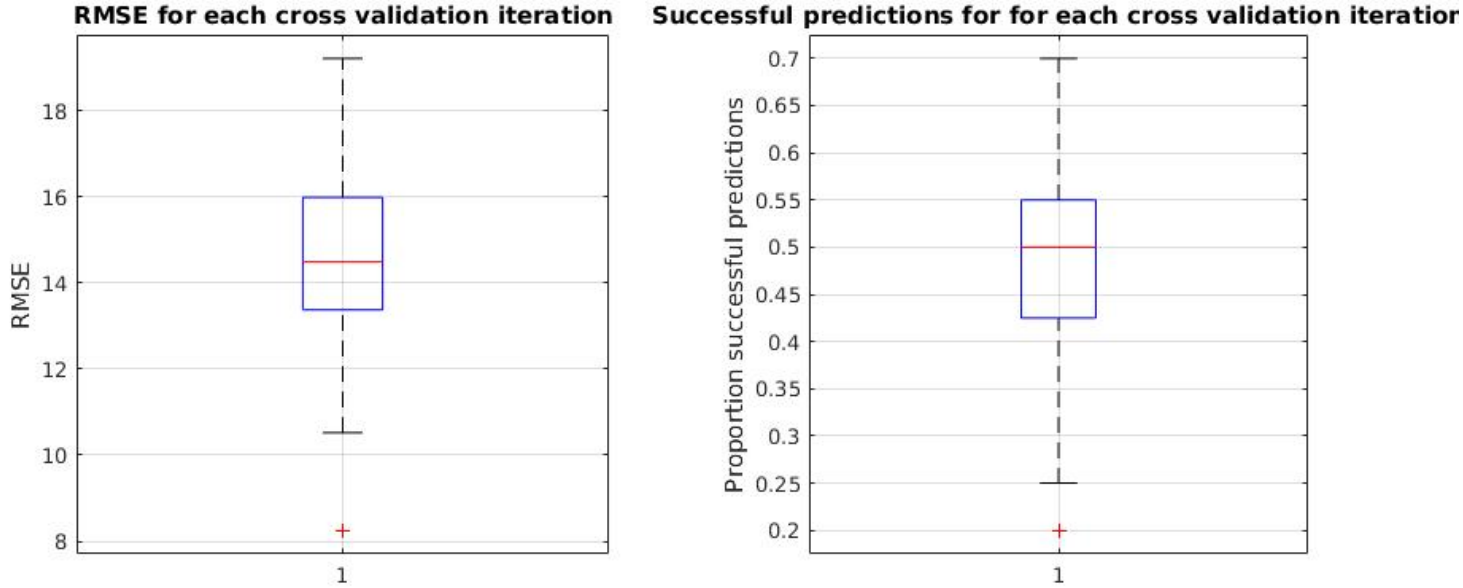

Figure 24

Discussion

Performance evaluation:
As shown above, the models trained on logged data performed better than those based on unlogged data. It can be because logging the data reduces the effect that large value outliers have. This is particularly important as the training data is small, thereby making it harder to spot outliers. Evidence of this can be seen in the very large predictions made in the unlogged models (figures 9 & 13) – likely a result of outliers in the training data.

In addition, models using objective SCORAD as the dependent variable performed better than those using total SCORAD. This is because it could be difficult for the subjective component in total SCORAD to accurately represent the severity of the disease as it depends on how the person feels. As such, the data that the model is trained on and compared to is less reliable.

Coefficients:
From the coefficient graphs, we see that the most important input was found to be the presence of an FLG mutation (FLG carrier input) which, when present, increased the SCORAD predictions by approximately 3.2. As this is a small change, it indicates that the inputs had little effect on the output. Thus, there is a small range of output values. Other biomarkers which had the most significant positive coefficients are skinType6 (1.3), MIP_1 (0.7), and IL_18 (0.59). Biomarkers which had the most significant negative coefficients are IL_2 (-1.2), IL_7 (-1.1), and IL_10 (-0.9).

Residuals:
The residuals for each model shows a linear increase in residual value with SCORAD. This indicates that the input data alone cannot accurately predict SCORAD and that there are missing attributes that have a large impact on SCORAD.

Comparison to an average predictor:
As seen in tables 7 & 8, some of the models perform significantly better than the average prediction model at 5% significance. However, we see that models that perform better with RMSE do not necessarily perform better with percentage successful predictions and vice-versa. This is probably due to small numbers of points which had extreme differences between the predicted and actual values affecting the RMSE more than the percentage successful predictions.

Overall, this means that although the other models may perform better than the null model, as they still have low accuracy values themselves (table 4), they are still not good models. The lack of performance is to be expected as the coefficients are small so the prediction stays within a small range regardless of input.

This is because, as mentioned earlier, in all but two of the continuous attributes the majority of the data was below the detection range and therefore, not accurate. As the model is only

as good as the input data (garbage in garbage out principle), it is not surprising to have poor performance.

In addition, the data set was too small. Of the 100 data points, only 53 were from patients with atopic dermatitis. Once testing and validation sets were extracted, only 60 points remained to train the model with 35 attributes.

Bibliography

Bibliography

1. Bos JD, Schram ME, Spuls IP, Leeflang MM, Lindeboom R, Schmitt J. EASI, (objective) SCORAD and POEM for atopic eczema: responsiveness and minimal clinically important difference. European Journal of Allergy and Clinical Immunology. 2011 September; 67(1).