REPEATED MEASURES ANALYSIS OF BINARY OUTCOMES: AN APPLICATION IN INJURY RESEARCH

by
Denise Svendsgaard Williamson
Department of Biostatistics
University of North Carolina

Institute of Statistics
Mimeo Series No. 2142T

December 1994

REPEATED MEASURES ANALYSIS OF BINARY OUTCOMES: AN APPLICATION IN INJURY RESEARCH

by

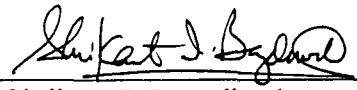
Denise Svendsgaard Williamson

A paper submitted to the faculty
of the University of North Carolina
in partial fulfillment of the requirements
for the degree of Master of Public Health
in the Department of Biostatistics

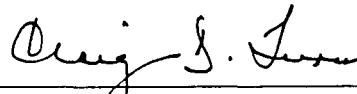
Chapel Hill

1994

Approved by:



Dr. Shrikant I. Bangdiwala



Dr. Craig D. Turnbull

REPEATED MEASURES ANALYSIS OF BINARY OUTCOMES: AN APPLICATION IN INJURY RESEARCH

by

Denise Svendsgaard Williamson

1. INTRODUCTION

Much of the focus of longitudinal data analysis concerns the time to occurrence of a unique event, for example, time to death or time to first heart attack. For repeated measures designs in which there is a multiply occurring event, there is a need for methods to model relationships between covariates and the event that is recurring. This is especially true for injury research. Because injuries are common in some populations, we often have multiply reoccurring events for one individual.

Complex methods for analyzing multiply recurring events such as injuries have rarely been applied to injury prevention research. This paper describes the application of existing methods for analyzing correlated data with binary outcomes to injury data in order to study the epidemiology of injuries.

The nature of the data, in which there are repeated measures for subjects, requires that the correlation among outcomes and covariates for a given subject be taken into account. The existence of both time-dependent and time-independent covariates must also be accommodated. A generalized estimating equation (GEE) approach, which takes into account these factors described above, is used to model the risk of injuries[7,10,11,12]. For comparison purposes, a sample survey method for modelling

ratios to incidence densities is applied to the data.[6] Also, the analysis of the data by traditional logistic regression methods, which do not account for the correlation described above, is discussed.

The specific dataset to which these methods will be applied is from a cohort study of rugby players, designed to examine the risk and protective factors for rugby injury. A description of the game of rugby is provided in Section 2. The study that collected the data is described in Section 3. Section 4 presents a descriptive analysis of the data. Section 5 describes the methods for analyzing the data. The results of the analyses are presented in Section 6. A discussion of other applications of this analysis and suggestions for future research is given in Section 7.

2. THE GAME OF RUGBY UNION

Rugby Union is a contact sport, typically played outdoors, on a grass field, or pitch. It is an amateur game played by two teams of 15 players, who are allowed to carry, kick, and throw the ball, which is similar to an American football. Players attempt to score points by placing the ball over the opponents' goal line (a try), or by kicking it over the crossbar (a goal). There are two halves of 40 minutes each.

Players may run with the ball and kick the ball in any direction, but may not throw or knock the ball toward the opponents' goal line. Team members play according to designated positions, generally described as forwards and backs. During a scrummage, or scrum, forwards of each side get into a formation against each other. Once the scrum is steady, the ball is put into the tunnel that is formed by the scrum. Forwards then push to gain an advantage until the ball becomes loose and is picked up by one team's backs, commencing an attack toward the opponent's goal line.

Rugby is a game with a high injury rate[9]. Protective gear is minimal, but many players use mouthguards[2]. Typical injuries include sprains, strains, and other soft tissue injuries[2]. Some players suffer multiple injuries to the same site[2]. Severe injuries are usually related to scrums and foul play.

Rugby is played in countries throughout the world. In particular, it is popular in Great Britain, France, Australia, South Africa, and New Zealand. In New Zealand, in fact, rugby is considered the national sport.

3. DESCRIPTION OF STUDY

The Rugby Injury and Performance Project (RIPP) is a prospective cohort study, conducted in Dunedin, New Zealand by the Injury Prevention Research Unit. The study is designed to evaluate the relationship of pre-season and within-season factors to rugby injuries. There have been few studies of this kind which follow both injured and non-injured players over a period of time, allowing the comparison of injuries to factors of interest. In New Zealand, rugby is a sport which has high rates of participation as well as injury[9]. This study was supported by a grant from New Zealand's Accident Rehabilitation and Compensation Insurance Corporation, who provided compensation for over 38,000 claims for sporting injuries in 1992[9].

Rugby union players, including both males and females, were recruited to participate in RIPP through their coaches and clubs prior to the 1993 club season. Baseline measurements were obtained for 350 players at the beginning of the season on factors such as whether the player had a current or chronic injury, whether the player had participated in off-season training, the player's average intake of alcohol, and demographic characteristics. Participants were telephoned weekly by a trained telephone interviewer assigned to the participant to conduct an interview regarding the participant's exposure to rugby and their injury outcomes for the previous week. Information collected includes how many games were played the previous week, whether any injuries were sustained during a team practice, game or in some other way, the level of stress the player experienced during the week, and whether the player was ill the previous week. Stress was measured by asking the player to rate how stressful the previous week was overall: not at all stressful, somewhat stressful, very stressful, or extremely stressful. If an injury caused the subject to either miss at least one team practice or one game or seek medical attention, additional information was collected on the specific injury(ies). Data were collected over 23 weeks of the regular club season, with most weeks achieving a response rate of 90-95%. Drops in the response rate (to 60-80%) coincided with a 3-week school holiday and a 1-week school holiday which occurred during the season[9].

