Supplement 1 for:

Derivation of the Buffalo Concussion Physical Examination Risk of Delayed Recovery (RDR) Score to identify Children at Risk for Persistent Post-Concussive Symptoms

Adam Cunningham, MA, MS; M Nadir Haider, MD, PhD

Other coauthors: John J Leddy, MD; Barry S Willer, PhD; Scott Darling, MD; Heidi Suffoletto, MD; Michael Freitas, MD and Rajiv Jain, MD

Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response Variables</td>
<td>2</td>
</tr>
<tr>
<td>Predictor Variables</td>
<td>4</td>
</tr>
<tr>
<td>Time-since-injury in Predictive Models</td>
<td>12</td>
</tr>
<tr>
<td>Multicollinearity</td>
<td>13</td>
</tr>
<tr>
<td>Results of Predictive Models</td>
<td>15</td>
</tr>
<tr>
<td>Development of the RDR-Score</td>
<td>20</td>
</tr>
<tr>
<td>Cross-validation and Summary</td>
<td>23</td>
</tr>
</tbody>
</table>

Implementation Note

All statistical analyses for this report were performed using the R Programming Language. Graphs were generated using the ggplot2 package, and data analysis performed using the flexsurv, muhaz, survival, and survminer packages. Since we are primarily interested in building a predictive model, we shall use the terms “predictor” and “response” throughout this report to refer to independent and dependent variables respectively. Given that there are 15 predictors, a traditional significance level of 0.05 is inappropriate, so we use a family-wide error rate of 0.05, and a Bonferroni correction to determine the significance level for each individual test. With 15 predictors, this yields a Bonferroni-corrected significance level of 0.05/15 = 0.0033. Cross validation is used to ensure that reported measures of model performance are not overstated. In particular, when variable selection techniques are used, cross validation is applied to the entire sequence of modeling steps including the selection of predictors.
1. Response Variables

There are two response variables for each participant:
1. Recovery Time (continuous)
2. Normal versus Delayed Recovery (binary)

Recovery Time

Recovery times are discrete variables, since values are discretized to the nearest day. For the purposes of this report however, Time to Recovery are considered as “suitably continuous”, and methods for analyzing continuous variables are used. Recovery times in the sample span a wide range, from a minimum of 6 days to a maximum of 268 days, so any error introduced in discretizing to the nearest day can be considered small relative to the magnitude of the variable.

As per clinical management protocols, examinations were conducted weekly for the first four weeks (the expected window of recovery), after which they were conducted every two weeks while the patient was referred for focused therapy. For patients with long recovery times, examinations could take place every four weeks as needed. This introduces an element of uncertainty as to exactly when “recovery” took place, with this uncertainty increasing for longer recovery times. Exact recovery times are therefore interval censored with recovery usually taking place sometime before being cleared for return-to-play.

However, the intention is for one of the models in this report to be used to develop a scoring system to predict the risk of persistent symptoms. If interval censoring is ignored, then the predicted recovery times from the model will be greater than the actual recovery times, having been biased high by the specific way in which the study was run with weekly examinations. The distribution of recovery times is shown in the figure below, binned into intervals of 5 days. The blue line shows a normal distribution fitted to the recovery times.

Figure 1.1: A histogram of recovery times shows a right-skewed distribution with a clear departure from normality

It can be seen in Figure 1.1 that recovery times are right-skewed, as is often found with time-to-event data. The mean recovery time is approximately 35 days, while the median is much less at 22 days, and there are also a small number of patients (17/270) with very long recorded recovery times of 100 days or more. It can also be seen that recovery times are far from normally distributed. This is confirmed by the Shapiro-Wilk test for normality ($p = 2.35 \times 10^{-23}$) that is statistically significant at a 0.05 level of significance.

In cases such as this where variables depart significantly from normality and equal variance, variable transformations can be explored to correct for these departures. A frequently used transformation for time-to-event data is to fit linear models to the logarithm of the response rather than directly to the response. This technique is used for the accelerated failure time models and is often found to correct for variance that increases with the magnitude of a variable. The blue line again shows a normal distribution fitted to the data.
Figure 1.2: A histogram of log recovery times is much closer to normally distributed than that for the recovery times

We note in Figure 1.2 that visually the logarithms of the recovery times are much closer to normally distributed than are the recovery times. Although the departure from normality is still statistically significant at a 0.05 level of significance, with a p-value for the Shapiro-Wilk test for normality of $5.02 \times 10^{-7}$, this is however much less than that seen for the recovery times. For this reason, tests and models involving recovery times in the following chapters will use the logarithm of the recovery times rather than the recovery times directly.

