Assessing potential risk factors of concussion using three-year collegiate athletes data

Muhammad Waleed[a], Guanlin Zhang[a]*

Abstract: Concussion, a type of mild Traumatic Brain Injury (mTBI), is considered one of the major public health issues. Studies are being conducted, especially in the sports arena, to determine its risks factors, and to consequently propose different strategies to minimize its risks. In this manuscript, we aim to study the effects of various risk factors on the incidence of concussion among college athletes, with a specific inclination towards determining the effects of gender and different game settings. To suffice our purpose, we use a three-year collegiate athletes data set comprising of yearly counts of reported concussion cases cross-classified against gender, game setting, and type of sports. We analyzed the dataset with the help of non-regression model methods (such as Chi-square test, the CMH test) as well as logistic regression modeling approach. Our results demonstrate that the incidence of concussion is significantly associated with the gender of athletes, game setting and type of sports. Furthermore, based on our fitted logistic model, we have sufficiently strong evidence to conclude that the interaction between game setting and type of sports is statistically significant.

Keywords: Concussion, Association, Chi-Square Test, Logistic Regression, Goodness of Fit

1. Introduction

According to the Centers for Disease Control and Prevention (CDC), a concussion is defined as "a type of traumatic brain injury caused by a bump, blow, or jolt to the head that can change the way your brain normally works. Sometimes a concussion can even occur from a blow to the body that causes the head to move rapidly back and forth (like a severe whiplash)." It is also referred to as a type of mild Traumatic Brain Injury (mTBI) [1]. Concussion has been labeled as a major public health issue due to its acute and potential long term effects associated with this injury [2].

Athletes are population at high risk of concussion, and cumulative effects of repeated concussions can have long term consequences [3, 4, 5]. According to University of Pittsburgh's Brain Trauma Research Center, more than $300,000$ sports-related concussions occur annually in the US, and the likelihood of suffering a concussion while playing a contact sport is estimated to be as high as 19 percent per year of play. More than $62,000$ concussions are sustained each year in high school contact sports and, among college football players, 34 percent have had one concussion and 20 percent have endured multiple concussions [6]. Some other recent studies have shown that approximately 1.6 to 3.8 million sport and recreation related concussions (SRC) occur every year in the U.S [2, 7].

Our goal of this project is to gain insight into the association between incidence of concussion and some of its potential risk factors using the data reported in a study, conducted by Covassin et al. in the year 2003, on the gender differences and the incidence of concussion among college athletes from participating institutions of National Collegiate Athletic

[a]*Department of Biostatistics, University of Kansas Medical Center, 3901 Rainbow Boulevard, Kansas City, KS 66103*
**Correspondence to:* gzhang2@kumc.edu

Association (NCAA) [5]. Covassin et al. found that, over the 1997-2000 academic years, a higher percentage of female athletes sustained concussions during games than the male athletes, and of all the sports under the study, women's soccer and men's lacrosse were found to have the highest injury rate of concussions [5]. In this manuscript, besides gender differences, we will investigate several risk factors including different types of sports and game setting over the 1997-2000 period. We will also attempt to study the involvement of other risk factors as well as the existence of potential interactions between risk factors. The topic we are presenting is of interest due to the following reasons. First of all, it has been suggested that concussions are more common in some collegiate sports than previously anticipated [8, 9]. Secondly, only a few studies have compared collegiate sports or identified which athletes are more at risk for suffering concussions [5, 9]. Finally, almost no studies have compared the difference of concussion incidence between actual games and practice occasions, and their interaction with different sports. We can argue that although practice is essentially a simulated game, for some sports the difference of nature or mechanics between practice and an actual game is more dramatic than the others. Those information above is useful in helping the NCAA committee on Competitive Safeguards and Medical Aspects of Sports recognize if there is a need to modify rules or equipment to help minimize the number of concussions sustained by collegiate athletes [5].

This manuscript has been organized in the following order: we describe a brief data description and method of analysis in the Section 2. Results have been described in the Section 3, followed by a discussion of results in the Section 4.