4. DESCRIPTIVE ANALYSIS OF DATA

The distribution of the data with respect to gender is described in the Table 1. Although rugby is a growing sport for women in New Zealand, men account for the majority of rugby players and represent approximately three-quarters of the data collected.

Table 1. Distribution of Gender

Gender	Number of players	Percent of players
Male	258	73.7
Female	92	26.3
TOTAL	350	100.0

Of the 350 players with baseline measurements, weekly data is available for 343 of them due to 7 who were not followed due to various reasons, e.g. could not be contacted by telephone at work or school. The distribution of the 343 players by gender is 254 males (74.1%) and 89 females (25.9%).

Of the 5923 player-weeks of data collected, there was no baseline data available for 5 subjects for whom 84 weeks of data were collected; therefore, these 84 weeks of data were removed from the analysis. The number of player-weeks of data available for subjects with baseline data is 5839 player-weeks, 4462 for males and 1377 for females. Thus, there is an average of 17.6 (4462 weeks/254 males) weeks of follow-up for males and 15.5 (1377 weeks/89 females) weeks of follow-up for females.

The distribution of grades is detailed in Table 2. Grades are analogous to leagues in baseball, where Senior A would be the highest league for males. Although some players occasionally play in a different grade, most players play in the same grade for the entire season. The grade shown below is the grade most often played in by the player, the modal grade.

Table 2. Distribution of Grade by Gender

Gender	Grade	Number of players	Percent of gender
Male	Senior A	95	37.4
	Senior B/Senior Reserve	38	15.0
	Under 21	64	25.2
	Schoolboys	53	20.9
	Other	4	1.6
	TOTAL MALES	254	*100.1
Female	Women	66	74.2
	Schoolgirls	23	25.8
	TOTAL FEMALES	89	100.0

* does not sum to 100.0% due to rounding

Table 3 indicates, as would be expected, that as a player's exposure to rugby games increases, the number of rugby injuries increases.

Table 3. Distribution of Game Injuries by Number of Games Played: Overall and by Gender

	Overall			Males			Females		
Number of games played in a week	Number of player-weeks	Number of player-weeks w/ 1 or more game injuries	% of total player-weeks w/ game injuries	Number of player-weeks	Number of player-weeks w/ 1 or more game injuries	% of total player-weeks w/ game injuries	Number of player-weeks	Number of player-weeks w/ 1 or more game injuries	% of total player-weeks w/ game injuries
0	1959	0	0.0	1369	0	0.0	590	0	0.0
1	3497	685	19.6	2799	563	20.1	698	122	17.5
2	318	69	21.7	244	53	21.7	74	16	21.6
3	46	11	23.9	34	9	26.5	12	2	16.7
4	17	5	29.4	15	5	33.3	2	0	0.0
5	1	1	100.0	1	1	100.0	0	0	N/A
6	1	0	0.0	0	0	N/A	1	0	0.0

From the table above, one could calculate an incidence density, i.e. the number of (weeks of) new game injuries per week exposed (i.e. per week played in one or more games). There were 771 weeks in which a player incurred one or more injuries in a game. Weeks in which a player played in one or more games total 3880. The incidence density is computed by dividing 771 by 3880, which is 0.20 weeks with game injuries per game week. The cumulative incidence of game injuries is approximately equal to the product of the incidence density times the length of the corresponding time period[3]. The cumulative incidence for an entire season, assuming 11.3 game weeks in a season (the average number of weeks a player played in one or more games in the RIPP data is 11.3), would be 11.3 x .20, or a total of 2.2 weeks with game injuries over an entire season. The cumulative incidence of game injuries computed for males is somewhat higher than for females, partly due to a higher incidence density for males, but also due to the fact that males play one or more games an average of 12.2 weeks and females play an average of 8.8 weeks. For males, the cumulative incidence of game injuries is 2.5 weeks with game injuries per season; for females, the cumulative incidence is 1.6 game injuries per season.

Table 4 shows the proportion of player-weeks for whom a positive response was obtained on covariates for males, females, and combined males and females. These covariates are relevant to the models discussed in subsequent sections. Factors appearing in Table 4 are time-dependent; that is, they

may vary from week to week for an individual. Stress was dichotomized based on a subject's response: not stressful if the player indicated the previous week was not at all stressful, stressful if the player rated the previous week as somewhat, very, or extremely stressful.

Table 4. Proportion of Player-Weeks with a Positive Response, by Time-Dependent Covariates

Group	Injured (in team practice, game, or other)	Stress	Injured in Game	Play one or more games	Miss game or team practice due to injury	Total Player-Weeks
Males	.20 (913/4461)	.30 (1349/4458)	.20* (631/3093)	.69 (3093/4462)	.12** (520/4439)	4462
Females	.15 (210/1377)	.34 (471/1375)	.18* (140/787)	.57 (787/1377)	.09** (118/1365)	1377
Overall	.19 (1123/5838)	.31 (1820/5833)	.20* (771/3880)	.66 (3880/5839)	.11** (638/5804)	5839