Normal versus Delayed Recovery

For the purposes of this study, with a sample consisting entirely of adolescents, a classification of persistent symptoms or Persistent Post-Concussion Symptoms (PPCS) means that the time-to-recovery was greater than or equal to 30 days. Persistent symptoms are therefore a binary categorical variable. For descriptive purposes, “persistent symptoms” is described in this report as having the two levels “Yes” (for recovery times greater than or equal to 30 days), or “No” (for recovery times of 29 days or less). Of the 270 patients in this study, 98 (36.3%) displayed persistent symptoms. We can therefore set a lower bound on the accuracy of any predictive model of 64%, simply by predicting that no patients have persistent symptoms regardless of any information we have about the patient.
2. PREDICTOR VARIABLES

The sample data contains 15 predictor variables, which fall into three broad categories:

**Demographic:** Sex (binary), age (continuous), and previous concussions (discrete).

**Injury:** Time (in days) since injury when the first examination was conducted (continuous, discretized), and whether or not the injury was Single/Low-velocity Injury (binary).

**Physical Examination:** Results of the ten clinical tests from the Buffalo Concussion Physical Exam (binary), classified in each case by whether or not the test result is normal or abnormal.

We note that:
- Time-since-injury is not a covariate in the same sense as the other predictors, but is the time at which the physical examination predictors were measured. This will be discussed further in “Time since injury in Predictive Models”.
- The physical examination predictors are, strictly speaking, time-dependent covariates that were measured once at time-since-injury. This will also be discussed further.

1. Sex

Patient sex is represented as a binary variable, Male or Female. Descriptive statistics are shown below. We note that the majority of patients are male (62%), with 38% being female. Female patients have a slightly higher proportion of persistent symptoms (42%) than males (33%). However, the difference in proportions between male and female patients is not statistically significant, with a Pearson chi-squared p-value of 0.183. There is a small but statistically significant difference in mean log recovery times, with female patients having longer recovery times on average. The p-value for the two-sample t-test is 0.000273, which is statistically significant at a Bonferroni-corrected significance level of 0.00333. This is consistent with the literature on the effect of sex on recovery time.

![Figure 2.1](image)

Figure 2.1: Sex has a statistically insignificant association with persistent symptoms but is significantly associated with log recovery times at a Bonferroni-corrected level

2. Age

It can be seen that the mean patient age is 14.9 years, with a median of 15.1 years. Patient ages span a range from 8.36 to 18.7 years, with most patients being of high school age. A linear regression model does not show age to be a statistically significant linear predictor of log recovery times. The p-value for a test of linear independence is 0.258, which is greater than a significance level of 0.05. A logistic regression model does not show age to be a statistically significant predictor of persistent symptoms either. The associated p-value is 0.179, which is also greater than a significance level of 0.05. Age is therefore unlikely to be included as a predictor in a scoring system for persistent symptoms.
3. Previous Concussions

The number of previous concussions is a discrete variable, with a minimum value in this data set of zero and a maximum value of seven. We note that the majority of patients (54%) have not had a previous concussion, and only a small proportion of patients (less than 6%) had three or more previous concussions. An increasing number of previous concussions is clearly associated both with a higher proportion of persistent symptoms and with longer recovery times. This is confirmed by the logistic regression test (p-value of 0.00907) and linear regression test (p-value of 0.024) for no association, both of which are statistically significant at a 0.05 level. The proportion of patients with persistent symptoms is similar for patients with 0, 1, or 2 previous concussions, only rising from 32% to 38%. Hence, 3 or more previous concussion was added as a predictor in the model.

For this new variable, the proportion of patients with persistent symptoms is statistically significantly different between patients with and without several previous concussions at a 0.05 level of significance. 35% of patients with 2 or less previous concussions displayed persistent symptoms, compared to 67% of patients with 3 or more previous concussions.
4. Single-impact/Low-velocity Injury or Multiple impact/High-velocity Injury

Single/Low-velocity or Multiple/High-velocity injury is a binary categorical variable and is obtained from the history provided to the physician who categorizes this injury. We note that most of the injuries were Single/Low-velocity Injury (n = 242), which is approximately 90% of the total. The Pearson chi-squared p-value is 0.00232, which is statistically significant at a Bonferroni-corrected level of significance of 0.00333. Recovery times are clearly longer on average for injuries that are Multiple/High-velocity Injury. This is confirmed by the two-sample t-test, which has a p-value of $1.93 \times 10^{-7}$. This is also statistically significant at a Bonferroni-corrected level of significance of 0.00333.

**Figure 2.4:** Single/Low-velocity Injury is significantly associated with both persistent symptoms and log recovery times at a Bonferroni-corrected level

5. Time-Since-Injury

**Figure 2.5:** Time since injury is significantly associated with both persistent symptoms and log recovery times at a Bonferroni-corrected level

Time-since-injury is measured in days from the initial injury until the first physical examination. We note that there is a clear positive correlation between time-since-injury and log recovery times. This is confirmed by the test for linear independence, which has a p-value of $2.3 \times 10^{-15}$. Time-since-injury is a
statistically significant predictor of persistent symptoms, with a logistic regression test of independence p-value of $3.49 \times 10^{-10}$.