2. Method

The data used for this project was originally reported in a study conducted by Covassin et al. in the year 2003 [5]. The primary objective of their study was to asses the difference in the incidence of concussions among collegiate athletes on the basis of gender. They obtained this data using the Injury Surveillance System (ISS) of the National Collegiate Athletic Association (NCAA). As per the definition by ISS, a reportable injury is "the one that occurs as a result of participation in an organized collegiate practice or game, requires medical attention by a team athletic trainer or physician, and results in

		Males			**Females**		
Academic Year	**Sport**	Concussion Cases	Athlete Exposures	Incidence Rate	Concussion Cases	Athlete Exposures	Incidence Rate
1997	Soccer	34	30966	1.0979	51	24981	2.0489
	Lacrosse	19	13486	1.4089	12	8762	1.3696
	Basketball	8	27706	0.2887	16	29413	0.5439
	Softball/ Baseball	22	51351	0.4284	9	26834	0.3354
	Gymnastics	0	227	0	1	8903	0.1123
1998	Soccer	27	19142	1.4105	47	22934	2.0494
	Lacrosse	15	9514	1.5766	7	7122	0.9829
	Basketball	21	39367	0.5334	30	38174	0.7859
	Softball/ Baseball	6	49207	0.1219	10	44280	0.2258
	Gymnastics	0	221	0	0	822	0
1999	Soccer	40	25636	1.5603	60	27167	2.2086
	Lacrosse	17	12177	1.3961	7	8531	0.8205
	Basketball	20	32836	0.6091	26	28992	0.8968
	Softball/ Baseball	25	80215	0.3117	28	75355	0.3716
	Gymnastics	0	1179	0	0	2083	0
	Total	254	393230	0.6459	304	354353	0.8579

Table 1. Total Game Concussions and Concussion Rates (per 1,000 Athlete Exposures) by Sport, Year, and Gender
Source: Covassin et al. [5]

Academic Year	Sport	Males: Concussion Cases	Males: Athlete Exposures	Males: Incidence Rate	Females: Concussion Cases	Females: Athlete Exposures	Females: Incidence Rate
1997	Soccer	10	127013	0.0787	12	87442	0.1372
	Lacrosse	11	76246	0.1443	14	33342	0.4199
	Basketball	16	118905	0.1346	24	104872	0.2289
	Softball/ Baseball	10	129623	0.0772	7	45291	0.1546
	Gymnastics	0	3897	0	4	31006	0.1290
1998	Soccer	8	77769	0.1029	9	77286	0.1165
	Lacrosse	4	56201	0.0712	3	31048	0.0966
	Basketball	31	164607	0.1883	26	140817	0.1846
	Softball/ Baseball	9	130988	0.0687	7	76076	0.0920
	Gymnastics	0	3265	0	0	9253	0
1999	Soccer	4	106241	0.0377	13	100855	0.1289
	Lacrosse	14	74115	0.1889	5	37765	0.1324
	Basketball	22	140637	0.1564	25	108480	0.2305
	Softball/ Baseball	9	213674	0.0421	17	129155	0.1316
	Gymnastics	0	9867	0	0	2083	0
	Total	148	1433048	0.1033	167	1032517	0.1617

Table 2. Total Practice Concussions and Concussion Rates (per 1,000 Athlete Exposures) by Sport, Year, and Gender
Source: Covassin et al. [5]

restriction of the student-athlete's participation for one or more days beyond the day of injury [5, 10]." Cerebral concussion is classified as a reportable injury according to the guidelines of ISS. The data under our consideration consists of the total yearly observed counts of cerebral concussions status (yes/ no) among collegiate athletes from each of the four US regions (East, South, Midwest, and West), and they have been cross-classified against four other categorical variable, namely gender (two categories: male, female), setting (two categories: actual game, practice), sport type (five categories: soccer, lacrosse, basketball, softball/baseball, gymnastics) and academic year (three categories: 1997-1998, 1998-1999, 1999-2000) [5]. Athlete-exposure refers to the number of instances "an athlete participating in one practice or game in which he or she is exposed to the possibility of athletic injury [5, 10]." This injury and athlete-exposure data was recorded by "certified athletic trainers from participating NCAA institutions on a weekly basis from the first day of pre-season practice to the final post-season game [5] ." Furthermore, it is worth mentioning that at least 10% of each US regions and the three NCAA Divisions (I, II, and III) are represented in this sample [5]. The observed data for the "actual game" and "practice" game setting has been classified by gender, sport and year in Tables 1 and 2, respectively. These observed data clearly suggests that concussions are rarely experienced by athletes participating in gymnastics. This may be due to the fact that gymnastics is not a contact sport, and hence we do not expect that athlete sustain a high number of head blows/ injuries. In this project, we will restrict our attention to the remaining four sports for data analysis purposes.