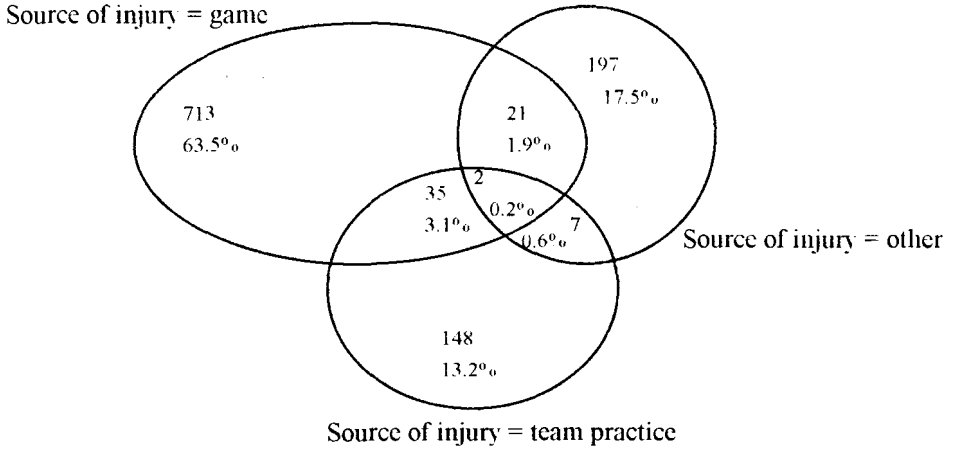
* Includes only those players who played in one or more games

** Does not include players who indicate not-applicable

There are a few player-weeks with missing data for some of the covariates indicated in Table 4. These weeks were not included in the calculation of the proportion of player-weeks with a positive response. The number of player-weeks these proportions are based on are included in parentheses under the proportions in the table.

Most of the player-weeks in which there was an injury (either in team practice, a game, or otherwise) involved a game injury (771/1123 = 69%). Injuries occurring during team practice accounted for 17% of all injury-weeks (192/1123) and injuries from other sources accounted for 20% of all injury-weeks (227/1123). These percentages exceed 100% since individuals can have injuries due to multiple combinations, e.g. an individual could be injured during team practice and during a game in a given week. Figure 1 describes the distribution of the different sources of injuries. The numbers in the figure indicate the number of player-weeks there were injuries accounted for by the source of injury (game, team practice, or other). The percentages indicate the percent of all injury-weeks accounted for by the source of injury.

Figure 1. Distribution of Sources of Injuries



Total Number of Injuries = 1123

Table 5 shows the proportion of individuals for whom a positive response was obtained on covariates for males, females, and combined males and females. These covariates are of relevance to models discussed in later sections. Factors appearing in Table 5 are time-independent; that is, they are constant for each individual. Although some players occasionally play in a different position, most players play in the same position for an entire season. The position shown in Table 5 is based on the position most often played.

Table 5. Proportion of Individuals with a Positive Response, by Baseline Covariates (Time-Independent)

Group	Participated in Off-season training	Current or Chronic Injury at start of season	Percent of Forwards (Position)	Total Individuals
Males	.80 (202/254)	.48 (122/253)	.53 (134/254)	254
Females	.56 (50/89)	.23 (19/83)	.55 (48/88)	89
Overall	.73 (252/343)	.42 (141/336)	.53 (182/342)	343

There are a few individuals with missing data for some of the covariates indicated in Table 5. These individuals were not included in the calculation of the proportion of individuals with a positive response. The number of individuals these proportions are based on are included in parentheses under the proportions in the table.

One of the aims of the study was to explore the relationship between pre-season and within-season factors in terms of rugby injuries. The following section discusses methods for modelling these relationships. Due to differences between males and females, all models discussed in subsequent sections are fit separately for males and females. Since the level of play varies from grade to grade, grade is controlled for where possible.

In subsequent sections, an understanding of the structure of the RIPP dataset will be useful. There is one observation, or record, of data per subject per week. Each observation has a variable which identifies the subject and a variable which indicates the week number of the study. If a subject could not be contacted in a given week, there is not an observation for that subject-week.

5. METHODS

As previously discussed, three methods were used to analyze this data. These methods were generalized estimating equations (GEEs), a survey sample method, and logistic regression. These methods are outlined below.

5.1 Generalized Estimating Equations

Generalized estimating equations is an approach for modelling longitudinal data for a general class of outcome variables, including Gaussian, Poisson, Binary, and Gamma outcomes[5]. GEEs take into account the correlation among outcomes for a given subject. Taking this correlation into account leads to increased efficiency relative to some "naive" estimators[7], such as those which assume that repeated observations from a subject are independent of one another. Both time-dependent and time-independent covariates can be accommodated. Strength is derived across subjects to estimate a "working" correlation matrix and hence account for the time dependence. The term "working" correlation matrix is used because it is not expected to be correctly specified, but consistent estimators of the regression coefficients and of their variances are available under the weak assumption that a weighted average of the estimated correlation matrices converge to a fixed matrix and under weak assumptions about the actual correlation among a subject's observations. It is not necessary that the observations for all subjects have the same correlation structure[7,10,11,12].

Unless the true correlation matrix is specified, a requirement of GEEs is that missing data are missing completely at random (MCAR). That is, whether an observation is missing cannot depend on previous outcomes. For this study, contact with each subject was attempted each week during the study, regardless of whether the subject played in a game or not (and regardless of previous outcomes). Many subjects were students, and two school holidays (one 3-week and one 1-week) fell during the 23 weeks of follow-up. During these school holidays, response rates dropped from about 90-95% to 60-80%. Clearly, missing data as a result of these school holidays would not be considered MCAR, but it is probably unrelated to the variables of interest and therefore would not be considered a threat to inferences using GEEs. Missing data for the RIPP dataset may be informative in some rare situations, e.g. a person could be hospitalized for an injury and not be contacted in a given week.

To implement the GEE approach, a known function of the marginal expectation of the outcome is specified to be a linear function of the covariates, and the variance is assumed to be a known function of the mean. Also, a "working" correlation matrix for the observations for each subject is specified. Regression coefficients are estimated by iteratively reweighted least squares. A SAS macro was used for all GEE models[5].