6. Orthostatic Intolerance (OI)

The proportion of patients with persistent symptoms is statistically significantly different between patients with and without OI. 27% of patients without OI displayed persistent symptoms, compared to 48% of patients with it. The Pearson chi-squared test for independence has a $p$-value of 0.000607, which is less than a Bonferroni-corrected significance level of 0.00333. Mean log recovery times are also statistically significantly different between patients with and without OI. The two-sample t-test for equal means has a $p$-value of 0.00176, which is also less than a Bonferroni-corrected significance level of 0.00333.

7. Neck Spasm

A small proportion of patients (7%) have abnormal neck spasm. The proportion of patients who displayed persistent symptoms is almost the same in patients with normal neck spasms (36%) as in patients with abnormal neck spasms (42%). The difference is not statistically significant, with a Pearson chi-squared $p$-value of 0.765. Mean log recovery times for patients with normal or abnormal neck spasms are also not statistically significantly different. The $p$-value of the two-sample t-test for equal means was 0.707.

8. Neck Tenderness

Approximately equal proportions of patients had normal (51%) and abnormal (49%) neck tenderness. The proportion of patients who displayed persistent symptoms is almost the same in patients with normal neck tenderness (35%) compared to abnormal neck tenderness (38%). The difference is not statistically significant, with a Pearson chi-squared $p$-value of 0.687. Mean log recovery times for patients with or without abnormal
Neck tenderness are not statistically significantly different. The $p$-value of the two-sample t-test for equal means was 0.545.

9. Neck Range of Motion

A small proportion of patients (12%) have abnormal neck range of motion. The proportion of patients who displayed persistent symptoms is similar in patients with normal (38%) or abnormal (27%) neck range of motion. The difference is not statistically significant, with a Pearson chi-squared $p$-value of 0.338. Mean log recovery times are not statistically significantly different between patients with normal versus abnormal neck range of motion. The $p$-value of the two-sample t-test for equal means was 0.756.

Figure 2.8: Neck range of motion is not significantly associated with either persistent symptoms or log recovery times.

10. Smooth Pursuits

The proportion of patients with normal smooth pursuits (41%) and abnormal smooth pursuits (59%) are fairly similar. A greater proportion of patients with abnormal smooth pursuits display persistent symptoms (44%) than do patients with normal smooth pursuits (26%). This difference is statistically significant at a Bonferroni-corrected significance level of 0.00333, with the Pearson chi-squared test for equal proportions having a $p$-value of 0.00325. Mean log recovery times are also statistically significantly different between patients with normal and abnormal smooth pursuits. The $p$-value of the two-sample t-test for equal means was 0.00222, which is also statistically significant at a Bonferroni-corrected significance level of 0.00333.

Figure 2.9: Smooth pursuits is significantly associated with both persistent symptoms and log recovery times at a Bonferroni-corrected level.
11. Horizontal Repetitive Saccades

The proportion of patients with normal horizontal saccades (41%) and abnormal horizontal saccades (59%) was fairly similar. The proportion of patients with abnormal horizontal saccades who display persistent symptoms (43%) is greater than those with normal horizontal saccades (27%). This difference is not statistically significant at a Bonferroni-corrected significance level of 0.00333. At a 0.05 level of significance, mean log recovery times are statistically significantly different between patients with normal and abnormal horizontal saccades.

12. Vertical Repetitive Saccades

The proportion of patients with normal vertical saccades (46%) and abnormal vertical saccades (55%) was very similar. The proportion of patients with abnormal vertical saccades who display persistent symptoms (44%) is greater than those with normal vertical saccades (28%). This difference is statistically significant at a 0.05 level of significance, with a Pearson chi-squared test p-value of 0.00994, although this was not statistically significant at a Bonferroni-corrected significance level of 0.00333.

13. Vestibulo-Ocular Reflex (VOR)

The proportion of patients with normal (54%) and abnormal VOR (46%) was very similar. A greater proportion of patients with abnormal VOR display persistent symptoms (48%) than do patients with normal VOR (27%). This difference is statistically significant at a Bonferroni-corrected significance level of 0.00333, with a Pearson chi-squared test for equal proportions p-value of 0.00043. The p-value of the two-sample t-test for equal means was 0.00426, although this was not statistically significant at a Bonferroni-corrected significance level of 0.00333.
14. Near Point of Convergence (NPC)

A relatively small proportion of patients (20%) have abnormal NPC. The proportion of patients who displayed persistent symptoms is almost the same in patients with normal (36%) or abnormal (39%) NPC. The difference is not statistically significant.