2.1. Statistical Analysis

As we described earlier, the main objective of our project is to determine whether the incidence of concussion is associated with gender and game setting, and if these variables interact. To suffice our purpose, we will analyze our data by utilizing both non-regression model tests as well as logistic model-fitting approach. Since all of our variables are nominal in nature, therefore, in our analysis without fitting any regression model, we will use both the chi-square test and the Cochran-Mentel-Haenszel (CMH) test to detect a general pattern of association. Instead of simply using the chi-square test, we also need to use the CMH test, because we need to control for possible variation which can potentially be caused by the type particular of sport or any academic year.

Using the model-based approach, we will fit logistic regression model to model the log odds/ probability of concussion (1 = “Yes”, 0 = “No”) with gender, game setting, sport type and academic year being the potential predictor variables in the model. For the model building purposes, we defined our potential explanatory variables as gender (G : 1 = “female”, 0 = “male”), game setting (S : 1 = “actual game”, 0 = “practice”), and the multi-category variables type of and academic year as

$$T_1 = \begin{cases} 1 & \text{if soccer} \\ 0 & \text{if baseball/ softball} \end{cases} \quad ; \quad Y_1 = \begin{cases} 1 & \text{if academic year 1997 – 1998} \\ 0 & \text{if academic year 1998 – 1999} \end{cases}$$

$$T_2 = \begin{cases} 1 & \text{if lacrosse} \\ 0 & \text{if baseball/ softball} \end{cases} \quad ; \quad Y_2 = \begin{cases} 1 & \text{if academic year 1999 – 2000} \\ 0 & \text{if academic year 1998 – 1999} \end{cases}$$

$$T_3 = \begin{cases} 1 & \text{if basketball} \\ 0 & \text{if baseball/ softball} \end{cases}$$

For the sake of analysis, we used academic year 1998 – 1999 as the reference group among the year categories, so that we can easily observe if the incidence of concussion varied from the start to the end of this study. For the multi-category sports variable, we chose baseball/ softball as the reference group because it has the largest observed sample sizes, which allows us to do comparisons with a stable group and consequently obtain stable results.

It is important to mention that the purpose of this study can also be achieved by constructing log-linear models (that allow us to model cell counts), instead of logistic regression models. Log-linear models, however, are more preferable when we have multiple response variables and our primary goal is to describe the associations and interactions in the joint distribution of those response variables. Furthermore, an increase in the number of explanatory variables leads to cumbersome log-linear models which may be difficult to interpret [11].

Statistical software SAS (Version 9.4) was used to perform the statistical analysis. Details of our analysis shall be presented in the upcoming sections.

	Academic Year			
Variable	1997 – 1998	1998 – 1999	1999 – 2000	**Total**
Concussion Status				
Yes	275 (0.03%)	260 (0.03%)	332 (0.03%)	867
No	935958 (99.97%)	984272 (99.97%)	1201499 (99.97%)	3121729
Gender				
Males	575296 (61.45%)	546795 (55.54%)	685531 (57.04%)	1807622
Females	360937 (38.55%)	437737 (44.46%)	516300 (42.96%)	1807622
Game Setting				
Game	213499 (22.80%)	229740 (23.33%)	290909 (24.21%)	734148
Practice	722734 (77.20%)	754792 (76.67%)	910922 (75.79%)	2388448
Type of Sport				
Soccer	270402 (28.88%)	197131 (20.02%)	259899 (21.63%)	727432
Lacrosse	131836 (14.08%)	103885 (10.55%)	132588 (11.03%)	368309
Basketball	280896 (30.00%)	382965 (38.90%)	310945 (25.87%)	974806
Softball/ Baseball	253099 (27.03%)	300551 (30.53%)	498399 (41.47%)	1052049
Total	936233	984532	1201831	3122596

Table 3. Frequency Distribution of Variables Across Different Academic Years (Percentages Are Across Each Year)

3. Results

For the data without gymnastics, the frequency distribution of each variable across different academic years has been presented in the Table 3. This table suggests that approximately 3% athlete-exposures (AE) sustain concussion injuries during each year. Furthermore, it shows that more male AE's were observed in comparison to female AE's and, as one should expect, most of the AE's were observed during practice sessions.

We are going to focus on the following three null hypotheses. Firstly, the incidence of concussion does not differ by gender conditioned an individual sport. Secondly, the incidence of concussion and gender / game setting are conditionally independent for each individual sport, conditioned on each academic year. Thirdly, the incidence of concussion among males and females in each sport does not vary across the levels of game setting (i.e. no interaction between gender and game setting), after controlling for academic years.