The link function needs to be correctly specified to make consistent inferences regarding regression coefficients. Choices for the link function include the identity link, the logarithm link, the logit link, and the reciprocal link. For the RIPP data, the logit link function was used as the appropriate link for all GEE models in order to parallel logistic regression.

The mean-variance relationship must also be specified. Choices include Gaussian, Poisson, binary, and gamma. For this analysis, since the variables of interest are dichotomized into binary format, the binary relationship was chosen for all GEE models.

Several choices are available for the working correlation matrix. One choice is to assume an identity matrix; that is, that repeated observations are uncorrelated. Another choice is an exchangeable correlation matrix; that is, that repeated observations for a subject are equally correlated. For this data, the choice of an autoregressive working correlation was made; it was assumed that the correlation between repeated observations decreases as a function of the time between two observations. Although estimates of

regression coefficients and their variances are robust to the choice of the working correlation matrix, choosing the working correlation matrix close to the actual one increases efficiency of the estimators relative to "naive" estimators.

One must also specify a variable which identifies each cluster. As indicated previously, for the RIPP data there exists an ID variable for each subject, which is part of each subject's weekly record. Naturally, the outcome variable and the covariates must also be specified.

Observations with data missing for any of the variables used in the model are deleted prior to running the GEE SAS macro. That is, only the week(s) a subject has missing data are deleted; weeks for which complete data is available for a subject are retained. However, subjects for whom there is only one complete observation (one week of data) are deleted since an autoregressive working correlation is assumed, and the dependence in correlation structure is 1, which must always be less than the minimum cluster size. The number of subjects deleted due to having only one complete observation (week) of data varies according to the model fit, but ranges from 0 to 5 subjects.

There are a few disadvantages of the GEE method. First of all, the regression model is assumed to be correct. Currently, there are no diagnostics available for checking this assumption. Also, specific sample size requirements are not specified in the literature, but problems with convergence develop with small sample sizes and/or strong collinearity problems. The number of subjects (clusters), as well as the sizes of the clusters, must also be considered. In fact, with the RIPP data, since separate models were fit for males and females, and females are about one-third the size of the male data, problems with convergence occurred for some of the models for females. This problem was thought to be due to the school girls who had an erratic playing schedule during the period of follow-up. All female models were run with school girls removed which resolved the problem with convergence. Therefore, the results presented in Section 6 do not include schoolgirls in order to have comparability across all models presented.

5.2 Survey Sample Method

Methods exist from sample survey techniques for producing direct estimates of adjusted incidence density ratios for risk factors of interest[6]. Incidence densities are calculated as the ratio of the number of events (e.g. game injuries) to the time at risk (e.g. weeks played in one or more games). They typically form the basis for comparisons among subgroups. Ratio estimation methods can be applied for analyzing the incidence of injuries. The application of this method has the advantage of avoiding the necessity of specifying the exact nature of the underlying variable distributions. However, adequate sample sizes are required so that sums comprising the numerator and denominator of the ratio are approximately normally distributed for the factor of interest. These methods have previously been used in sample surveys, where estimates of rates or proportions for population subgroups are desired, and large sample sizes support the analysis.

This method has been applied elsewhere to a dataset with features similar to RIPP data[6]. In the previous application of this method, there were repeat occurrences of an outcome for a given subject, at-risk periods were subject to random variability, and there were covariates that varied over the course of follow-up.

In this approach, incidence densities are first estimated for each cell of a cross-classification of the covariates. The corresponding estimated covariance matrix of the cell-specific ratio estimates is computed via a first-order Taylor series approximation of the variance of the estimates about their expected values[6]. Linear models can then be fit to the logarithms of the ratio estimates by the application of weighted least squares methods in order to estimate the simultaneous effects of the covariates. Although this method is usually used in sample surveys ,where variances are computed according to complex sampling designs, this method is applicable to simple random sampling from a population with replacement.

The SUDAAN software package was used to implement this analysis method[8]. Estimates of cell-specific ratios and the corresponding covariance matrix were estimated from PROC RATIO in SUDAAN. SAS was then used to fit linear models to the ratios via weighted least squares.

Advantages of the sample survey method are that the full covariance matrix is used as opposed to a working approximation of the covariance matrix in GEEs. Also, no assumptions are needed regarding the structure of the correlations among repeated observations on a subject. Disadvantages of this method include the inability to include many covariates due to minimum cell size requirements for cross-classifications of covariates. The required assumption that the sample was selected from a corresponding population with replacement following a probability sampling scheme could also be considered a disadvantage.

The assumption of simple random sampling from a population with replacement may not be reasonable for the RIPP data. Rugby players were recruited for the study, and there could be selection bias if players volunteered to participate based on their injury status or risk of injury. Also, the application of this method to the RIPP data meant that no more than three covariates could be modelled, and only for males due to minimum cell size recommendations.

5.3 Logistic Regression

Logistic regression methods are widely used, and there is much literature describing this method[4]. The method of logistic regression assumes observations are independent. In the RIPP dataset, this assumption clearly cannot be made since there are repeated observations on a subject and they may be correlated. However, the application of this method is described in order to determine how conclusions would differ from the GEE approach.

Logistic regression was applied to this data using SAS PROC LOGISTIC. SAS automatically removes any observations with missing model variables from analysis.

Disadvantages of this method are that the assumption of independence cannot be ignored. The correlation among outcomes on a given subject cannot be accounted for using logistic regression. Also, the application of logistic regression assumes the sample is representative of a corresponding larger target population. Advantages are that there are established methods for assessing the fit of the model and providing diagnostics on the model for logistic regression and a wide variety of software is available with logistic regression capabilities.