Figure 2.12: Near point of convergence is not significantly associated with either persistent symptoms or log recovery times

15. Tandem Gait

A fairly similar proportion of patients have normal tandem gait (58%) as have abnormal tandem gait (42%). At a 0.05 level of significance, the proportion of patients with normal tandem gait who displayed persistent symptoms (31%) is not statistically significantly different to those with abnormal tandem gait (43%). The Pearson chi-squared test for equal proportions had a value of 0.068. Mean log recovery times for patients with normal or abnormal near point of convergence are not statistically significantly different. The p-value of the two-sample t-test for equal means was 0.065.

Figure 2.13: Tandem gait is not significantly associated with either persistent symptoms or log recovery times
Summary

There are very few missing values in the sample data, with only 7 values missing for OI, while the rest of the data is complete. Observations containing missing values are included when generating descriptive statistics for each predictor, but are removed when constructing predictive models. This yields a training set of 263 observations. The following graph summarizes the test for independence results for each predictor. P-values are plotted on a log scale, so the longer bars correspond to smaller p-values. At a 0.05 level of significance, age, convergence, neck range of motion, neck spasm, neck tenderness, and tandem gait have no statistically significant association with either log recovery times or with persistent symptoms. They are therefore unlikely to be good predictors of persistent symptoms.

Figure 2.14: Tests of independence between predictor and response variables

![Graph showing tests of independence between predictor and response variables.](image-url)
3. TIME-SINCE-INJURY IN PREDICTIVE MODELS

Time-since-injury was the predictor most significantly associated with both persistent symptoms (p = 3.49×10^-10) and log recovery times (p = 2.3×10^-15). Representing this information correctly in predictive models is therefore critical to model fit and performance. The simple linear regression model below predicts recovery time from time-since-injury, and shows the strong association that exists between the two variables.

**Figure 3.1. Linear regression model predicting recovery time from time-since-injury**

The red line shows the fitted model, the gray band a 95% confidence interval for the fitted regression line, and the dashed black line the lower limit for recovery time (i.e. time-since-injury).

The model has the approximate form: Recovery time = 14 + 3.5*time-since-injury (in days)

**Time-since-injury in Cox PH Models**

Time-since-injury represents a delayed entry time, and would seem to be represented most naturally in time-to-event models as a left-truncation point. A key assumption however when using left-truncation in Cox PH models is that the event times and the delayed entry times are independent, given the covariates. Since recovery times are strongly dependent on time-since-injury, an alternative approach is to include time-since-injury directly as a predictor in Cox PH models. To explore whether hazard ratios are constant across time for patients with different time-since-injury, we plot estimated hazard rates over time for patients in three groups: time-since-injury between 0 and 5 days (148 patients), between 6 and 10 days (83 patients), and greater than 10 days (32 patients). Estimated hazard ratios for patients with different time-since-injury are were not constant over time. Hence, a stratified Cox PH Model will be used going forward.

**Time-Since-Injury in Accelerated Failure Time Models**

As with Cox PH models, a theoretically well motivated way to represent time-since-injury is as a delayed entry time. Hence, a question of interest is whether any AFT models could adequately model the relationship seen in the sample data between recovery times and time-since-injury through left-truncation alone. The Weibull and log-logistic distributions both have closed-form survival functions, so the relationship between time-since-injury and median recovery times can be calculated analytically for these models.

**Time-Since-Injury in Binomial Generalized Linear Models**

Since there is a one-to-one correspondence between AFT models and bGLMs, different choices of error distribution for AFT models yield different link functions for their associated binomial GLMs, while the linear predictor for a bGLM is a linear transformation of that for the associated AFT model. Representing time-since-injury as a predictor in AFT models therefore naturally leads to also representing time-since-injury as a predictor in binomial GLMs.
4. MULTICOLLINEARITY

We note that multicollinearity is a function of the design matrix alone (i.e., just the predictor data), and not of the response. Before we develop predictive models, it is therefore worth examining this aspect of the data to see if any problems may occur. Given the nature of the predictors, which include multiple clinical tests of the same neurological systems, it is reasonable to assume that correlations between predictors will exist. Two primary ways to reduce the effects of multicollinearity are:

1. Remove highly correlated predictor variables from the model.
2. Use partial least squares regression or principal component analysis to reduce the number of predictors to a smaller, uncorrelated set.