Although our data is quite large, we will still used $\alpha = 5\%$ level of statistical significance to perform analysis. Any particular hypothesis will be rejected if the P-value for the appropriate test statistic is observed to be less than 0.05.

3.1. Analysis without Fitting Regression Models

Before conducting analysis to test our main hypotheses, we performed a preliminary analysis to detect if there is any association between the concussion and each of the four explanatory variables (gender, game setting, type of sport). This was done using the Pearson's X^2 and likelihood-ratio G^2 chi-squared tests, since all of our variables are nominal in nature and estimated expected cell counts for all cells are greater than 5. Results of these tests have been summarized in the Table 4. It can be clearly observed that the incidence of concussion is associated with gender, game setting and type of sport.

	Pearson's X^2		**Likelihood-Ratio G^2**	
Variable	Observed Statistic	P-Value	Observed Statistic	P-Value
Gender	47.2255	< 0.0001	46.4949	< 0.0001
Game Setting	800.1651	< 0.0001	648.5783	< 0.0001
Type of Sport	130.5350	< 0.0001	132.8899	< 0.0001
Academic Year	1.5334	0.4646	1.5279	0.4648

Table 4. Pearson X^2 and Likelihood-Ratio G^2 Chi-Square Test of Association between Concussion and Other Variables

In order to test the hypothesis that the incidence of concussion across different sports does not differ by gender, we conducted Pearson's chi-square test after sub-setting data for each individual sport, and subsequently testing the marginal association between gender and concussion on the resulting data. Results have been summarized in the Table 5. Based on these results, we have sufficient evidence to conclude that concussion is marginally associated (by collapsing data across different academic years) with gender, conditioned on whether athletes play soccer, basketball or softball/ baseball. On the other hand, we did not observe any marginal association (across different academic years) between concussion and gender for those college athletes who play lacrosse. In the Table 5, we have also provided the Pearson's chi-square test statistic and its corresponding P-values as reported by Covassin et al. [5]. We observe[†] that our results do not exactly match with the ones obtained by Covassin et al. [5]. The reason behind this difference in results is still unknown to us.

The conditional association between concussion and gender/ game setting for each sport, after controlling for academic years, was tested using the Cochran-Mentel-Haenszel (CMH) test of conditional independence. It must be noted that the CMH test is appropriate when the assumption of homogeneous association is satisfied and the potential value $\sum_{k=1}^{3} (n_{11k} - \mu_{11k})$ is greater than 5 (in magnitude). For each sport, we found that the magnitude of potential value was indeed greater than 5. Furthermore, the assumption of homogeneous association was tested using the Breslow-Day

[†] In fact, we reach contrasting conclusions for the athletes who play softball/ baseball.

Variable	Our Result: Observed Statistic	Our Result: P-Value	Covassin et al. [5]: Observed Statistic	Covassin et al. [5]: P-Value
Soccer	25.2377	< 0.0001	12.9900	0.0003
Lacrosse	0.5579	0.4552	1.6300	0.2017
Basketball	9.0881	0.0026	5.1400	0.0233
Softball/ Baseball	8.6754	0.0032	0.9900	0.3197

Table 5. Comparison of Pearson X^2 Test of Marginal Association (Across Different Academic years) between Concussion and Gender, Conditioned on Each Individual Sport

(BD) test statistic, which is also applicable for our data because all estimated expected cell counts were greater than 5 in each partial table.

The P-values for the BD and CMH tests have been reported in the Table 6, and we make the following inferences:

1. Based on the results of BD test, we conclude that the association between concussion and gender/ game setting is homogeneous across all academic years, regardless of the type of sport.
2. These results also suggest a statistically significant conditional association between gender and concussion for those athletes who play soccer, basketball or softball/baseball, after controlling for academic years.
3. It must be noted, however, that we do not have sufficient evidence of conditional association between gender and concussion, for those athletes who play lacrosse. This is probably due to the fact that the P-value for the BD test is very close 0.05 (i.e. marginally insignificant), and in reality the association between gender and incidence of concussion for these athletes may not be homogeneous across all three academic year. This violates the assumption of homogeneous association required to use the CMH test, and consequently CMH test may not be powerful enough to detect any pattern of association.
4. Lastly, we have sufficient evidence to conclude that the conditional association between the incidence of concussion and game setting regardless of the type of sport is statistically significant, after controlling for academic years.