6. RESULTS

Several models are described in this section. As stated earlier, RIPP was designed to examine the risk and protective factors of injury. From a public health perspective, the interest would be in modelling the likelihood of injury. A related outcome of interest, which may be of importance to a coach, might be the probability of a particular player playing a game in a given week. Another outcome, which would combine these two outcomes, would be the probability of missing a game due to a previous injury in a given week.

First, I will describe the five GEE models and provide the results. Then, I will discuss the model to which the sample survey method was applied, and present the results of this analysis. Finally, I will present the results of the five models with logistic regression applied to them, and compare these results to the GEE results. All models were fit separately for males and females.

6.1 GEE Model Results

The generalized estimating equation (GEE) approach to logistic regression was used to fit the following five models. As stated earlier, the female models do not include school girls due to their erratic playing schedule.

Model 1 estimates the likelihood of injury during a game in one week (week k) as a function of baseline (time-independent) covariates, including off-season training participation, injury status at pre-season, position, and grade, and covariates in week k-1 and week k. Covariates in week k-1 include whether the player was injured during a game, team practice, or in some other way, whether the player was ill, and whether the player rated himself/herself as stressed during that week (where the player is considered stressed if he/she rated themselves as somewhat, very, or extremely stressed). Week k covariates include whether the player was ill and whether the player was under stress. Figure 2 provides a description of the outcome variable and covariates included in this model.

Figure 2. Description of Model 1

Outcome: INJURED IN GAME (1 - injured during a game in week k/0 - not injured in game in week k)

Predictors:	**Week k - 1**	**Week k**	**Baseline (Time-independent)**
	Injured, k-1 (1-yes 0-no)	Ill, k (1-yes 0-no)	Off-season Trng (1-yes 0-no)
	Ill, k-1 (1-yes 0-no)	Stress, k (1-yes 0-no)	Position (1-back 0-forward)
	Stress, k-1 (1-yes 0-no)		Current Chronic Injury (1-yes 0-no)
			Senior A (1 - Sr. A/0 - not Sr. A)*
			Senior B (1 - Sr. B 0 - not Sr. B)*
			Under 21(1-under21 0-not under 21)*
			School Boy (1-sch boy 0-not sch boy)*

* *grade indicators; for model applied to males only*

Table 6 provides results from Model 1. An intercept was included in the model, but to simplify the table, intercept results have not been included. For males, 246 clusters and 2697 weeks of data were read by the GEE macro. Cluster size ranged from 2 to 19 and the average cluster size was 11 weeks. For females, 58 clusters and 597 weeks of data were read. The range of cluster sizes were 2 to 16 and the average cluster size was 10 weeks. The output provided by the GEE include the regression coefficient, the naive standard error, and the robust standard error. The naive standard error is calculated on the assumption that the working correlation matrix is the true correlation matrix. The calculation of the robust standard error does not make this assumption. The odds ratios (OR) in the table below are the exponentiated regression coefficients. The 95% confidence intervals and the p-values (2-sided) are based on the robust standard error since I do not want to make the assumption that the working correlation is the true correlation.

Table 6. GEE Results for Model 1

	Males			Females		
Covariate	OR	95% CI	p-value	OR	95% CI	p-value
Injured, k-1	1.81	(1.44,2.28)	< 0.001	1.01	(0.54,1.90)	0.984
Ill, k-1	1.41	(1.09,1.82)	0.009	1.44	(0.83,2.52)	0.197
Stress, k-1	1.06	(0.86,1.30)	0.596	0.88	(0.53,1.44)	0.610
Ill, k	0.97	(0.73,1.27)	0.803	0.76	(0.38,1.52)	0.441
Stress, k	1.22	(0.97,1.54)	0.087	1.01	(0.65,1.57)	0.968
Off-season Training	1.17	(0.88,1.55)	0.276	1.21	(0.74,1.98)	0.441
Position	0.86	(0.69,1.08)	0.208	0.98	(0.61,1.57)	0.928
Current/Chronic Injury	1.18	(0.94,1.47)	0.153	0.98	(0.50,1.93)	0.944
Senior A	1.49	(0.47,4.75)	0.503			
Senior B	1.40	(0.43,4.58)	0.582			
Under 21	1.52	(0.47,4.91)	0.484			
School Boy	1.31	(0.40,4.31)	0.653			

These results indicate that, for males, injured in week k-1 and ill in week k-1 are strong predictors for whether an individual will be injured during a game in week k. Both covariates have an odds ratio greater than 1, indicating that an individual would be more likely to be injured in a game if either of these factors are present. Other covariates, although not significant predictors, have estimated odds ratios in the direction one would generally expect. For example, the odds ratio for stressed in week k is over 1, indicating one may be more likely to be injured in a game if one is under stress. Although none of the predictors for females are significant, the direction of the odds ratios are the same as males for most of the predictors.

Model 2 estimates the likelihood of playing in one week (week k) as a function of baseline (time-independent) covariates and covariates in week k-1 and week k. All covariates remain the same as those described for Model 1. Figure 3 provides a description of the outcome variable and covariates included in this model.