**Independence Tests Between Predictor Variables**

*Figure 4.1: Tests of independence between predictor variables*

Every predictor variable is plotted against every other predictor, with colored bars indicating proportions of abnormal test results, or male patients in the case of sex. Labels on the diagonal refer to the x-axis in the same column and the y-axis in the same row. P-values for tests of independence are color-coded on each graph. Red denotes significance at a Bonferroni-corrected level, corrected for 120 pairwise tests.
We observe the following:

1. Age and several previous concussions are not significantly associated with any other predictors at a 0.05 level of significance.
2. Convergence is not significantly associated with most other predictors at a 0.05 level of significance, and with no other predictors at a Bonferroni-corrected level of significance.
3. Single/Low-velocity injury and time-since-injury have little significant association with most other predictors. This is particularly relevant for time-since-injury, as it means that the symptoms for patients entering the study a longer time after their injury are neither better nor worse than those for patients whose injury was more recent. This is explored in more detail later.
4. Sex is not significantly associated with most other predictors, with the main associations being with OI, single/low-velocity injury, and time-since-injury.
5. Horizontal saccades, vertical saccades, smooth pursuits and VOR form a cluster of related predictors, with each predictor significantly associated with the others at a Bonferroni-corrected level of significance.
6. Tandem gait is less strongly related to this cluster, being significantly associated with vertical saccades and VOR at a Bonferroni-corrected level of significance, and with horizontal saccades and smooth pursuits at a 0.05 level of significance.
7. OI is further out of this cluster, being significantly associated with VOR at a Bonferroni-corrected level of significance, and to horizontal saccades, vertical saccades, and tandem gait at a 0.05 level of significance, but not with smooth pursuits.
8. Neck tenderness is significantly associated with both neck range of motion and neck spasms at a Bonferroni-corrected level of significance, although neck range of motion and neck spasms are not significantly associated with each other.

**Time-Dependent Covariates**

We are aware that the predictors from the initial physical examination are in fact *time-dependent covariates*. Whereas the values of the predictors age, sex, several previous concussions and type of injury are fixed at the time of the concussive injury, the results of the physical examination tests are dependent on the recovery process and can be expected to vary depending on how long after the injury the initial examination is performed. Time-dependent covariates are known to cause problems with regression models in several ways. Firstly, it is far from obvious how to model the functional form of a time-dependent covariate, and doing so introduces new complexity to models with an associated risk of overfitting. Secondly, with time-dependent covariates, the ability to predict in Cox regression models is usually lost. The existence of a measurement at a particular time implies that a subject is still in the study, and has not yet recovered. This may not be a particular issue however, given that tests are only performed once at entry to the study. Since left-truncation occurs at the same time as tests are conducted, the inclusion of patients in the study at this time is already explicitly modeled using left-truncation. This is a difficult issue to deal with rigorously in time-to-event models. Correlation between physical exam components and time-since-injury was analyzed and no predictor was significantly time dependent after correction. In the absence of any clear techniques to incorporate this information apart from the use of left-truncation, we do not attempt to explicitly model time-dependent covariates in this analysis.
5. RESULTS OF PREDICTIVE MODELS

Results of each model fit are presented in this section. The best performing model out of all classes of model will then be used to develop a scoring system for persistent symptoms. Leave-one-out cross validation is used on all models to estimate performance on new data. Unless otherwise indicated, references to model performance refer to cross validated results rather than results on training data.

Model Coefficients for Models Based on All Predictors

Figure 5.1: Model coefficients for models based on all predictors
It can be seen that for models which include time-since-injury as a predictor (the AFT and bGLM), this is the most significant predictor of persistent symptoms at a Bonferroni-corrected level of significance. Orthostatic intolerance and single/low-velocity injury are also statistically significant predictors in all three models at either a Bonferroni or a 0.05 level of significance. The bGLM with a corresponding log-log link function has the greatest number of statistically significant predictors, while the Cox PH model has the least. A reason may be that stratifying by time-since-injury in the Cox PH model reduces power by reducing the number of observations available to estimate the hazard rate for each stratum.

Model Coefficients for Stepwise-Selected Models Based on All Predictors

Figure 5.2: Coefficients for stepwise-selected models based on all predictors

It can be seen that the number of predictors in these models is less than half that of the full models. These models would therefore result in a simpler scoring system. The most significant predictors in the full models of the previous section have all been retained in the stepwise selected models (time-since-injury, OI, simple injury), but the order of the less significant predictors has changed slightly as predictors have been dropped from the models. Although not shown here, the number of predictors in the AFT models with higher AIC (log-logistic, lognormal and Weibull) was greater (7 predictors in each model).

Interaction Terms

There are several reasons for considering pairwise interaction terms between the predictors. The first is that statistically significant interactions between predictors may well exist. There are however pitfalls involved with using interaction terms, particularly the risk of overfitting the data. With 15 predictors there are 105 pairwise interaction terms, and only 270 observations in the sample data. Including all these terms in a model would severely violate the suggestion that the number of observations used to fit a regression model should exceed the number of predictors (or candidate predictors if using variable selection) by a factor of at least ten.
**Figure 5.3: Interaction terms between predictor variables in two-variables models**

![Table showing interaction terms](image)

Significance is shown using both a 0.05 level of significance and a Bonferroni-corrected level of significance of 0.05/105 = 0.000476 since there are 105 separate interaction terms being tested.