Sport	X Variable — Gender: Breslow-Day Test	Gender: CMH Test	Game Setting: Breslow-Day Test	Game Setting: CMH Test
Soccer	0.6548	< 0.0001	0.4240	< 0.0001
Lacrosse	0.0653	0.4064	0.1268	< 0.0001
Basketball	0.4496	0.0022	0.2735	< 0.0001
Softball/ Baseball	0.5477	0.0015	0.2322	< 0.0001

Table 6. P-Values for Breslow-Day Test of Homogeneous Association and CMH Test of Conditional Association between Concussion with Gender and Game Setting for Each Type of Sport Across Different Academic Yeas

To test the interaction hypothesis, we will rely on the model-based approach described in the next sub-section.

3.2. Model Based Approach

The purpose of this section is to choose an appropriate logistic regression model, with which we can model the odds of concussion on the basis of its potential risk factors. To suffice our purpose, we used various in-built selection procedures (forward, backward and stepwise) in the `PROC LOGISTIC` using selection and deletion criterion of 0.05 each. Before describing our model fitting results, we would like to mention that we are going to use the rule of thumb "20:1 for the ratio of $\min$(events, non-events) to number of candidate parameters in the model [12]." Let logit$(\widehat{\pi})$ denote the odds of concussion. Based on the results of different selection procedures, we obtain a hierarchical logistic regression model that includes main-effect of gender and an interaction between type of sports with game setting. In terms of the dummy variables defined in the Section 2.1, the fitted logistic model can be expressed as follows:

$$\text{logit}(\widehat{\pi}) = \widehat{\beta}_0 + \widehat{\beta}_1 G + \widehat{\beta}_2 S + \widehat{\beta}_3 T_1 + \widehat{\beta}_4 T_2 + \widehat{\beta}_5 T_3 + \widehat{\beta}_6 (S * T_1) + \widehat{\beta}_7 (S * T_2) + \widehat{\beta}_8 (S * T_3)$$

The maximum likelihood estimates of the above model parameters along with their corresponding 95% Wald's confidence intervals have been summarized in the Table 7. Interpretation of the estimated model parameters for each predictor variable (in terms of odds of sustaining concussion) has been presented in the Appendix A. Some more discussion on our fitted model and justification of our model choice shall be presented in the Section 4.

Parameter	Estimate	ASE	P-value	95% Wald's Confidence Interval
Intercept (β_0)	-9.5451	0.1334	< 0.0001	$(-9.8066, -9.2827)$
Gender (β_1)	0.3355	0.0685	< 0.0001	$(0.2012, 0.4698)$
Game Setting (β_2)	1.2880	0.1643	< 0.0001	$(0.9659, 1.6101)$
Type of Sport				
Soccer (β_3)	0.1371	0.1868	0.4629	$(-0.2289, 0.5031)$
Lacrosse (β_4)	0.7130	0.1912	0.0002	$(0.3383, 1.0878)$
Basketball (β_5)	0.7835	0.1548	< 0.0001	$(0.4802, 1.0869)$
Game Setting*Type of Sport				
Game*Soccer (β_6)	1.5738	0.2207	< 0.0001	$(1.1412, 2.0063)$
Game*Lacrosse (β_7)	0.7427	0.2441	0.0023	$(0.2643, 1.2211)$
Game*Basketball (β_8)	-0.0972	0.2054	0.6359	$(-0.4998, 0.3053)$

Table 7. Maximum Likelihood Parameter Estimates, Asymptotic Standard Error and 95% Wald Confidence Interval for Our Logistic Regression Model

4. Discussion

The purpose of this project was to assess potential risk factors of concussion among collegiate athletes using a three-year data. The data set under our consideration was reported in a 2003 study by Covassion et al. [5] in which their primary objective was determine any differences in the incidence of concussion among college athletes on the basis of gender. A brief discussion of our findings is presented below:

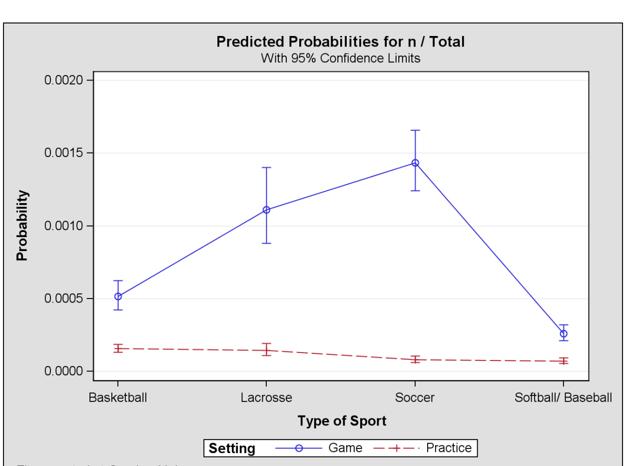
1. Using the Pearson's X^2 chi-square test, we inferred that the association between the incidence of concussion and the gender is statistically significant, conditioned on any individual sport. This conclusion is consistent with the data presented in the Tables 1 and 2, where incidence rates per 1,000 athlete exposures for females are greater than those for males (in almost all cases). Weaker neck strength for females is likely to be a major contributing factor behind their increased concussion incidence rate [5].

2. We also used the CMH test statistic to find that the conditional association between concussion and gender/ game setting, after controlling for academic years, is statistically significant for all sports except lacrosse. For the athletes who play lacrosse, we found that the conditional association between gender and concussion is not statistically significant. We used BD test of homogeneous association to find that the association between gender and incidence of concussion for the athletes who play lacrosse may not be homogeneous across all three academic year. This also gives us a hint that there might be interaction between predictors if we use model-based analysis approach.

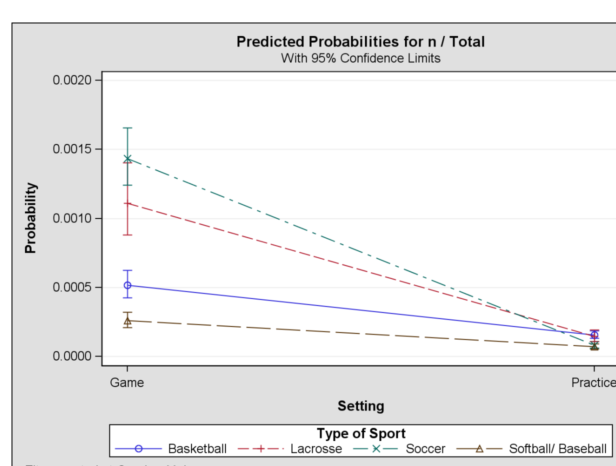
Model	**Deviance** Statistic	P-Value	AIC
$G+S+T+Y$	152.6875	<0.0001	385.4810
$G*S+T+Y$	151.8411	<0.0001	396.6345
$G*T+S+Y$	146.7646	<0.0001	385.5580
$G*Y+S+T$	151.3548	<0.0001	388.1482
$G+S*T+Y$	54.7229	0.0303	293.5164
$G+S*Y+T$	151.0311	<0.0001	387.8245
$G+S+T*Y$	138.3009	<0.0001	383.0944
$G*S*T+Y$	40.9519	0.0877	293.7453
$G*S*Y+T$	145.7673	<0.0001	392.5607
$G*T*Y+S$	124.5111	<0.0001	391.3046
$G+S*T*Y$	28.9251	0.1828	295.7186
$G+S*T$	56.3077	0.0359	291.1011

Table 8. Goodness-of-Fit Tests for Various Logistic Models

3. Using various selection procedures, we found that the ‘best’ hierarchical logistic regression model for our data is the one with main effect for gender, and an interaction between game setting and type of sports. We recall that the criterion of model selection is: (a) model parsimony and (b) adequate fit of the data. In other words, during model selection stage, we primarily attempt to choose simplest possible model which adequately fits the given data [11]. In the Table 8, we have presented AIC values for each fitted model. To choose appropriate model for our data on the basis on these values, we observe that our model is appropriate, since its corresponding AIC value is the lowest among all reported models [11].
Furthermore, we tested adequacy of each fitted model by using deviance test statistic. For each model, we tested the null hypothesis H_0 : fitted model fits the data as well as the saturated model, against the alternative H_1 : fitted model does not fit the data as well as the saturated model. The deviance test results for these model have also been presented in the Table 8. We observe that we have (marginal) evidence to reject the null hypothesis for our chosen model. It is important to mention that we still prefer this model, instead of the other models showing an evidence of adequate fit, because we want to choose a parsimonious model which is much easier to interpret. In addition, it is quite possible that those models are exhibiting adequate fit only due to over-fitting.