Figure 3. Description of Model 2

Outcome: PLAY (1 - play one or more games in week k/0 - not play a game in week k)

Predictors:	**Week k - 1**	**Week k**	**Baseline (Time-independent)**
	Injured, k-1 (1-yes/0-no)	Ill, k (1-yes/0-no)	Off-season Trng (1-yes/0-no)
	Ill, k-1 (1-yes/0-no)	Stress, k (1-yes/0-no)	Position (1-back/0-forward)
	Stress, k-1 (1-yes/0-no)		Current Chronic Injury (1-yes/0-no)
			Senior A (1 - Sr. A/0 - not Sr. A)*
			Senior B (1 - Sr. B/0 - not Sr. B)*
			Under 21(1-under21/0-not under 21)*
			School Boy (1-sch boy/0-not sch boy)*

* *grade indicators; for model applied to males only*

Table 7 provides results from Model 2. As with Model 1, an intercept was included in the model, but intercept results have not been included in the table. For males, 253 clusters and 3887 weeks of data were read by the GEE macro. Cluster size ranged from 2 to 24 and the average cluster size was 15 weeks. Note that although there were 23 weeks of follow-up during the regular club season, 8 players were followed for 24 or 25 weeks because they extended their season by changing grades at the end of the season. For females, 64 clusters and 975 weeks of data were read. The range of cluster sizes were 5 to 19 and the average cluster size was 15 weeks.

Table 7. GEE Results for Model 2

Covariate	Males OR	Males 95% CI	Males p-value	Females OR	Females 95% CI	Females p-value
Injured, k-1	0.65	(0.55,0.78)	< 0.001	0.61	(0.42,0.88)	0.009
Ill, k-1	1.10	(0.88,1.36)	0.401	0.99	(0.68,1.45)	0.968
Stress, k-1	0.87	(0.73,1.05)	0.144	1.09	(0.82,1.44)	0.569
Ill, k	0.72	(0.58,0.88)	0.002	0.57	(0.39,0.83)	0.003
Stress, k	1.16	(0.96,1.41)	0.128	1.28	(0.93,1.77)	0.126
Off-season Training	1.13	(0.77,1.65)	0.529	1.33	(0.73,2.43)	0.358
Position	0.86	(0.66,1.12)	0.254	0.81	(0.46,1.43)	0.472
Current/Chronic Injury	0.77	(0.59,1.01)	0.063	0.62	(0.33,1.16)	0.139
Senior A	1.90	(0.81,4.42)	0.139			
Senior B	1.18	(0.50,2.78)	0.704			
Under 21	1.30	(0.55,3.09)	0.548			
School Boy	1.04	(0.44,2.48)	0.920			

These results indicate that, for both males and females, injured in week k-1 and ill in week k are strong predictors for whether an individual will play a game in week k. Both covariates have an odds ratio of less than 1, indicating that an individual would be less likely to play if either of these factors are present. The odds ratio for current or chronic injury could be considered borderline significant (especially for males); since it is less than 1, it indicates one would be less likely to play in week k if one had a current or chronic injury. Other covariates, although not significant predictors, have estimated odds ratios in the direction one would expect. For example, the odds ratio for stressed in week k is over 1, indicating one is more likely to play if one is under stress. This may make sense if one considers that playing in a game may be a cause of stress. The odds ratio for off-season training is greater than 1, indicating that one is more likely to play in a game in week k if one participated in off-season training prior to the season. Odds ratios for grades Senior A through School Boy are suggestive of a trend. In fact, Senior A players do play more often than other grades as this trend would indicate. It is interesting to note that stress in week k-1 has an odds ratio in a different direction for males compared to females for Model 1 and Model 2, and is the only variable to differ in direction of the odds ratio between males and females in both models.

Model 3 is the same as Model 2, except it includes a covariate from week k-2. The likelihood of playing in one week (week k) is estimated as a function of baseline (time-independent) covariates and covariates in week k-2, week k-1, and week k. The only covariate included from week k-2 is whether a

player was injured in a game, team practice, or in some other way. Baseline covariates and covariates in week k-1 and week k remain the same as in previous models. Figure 4 provides a description of the outcome variable and covariates included in this model.

Figure 4. Description of Model 3

Outcome: PLAY (1 - play one or more games in week k/0 - not play a game in week k)

Predictors:	**Week k - 2**	**Week k - 1**	**Week k**	**Baseline (Time-independent)**
	Injured,k-2 (1-y 0-n)	Injured,k-1 (1-y 0-n)	Ill, k (1-y 0-n)	Off-season Trng (1-y 0-n)
		Ill, k-1 (1-y 0-n)	Stress, k (1-y 0-n)	Position (1-back 0-forward)
		Stress, k-1(1-y 0-n)		Current Chronic Injury (1-y 0-n)
				Senior A (1 - Sr. A 0 - not Sr. A)*
				Senior B (1 - Sr. B 0 - not Sr. B)*
				Under 21(1-under21 0-not under 21)*
				School Boy (1-sch boy 0-not sch boy)*

* *grade indicators; for model applied to males only*

Table 8 provides results from Model 3. As before, an intercept was included in the model, but intercept results have not been included in the table. For males, 249 clusters and 3410 weeks of data were read by the GEE macro. Cluster size ranged from 2 to 23 and the average cluster size was 14 weeks. For females, 64 clusters and 864 weeks of data were read. The range of cluster sizes were 2 to 18 and the average cluster size was 14 weeks. The reduction in weeks read from Model 2 to Model 3 was due to missing data; that is, when the covariate for week k-2 was calculated, if a player was not contacted in week k-2, there was no data for this covariate, so the observation was deleted.