It can be seen that the only interaction terms that are statistically significant at a 0.05 level of significance are those involving VOR/OI, and sex/neck tenderness. No interaction terms are significant at a Bonferroni-corrected level of significance. This is an interesting result given that OI and VOR were both found to be individually significant predictors of persistent symptoms. The subsequent models will include all predictors plus all pairwise interaction terms. This makes 105 variables in total for the Cox PH model (14 main effects plus 91 interaction terms), and 120 variables for the AFT and binomial GLMs which include time-since-injury as a predictor (15 main effects plus 105 interaction terms). Stepwise selection using the AIC is then employed to see if any interaction terms are considered sufficiently predictive to be included.
Model Coefficients for Stepwise-Selected Models Based on All Interactions

*Figure 5.4: Coefficients for stepwise-selected models based on all pairwise predictor interactions*

It can be seen that very few interaction terms are included in the stepwise selected models. This is expected given that only two interactions were individually significant at a 0.05 level when included in two-variable models. Only two interaction terms have been included in the Cox PH model and binomial GLM, and none in the inverse-Weibull model. Although not shown, these models have less terms than the other AFT models (7 in the log-logistic AFT, 9 in the lognormal AFT, and 13 in the Weibull AFT), and in the other binomial GLMs (15 in the models with logit and probit links, and 12 in that with a cloglog link). Hence, the models shown are the simplest in their class.

**Summary**

The table below summarizes the results of modeling. Pairs of numbers in a cell refer to performance on training and test data, with the upper number showing performance on the training data, and the lower number the performance using cross-validation. For AFT and binomial GLM models, results are only given for the model with the lowest AIC. This was the inverse-Weibull in every case, and the binomial GLM with a corresponding log-log link in every case except the stepwise-selected model include all pairwise predictor interactions.
Table 5.1: Summary of model performance

<table>
<thead>
<tr>
<th>Model (best AFT)</th>
<th>Binomial GLM (Clog-log)</th>
<th>AFT Model (Inverse-Weibull)</th>
<th>Cox PH Model (Stratified)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy (%)</td>
<td>Sensitivity (%)</td>
<td>Specificity (%)</td>
</tr>
<tr>
<td>Models with Main Effects Only</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Null model</td>
<td>64</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Demographic</td>
<td>73</td>
<td>47</td>
<td>87</td>
</tr>
<tr>
<td>Exam</td>
<td>67</td>
<td>35</td>
<td>86</td>
</tr>
<tr>
<td>All predictors</td>
<td>76</td>
<td>54</td>
<td>89</td>
</tr>
</tbody>
</table>

Models Including Interaction Terms

| All interactions | 92          | 87            | 96             | 338                  | 84          | 66            | 95             | 2167                 | 80          | 71            | 85             | 1965                  |
| All interactions | 77          | 57            | 89             | 266                  | 73          | 43            | 92             | 2150                 | 76          | 54            | 89             | 1901                  |
| All interactions | 78          | 56            | 91             | 263                  | 73          | 43            | 92             | 2150                 | 75          | 45            | 92             | 1900                  |

Binomial GLM Assumptions

Before we use this model to devise a scoring system, we perform the Hosmer-Lemeshow test of goodness of fit commonly used in regression models.

Figure 5.5: Binomial GLM test of goodness of fit

The p-value for the Hosmer-Lemeshow tests is greater than 0.05 for all binomial GLMs, hence we can be confident that our model is not violating assumptions of logistic regression.
6. DEVELOPMENT OF THE RDR-SCORE

The following table summarizes the details of the “best” model.

Table 6.1: Details of “best” model

| Variable                              | Coeff.  | Std. Err. | z       | P > |z|  | 95% CI          |
|---------------------------------------|---------|-----------|---------|-----|---|----------------|
| Time-since-injury                     | 0.2585  | 0.0363    | 7.125   | 1.04×10^{-12} | 0.1874 | 0.3296          |
| Orthostatic intolerance               | 1.3051  | 0.3716    | 3.512   | 0.0004 | 0.5768 | 2.0333          |
| Vestibulo-ocular reflex               | 1.2874  | 0.3714    | 3.466   | 0.0005 | 0.5595 | 2.0154          |
| > 2 previous concussions              | 1.0754  | 0.4545    | 2.366   | 0.0180 | 0.1846 | 1.9663          |
| Tandem gait                           | 0.2931  | 0.2596    | 1.130   | 0.2586 | -0.2155 | 0.8018          |
| High/multiple impact injury           | 0.5003  | 0.4724    | 1.059   | 0.2896 | -0.4256 | 1.4262          |
| OI*VOR                               | -1.1326 | 0.4969    | -2.279  | 0.0226 | -2.1064 | -0.1687          |
| High impact*tandem gait               | 1.2672  | 0.7760    | 1.633   | 0.1025 | -0.2538 | 2.7882          |
| Intercept                             | -3.6921 | 0.4106    | -8.991  | < 2×10^{-16} | -4.4970 | -2.8872          |

The following transformation are made to form a simpler and more intuitive system:

1. Integer coefficients: A scoring system based on integer values would be simplest to use and understand. The main issue here is transforming the model coefficients to a set of integers while minimally changing the model predictions.