(a) Game setting on x-axis and type of sport as grouping variable



(b) Type of Sport on x-axis and game setting as grouping variable

Figure 1. Interaction plot depicting estimated effects of type of sport and game setting on the estimated probabilities of concussion, along with 95% confidence limits.

4. In the Table 7, we note that most of our fitted model parameters are statistically significant (P-values less than 0.05 and the corresponding confidence intervals not containing 0 in between). Based on our fitted model, we were unable to reject the null hypothesis of no interaction between gender and game setting for each sport, after controlling for academic years.
5. In order to visualize the interaction between type of sports and game setting more clearly, we plotted the predicted probabilities of concussion for each combination along with their respective 95% Wald's confidence interval. In the Figure 1(a), we observe that there exists a statistically significant difference between the predicted probabilities for various sports in the competitive game situation, however difference between sports during practice session is not quite apparent. On the other hand, we observe in the Figure 1(b) that there exists a statistically significant between the predicted probabilities of concussion for all sports among the two different game settings. Since the predicted probabilities are very close to zero, it can be argued that the Wald's method should not be used for analysis purposes due to its poor performance near the boundaries of parameters space [11]. For our project, we used Wald's method because its results were very close to the ones obtained using other approaches, such as the profile likelihood approach.

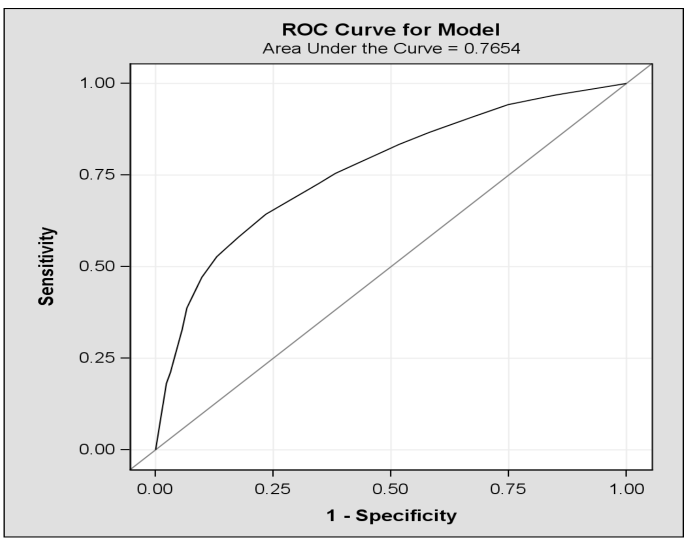


Figure 2. ROC curve for the fitted model

6. We plotted an ROC curve, given in the Figure 2, in order to asses the prediction power of our fitted logistic regression model. The area under the curve was found to be 0.7654, which shows that our model provides a better predictions than random guessing [11].
7. Since the interaction term parameter Game*Basketball was found to be insignificant for the fitted model, one might attempt to fit a more simpler by considering the dummy variables for each multi-category variable (defined in the Section 2.1) as separate variables, and using them to fit logistic regression models via the model selection procedure described earlier. We were able to obtain a more simpler model having the form

$$\text{logit}(\widehat{\pi}) = \widehat{\beta}_0^* + \widehat{\beta}_1^* G + \widehat{\beta}_2^* S + \widehat{\beta}_3^* T_1 + \widehat{\beta}_4^* T_2 + \widehat{\beta}_5^* T_3 + \widehat{\beta}_6^* \left(S * T_1\right) + \widehat{\beta}_7^* \left(S * T_2\right).$$

For this model, we found P-value for the deviance test statistic to be equal to 0.0432. Therefore, using deviance test statistic, we also found that this simpler model exhibit (marginally) inadequate fit. The AIC (= 289.3261) for this model for lower than our fitted model and area under its ROC, being equal to 0.765, was also similar to our model.