Table 8. GEE Results for Model 3

	Males			Females		
Covariate	OR	95% CI	p-value	OR	95% CI	p-value
Injured, k-2	0.74	(0.62,0.88)	< 0.001	1.17	(0.77,1.76)	0.459
Injured, k-1	0.61	(0.51,0.74)	< 0.001	0.63	(0.42,0.96)	0.030
Ill, k-1	1.11	(0.87,1.40)	0.406	1.04	(0.69,1.58)	0.842
Stress, k-1	0.81	(0.66,0.99)	0.038	1.15	(0.85,1.55)	0.374
Ill, k	0.69	(0.55,0.86)	0.001	0.62	(0.43,0.92)	0.016
Stress, k	1.16	(0.94,1.43)	0.177	1.16	(0.84,1.59)	0.368
Off-season Training	1.14	(0.77,1.71)	0.509	1.32	(0.72,2.42)	0.379
Position	0.84	(0.63,1.11)	0.215	0.82	(0.45,1.48)	0.509
Current/Chronic Injury	0.80	(0.60,1.06)	0.116	0.72	(0.37,1.40)	0.332
Senior A	1.87	(0.79,4.42)	0.150			
Senior B	1.12	(0.47,2.70)	0.795			
Under 21	1.36	(0.57,3.26)	0.490			
School Boy	1.09	(0.46,2.61)	0.842			

Results for Model 3 are very similar to Model 2. The small changes in the estimated odds ratios for the covariates present in both models indicate collinearity is not likely to be a problem. As with Model 2, injured in week k-1 and ill in week k are strong predictors for whether an individual will play a game in week k for both males and females. For males, injured in week k-2 is also a strong predictor, though not as strong in terms of significance or in terms of the estimated odds ratio as injured in week k-1. This makes sense because as the time between weeks increases, one would expect covariates to have less impact on the outcome.

Model 4 is the same as Model 2, except the outcome is whether an individual missed a game or team practice in week k due to a previous injury. The likelihood of missing a game or team practice in one week (week k) is estimated as a function of baseline (time-independent) covariates and covariates in week k-1 and week k. These covariates are the same as those described in previous models. Figure 5 provides a description of the outcome variable and covariates included in this model.

Figure 5. Description of Model 4

Outcome: MISS (1 - miss a game or team practice in week k due to previous injury/0 - not miss)

Predictors:	**Week k - 1**	**Week k**	**Baseline (Time-independent)**
	Injured, k-1 (1-yes/0-no)	Ill, k (1-yes/0-no)	Off-season Trng (1-yes/0-no)
	Ill, k-1 (1-yes/0-no)	Stress, k (1-yes/0-no)	Position (1-back/0-forward)
	Stress, k-1 (1-yes/0-no)		Current Chronic Injury (1-yes/0-no)
			Senior A (1 - Sr. A/0 - not Sr. A)*
			Senior B (1 - Sr. B/0 - not Sr. B)*
			Under 21(1-under21/0-not under 21)*
			School Boy (1-sch boy/0-not sch boy)*

* *grade indicators; for model applied to males only*

Table 9 provides results from Model 4. As before, an intercept was included in the model, but intercept results have not been included in the table. For males, 253 clusters and 3867 weeks of data were read by the GEE macro. Cluster size ranged from 2 to 24 and the average cluster size was 15 weeks. For females, 64 clusters and 969 weeks of data were read. The range of cluster sizes were 5 to 19 and the average cluster size was 15 weeks.

Table 9. GEE Results for Model 4

Covariate	Males OR	Males 95% CI	Males p-value	Females OR	Females 95% CI	Females p-value
Injured, k-1	3.62	(2.84,4.60)	< 0.001	4.28	(2.74,6.69)	< 0.001
Ill, k-1	0.81	(0.59,1.12)	0.197	0.90	(0.59,1.39)	0.653
Stress, k-1	1.17	(0.92,1.49)	0.211	1.07	(0.78,1.47)	0.660
Ill, k	0.92	(0.69,1.21)	0.535	0.79	(0.51,1.22)	0.280
Stress, k	1.17	(0.94,1.46)	0.171	1.03	(0.66,1.60)	0.904
Off-season Training	0.74	(0.40,1.36)	0.332	2.02	(0.84,4.87)	0.116
Position	1.59	(1.06,2.40)	0.026	0.87	(0.38,2.02)	0.749
Current/Chronic Injury	1.36	(0.90,2.06)	0.150	3.57	(1.49,8.55)	0.004
Senior A	0.52	(0.09,3.10)	0.472			
Senior B	0.32	(0.05,2.07)	0.234			
Under 21	0.38	(0.06,2.43)	0.308			
School Boy	0.25	(0.04,1.61)	0.144			

As one would hypothesize, injured in week k-1 is a strong predictor for whether an individual will miss a game or team practice in week k due to previous injury. A player is much more likely to miss a game or team practice in week k due to a previous injury if the player was injured in week k-1. For males, position would also appear to be a strong predictor. Backs are more likely than forwards to miss a game or team practice due to a previous injury. A possible reason for this may be that backs, who typically are the players running with the ball, are less able to play with an injury than forwards. However, this pattern is not found for females. These results also indicate that players with a current or chronic injury at the start of the season are more likely to miss a game or team practice in week k due to a previous injury, especially females who are almost 4 times as likely to miss a game or team practice if they had a current or chronic injury at the start of the season. This extreme (and significant) odds ratio for females may warrant further study.

Model 5 is the same as Model 4, except it includes a covariate from week k-2. The likelihood of missing a game or team practice in one week (week k) due to a previous injury is estimated as a function of baseline (time-independent) covariates and covariates in week k-2, week k-1, and week k. These covariates are the same as described in Model 3. A description of the outcome variable and covariates included in this model are provided in Figure 6.