2. Positive coefficients: The coefficient for having a single/low-velocity injury is negative in the binomial GLM. A scoring system should reverse this so that the reference level for each predictor is the lowest risk level.

3. Zero minimum: The intercept term means that the linear predictor does not equal zero when all predictors have a value of zero. A minimum value of zero for the scoring system would be easier to understand, and correspond to the lowest possible risk.

To convert all coefficients for main effects to positive values, a new model was generated with single/low-velocity injury as the reference level. The new coefficients are shown below, with “multiple/high-velocity” being the opposite of “single/low-velocity”.

Figure 6.1: Predictors chosen for the stepwise-selected corresponding log-log binomial GLM based on all predictor interactions with minimum risk reference levels

An integer-based scoring system can be created by scaling all estimated coefficients by a multiplier \( m \). The table below shows the scores assigned to each predictor for each value of \( m \). The last two rows show the root mean squared error per predictor and the maximum score for each system.
Table 6.2: Integer scores for each predictor based on different values of the multiplier m

<table>
<thead>
<tr>
<th>Predictor multiplier m</th>
<th>2.487</th>
<th>3.832</th>
<th>6.277</th>
<th>10.233</th>
<th>14.055</th>
</tr>
</thead>
<tbody>
<tr>
<td>OI</td>
<td>3</td>
<td>5</td>
<td>8</td>
<td>13</td>
<td>18</td>
</tr>
<tr>
<td>VOR</td>
<td>3</td>
<td>5</td>
<td>8</td>
<td>13</td>
<td>18</td>
</tr>
<tr>
<td>Over 2 previous concussions</td>
<td>3</td>
<td>4</td>
<td>7</td>
<td>11</td>
<td>15</td>
</tr>
<tr>
<td>Tandem gait</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Multiple/high velocity injury</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>VOR*OI</td>
<td>-3</td>
<td>-4</td>
<td>-7</td>
<td>-12</td>
<td>-16</td>
</tr>
<tr>
<td>High velocity*tandem gait</td>
<td>3</td>
<td>5</td>
<td>8</td>
<td>13</td>
<td>18</td>
</tr>
<tr>
<td>RMS error</td>
<td>0.056</td>
<td>0.021</td>
<td>0.013</td>
<td>0.008</td>
<td>0.005</td>
</tr>
<tr>
<td>Maximum score</td>
<td>11</td>
<td>18</td>
<td>29</td>
<td>46</td>
<td>64</td>
</tr>
</tbody>
</table>

The decrease (on average) in the root mean squared error per predictor as \( m \) increases shows that, if we allow \( m \) to be arbitrarily large, the transformed values can be made as close to the original as desired. The disadvantage of increasing \( m \) is that the scores also become larger, leading to a more complicated scoring system. Ideally, the value of \( m \) should be chosen so that (i) the estimated probabilities calculated using the scoring system are close to those using the binomial GLM, and (ii) the integers used in the scoring system are low. To check (i), we plot below the predicted estimated using the binomial GLM against scores calculated on the sample data. The actual outcomes (persistent/non-persistent symptoms) for each patient are color coded.

It can be seen that when \( m = 3.832 \), the score coefficient of 0.9905 for time-since-injury is fortuitously close to the integer 1. Using this value of \( m \) would therefore mean that time-since-injury would not need to be multiplied by a non-integer value, but could simply be included in the score without modification. Assigning a score to a patient is simply done by summing the scores for each predictor. The interaction terms in this scoring system can be thought of as adjustments for additional or less risk.
Table 6.3: Final integer-based scoring system for estimating the risk of Delayed Recovery

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time-since-injury</td>
<td>1 point per day</td>
</tr>
<tr>
<td>High-velocity/multiple impact</td>
<td>2 points</td>
</tr>
<tr>
<td>More than 2 previous concussions</td>
<td>4 points</td>
</tr>
<tr>
<td>OI</td>
<td>5 points</td>
</tr>
<tr>
<td>VOR</td>
<td>5 points</td>
</tr>
<tr>
<td>Tandem gait</td>
<td>1 point</td>
</tr>
<tr>
<td>OI and VOR</td>
<td>-4 points</td>
</tr>
<tr>
<td>High-velocity/multiple impact and tandem gait</td>
<td>5 points</td>
</tr>
</tbody>
</table>

The final part of the scoring system is deciding how to map scores to risk categories. Overall accuracy is dependent on which scores are chosen as cutoffs for the risk categories. The binomial GLM had 78% training accuracy, with 51% sensitivity and 90% specificity using a cutoff probability of 0.5. Choosing a cutoff of 0.3 for low risk decreases overall accuracy and specificity, but increases sensitivity considerably (from 51% to 72%). From the graph, it can be seen that most patients with a score of 10 or less (corresponding to a probability of 0.30) did not develop persistent symptoms, while most patients with a score of 15 or greater did (corresponding to a probability of 0.70). Patients with scores from 11 to 14 appear to be a mixture of those who did and did not develop persistent symptoms. This suggests that we can classify patients into low-, medium-, or high-risk categories using scores of 10 and 15 to mark the boundaries of low/medium and medium/high risk respectively.