References

1. "What is Concussion?" Brain Injury Association of America. Accessed: Dec. 06, 2017. http://www.biausa.org/concussion/whatisaconcussion
2. Collins M. W., Kontons A. P., and D. O. Okonkwo, et al. *Concussion is Treatable: Statements of Agreements from the Targeted Evaluation and Active Management(TEAM) Approaches to Treating Concussion Meeting held in Pittsburg, October 15-16, 2015.* Neurosurgery, 79(6): 912-929, 2016.
3. R. W. Evans *The postconcussion syndrome: 130 years of controversy.* Semin Neurol., 14: 32-39, 1994.
4. Gronwell D. M. A. and P. Wringtson. *Cumulative effect of concussion.* Lancet, 2: 99-997, 1975.
5. Covassin T., Swanik C. B., and M. L. Sachs. *Sex Differences and the Incidence of Concussions Among Collegiate Athletes*, Journal of Athletic Training, 38(3): 238-244, 2003.
6. "Concussion.'' American Association of Neurological Surgeons. Accessed: Dec. 06, 2017. http://www.aans.org/Patients/Neurosurgical-Conditions-and-Treatments/Concussion
7. Langois J. A., Rutland-Brown W. and M. M. Wald. *The epidemiology and impact of traumatic brain injury: a brief overview.* J Head Trauma Rehabil., 21: 375-378, 2006.
8. Barnes B. C., Cooper L., Kirkendall D. T., McDermott T. P., Jordan B. D. and E. E. Garrett Jr. *Concussion history in elite male and female soccer players.* Am J Sports Med. 26: 433-438, , 1998.
9. Boden B. P., Kirkendall D., and W. E. Garrett Jr. *Concussion incidence in elite college soccer players.* Am J Sports Med., 26:238-241, 1998.
10. National Collegiate Athletic Association.. *NCAA Injury Surveillance System for Academic Years 1997 – 2000.* Indianapolis, IN: National Collegiate Athletic Association; 2000.
11. Agresti, A. *Categorical Data Analysis.* John Wiley & Sons, Inc. 3rd Edition, 2013.
12. van der Ploeg, T., Austin P. C. and E. W. Steyerberg. *Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints.* BMC medical research methodology, 14 (1): 137, 2014.

A. Interpretation of the estimated parameters in the fitted logistic regression model

1. Gender (G): *Controlling for game setting and all four types of sports, the estimated odds of sustaining concussion for the female athletes are* $\exp\left\{\widehat{\beta}_1\right\} = \exp\{0.3355\} = 1.3986$ *times the estimated odds of sustaining concussion for the male athletes.*
2. Game Setting (S): *Controlling for gender (G), the estimated odds of sustaining concussion during actual game for those athletes who play*
   (i) softball/ baseball are $\exp\left\{\widehat{\beta}_2\right\} = \exp\{1.2880\} = 3.6255$
   (ii) soccer are $\exp\left\{\widehat{\beta}_2 + \widehat{\beta}_6\right\} = \exp\{1.2880 + 1.5738\} = 17.4929$
   (iii) lacrosse are $\exp\left\{\widehat{\beta}_2 + \widehat{\beta}_7\right\} = \exp\{1.2880 + 0.7427\} = 7.6194$
   (iv) basketball are $\exp\left\{\widehat{\beta}_2 + \widehat{\beta}_8\right\} = \exp\{1.2880 - 0.0972\} = 3.2897$
   times the estimated odds of sustaining concussion for such athletes during practice session.
3. Soccer (T_1): *Controlling for gender, the estimated odds of sustaining concussion for those athletes who play soccer*
   (i) during practice session are $\exp\left\{\widehat{\beta}_3\right\} = \exp\{0.1371\} = 1.1469$
   (ii) during actual game are $\exp\left\{\widehat{\beta}_3 + \widehat{\beta}_6\right\} = \exp\{0.1371 + 1.5738\} = 5.5339$
   times the estimated odds of sustaining concussion for those athletes who play softball/ baseball.
4. Lacrosse (T_2): *Controlling for gender, the estimated odds of sustaining concussion for the athletes playing lacrosse*
   (i) during practice session are $\exp\left\{\widehat{\beta}_4\right\} = \exp\{0.7130\} = 2.0401$
   (ii) during actual game are $\exp\left\{\widehat{\beta}_4 + \widehat{\beta}_7\right\} = \exp\{0.7130 + 0.7427\} = 4.2874$
   times the estimated odds of sustaining concussion for those athletes who play softball/ baseball.

5. Basketball (T_3): *Controlling for sex, the estimated odds of sustaining concussion for the athletes playing basketball*
   (i) during practice session are $\exp\left\{\widehat{\beta}_5\right\} = \exp\{0.7835\} = 2.1891$
   (ii) during actual game are $\exp\left\{\widehat{\beta}_5 + \widehat{\beta}_8\right\} = \exp\{0.7835 - 0.0972\} = 1.9864$

   times the estimated odds of sustaining concussion for those athletes who play softball/ baseball.