Figure 6.3: Classification of patients using the scoring system based on all training data

<table>
<thead>
<tr>
<th>Score</th>
<th>Low risk</th>
<th>Medium risk</th>
<th>High risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>122 (73%)</td>
<td>38 (23%)</td>
<td>6 (4%)</td>
</tr>
<tr>
<td></td>
<td>No persistent symptoms</td>
<td>Persistent symptoms</td>
<td></td>
</tr>
<tr>
<td>Count</td>
<td>27 (28%)</td>
<td>27 (28%)</td>
<td>43 (44%)</td>
</tr>
</tbody>
</table>

The proportions in each risk category were checked and found to be almost identical to those obtained using the estimated probabilities generated by the binomial GLM using cutoff probabilities of 0.3 and 0.7 (proportions were identical in the low category, and only differed by one or two patients in the medium and high-risk categories). This suggests that any discretization errors introduced by using the scoring system are minimal.
7. CROSS-VALIDATION AND SUMMARY

The first step in this process checks that the predictors chosen for the model trained on all the data are the same ones that would be chosen using leave-one-out cross validation. This was done by counting the number of stepwise-selected models that included each predictor (and their interaction pairs) using leave-one-out cross validation. We note that the OI*VOR interaction term was counted as either OI*VOR or VOR*OI, and was actually selected for every model. The predictors used to develop the scoring system are present in nearly every model. The exception is the multiple/high-velocity*tandem gait interaction term, which was not selected using the AIC in just 2 models (less than 1%). A small number of models (13% or less) included predictors not present in the scoring system, but these are all of low statistical significance. It would have been of more concern if the scoring system included predictors which only appeared in a few stepwise selected models, which did not occur. It seems therefore that stepwise selection of predictors for the model is reasonably robust, and that the predictors used in the scoring system have been validly included.

The next step is to calculate the scores that would be generated using leave-one-out cross validation for each model. Ideally the variation in the maximum likelihood estimates for the model coefficients will be small enough that the rounding process will lead to these matching the scores generated using all the sample data, but again this is a matter for experimental verification. The scoring system therefore seems relatively robust. The main reason for this is that the rounding process used to convert coefficients to integer scores is tolerant of coefficient changes that are small relative to the magnitude of the coefficient. To calculate the cross validated scores we generate a scoring system for every patient on a training set with the patient removed, then scored the patient according to that system. The scoring system is generated using stepwise-selected models based on all predictor interactions, so the models may differ in the predictors chosen. Scores for each predictor are found by dividing by the coefficient for time-since-injury and rounding, which fixes the score for time-since-injury at one for every model.

Figure 7.1. Cross-validation of the BCPE RDR-Score using Leave-one-out method

For the model tested on new data using cross validation, 69% of patients without persistent symptoms fall into the low-risk category, and 67% of patients with persistent symptoms fall into either the medium- or high-risk categories. Performance on test data is therefore somewhat lower than on training data.

Conclusions

The goal of this study was to develop a scoring system to quantify the risk of persistent symptoms for adolescents presenting within 14 days of injury. The strategy adopted has been to fit predictive statistical
models to the patient data collected by the University at Buffalo Concussion Management Clinic, then use the parameters of the model to develop the scoring system.

Two factors have complicated the modeling process when compared to typical regression modeling: the presence of a delay between the concussive injury and the initial physical examination (referred to as time-since-injury), and the presence of time-dependent covariates in the results of the initial physical examination. The approach taken in this report has been to “let the data decide” how to incorporate this information into models. Analysis of the effect of these factors on the responses justified including them as predictors, which has allowed simpler and more understandable models to be developed.

The existence of a long tail in the distribution of recovery times has also been the subject of extensive analysis. Models based on the right-skewed extreme value (maximum) distribution were found to best accommodate this long tail. The inverse-Weibull AFT model and binomial GLM with a corresponding log-log link function, which are based on this distribution, were found to provide the best fit and simplest stepwise selected models for recovery times and persistent symptoms respectively.

A scoring system has been developed based on this binomial GLM that is simple, easy to use, and easy to understand. It employs three questions about a patient’s injury and concussion history, and three clinical tests. These tests and demographic characteristics make sense clinically and statistically, and for the most part are the same predictors that would be selected by a simple univariate analysis of statistical significance.