Figure 6. Description of Model 5

Outcome: MISS (1 - miss a game or team practice in week k due to previous injury/0 - not miss)

Predictors:	**Week k - 2**	**Week k - 1**	**Week k**	**Baseline (Time-independent)**
	Injured.k-2 (1-y 0-n)	Injured.k-1 (1-y 0-n)	Ill. k (1-y 0-n)	Off-season Trng (1-y 0-n)
		Ill. k-1 (1-y 0-n)	Stress. k (1-y 0-n)	Position (1-back 0-forward)
		Stress. k-1(1-y 0-n)		Current Chronic Injury (1-y 0-n)
				Senior A (1 - Sr. A 0 - not Sr. A)*
				Senior B (1 - Sr. B 0 - not Sr. B)*
				Under 21(1-under21 0-not under 21)*
				School Boy (1-sch boy 0-not sch boy)*

* *grade indicators; for model applied to males only*

Table 10 provides results from Model 5. As before, an intercept was included in the model, but intercept results have not been included in the table. For males, 249 clusters and 3391 weeks of data were read by the GEE macro. Cluster size ranged from 2 to 23 and the average cluster size was 14 weeks. For females, 64 clusters and 860 weeks of data were read. The range of cluster sizes were 2 to 18 and the average cluster size was 13 weeks.

Table 10. GEE Results for Model 5

	Males			Females		
Covariate	OR	95% CI	p-value	OR	95% CI	p-value
Injured, k-2	1.42	(1.14,1.78)	0.002	1.76	(1.07,2.88)	0.025
Injured, k-1	3.95	(3.07,5.07)	< 0.001	5.20	(3.13,8.61)	< 0.001
Ill, k-1	0.80	(0.58,1.11)	0.180	0.89	(0.54,1.46)	0.646
Stress, k-1	1.21	(0.95,1.54)	0.128	0.98	(0.70,1.38)	0.928
Ill, k	1.02	(0.77,1.35)	0.881	0.78	(0.43,1.42)	0.424
Stress, k	1.16	(0.92,1.47)	0.200	1.06	(0.66,1.70)	0.810
Off-season Training	0.68	(0.35,1.30)	0.238	2.97	(1.07,8.24)	0.037
Position	1.69	(1.08,2.63)	0.021	0.66	(0.26,1.70)	0.390
Current/Chronic Injury	1.38	(0.89,2.14)	0.153	3.13	(1.22,8.03)	0.018
Senior A	0.50	(0.08,3.21)	0.465			
Senior B	0.32	(0.05,2.17)	0.242			
Under 21	0.36	(0.05,2.50)	0.303			
School Boy	0.23	(0.03,1.65)	0.144			

Results for Model 5 are quite similar to the results for Model 4. As before, the small changes in the estimated odds ratios for the covariates present in both models indicate collinearity is not likely to be a problem. As with Model 4, injured in week k-1 is a strong predictor for whether an individual will miss a game or team practice in week k due to a previous injury for both males and females. Injured in week k-2 is also a strong predictor for this outcome for both males and females, though not as strong in terms of significance or in terms of the estimated odds ratio as injured in week k-1.

6.2 Sample Survey Method Results

Due to sample size requirements for the sample survey method[6], only males were modelled. Incidence density of game injuries was defined as number of weeks with game injuries per weeks a player played in one or more games. The cell-specific sample sizes in the injury in week k-1 by ill in week k-1 by stress in week k were adequate for supporting the large-sample approximations this method requires. Estimated ratios and their covariance matrix were computed for the cross classification of the covariates described above. A model was then fit to estimate the simultaneous effects of the covariates. Results provided include the estimated regression coefficient and its standard error and p-value, and the estimated incidence density ratio (IDR) and its 95% confidence interval. This method also provides a goodness of fit test. Cell sizes and estimated ratios are provided for each cross-classification of covariates in Table 11. Results of this analysis are provided in Table 12.

Table 11. Cell sizes and Estimated Ratios for Each Cross-Classification of Covariates

Injury, k-1	Ill, k-1	Stress, k	Number of Players (1)	Number of Player-Weeks with Game Injuries (2)	Number of Player-Weeks a player played in one or More Games (3)	Estimated Ratio (Incidence Density) (2) / (3)
y	y	y	33	13	31	0.42
y	y	n	64	24	65	0.37
y	n	y	125	54	145	0.37
y	n	n	193	79	284	0.28
n	y	y	84	24	94	0.26
n	y	n	151	39	195	0.20
n	n	y	202	96	539	0.18
n	n	n	244	211	1374	0.15

Table 12. Results for Survey Sample Method

Effect	Beta	SE	p-value	IDR	95% CI
Intercept	-1.90	0.07	< .001	----	
Injury, k-1	0.64	0.08	< .001	1.90	(1.62, 2.23)
Ill, k-1	0.27	0.09	.003	1.32	(1.10, 1.58)
Stress, k	0.22	0.08	.006	1.25	(1.07, 1.46)

Goodness of fit chi-square (4 d.f.) = 0.74 $p = 0.95$

Odds ratios can be used to estimate incidence density ratios under certain circumstances[3]. Model 1 most closely parallels the outcome for the survey sample method. For injury in week k-1, the estimated OR is 1.81 (1.44, 2.28); for ill in week k-1, the estimated OR is 1.41 (1.09, 1.82); and for stress in week k, the estimated OR is 1.22 (0.97,1.54). These estimates are similar to the estimated IDRs produced by the sample survey method.

6.1 Logistic Regression Results

Logistic regression was applied to the 5 GEE models discussed previously. Their results are presented in Tables 13 - 17.

Table 13. Logistic Regression Results for Model 1

	Males			Females		
Covariate	OR	95% CI	p-value	OR	95% CI	p-value
Injured, k-1	2.22	(1.79,2.76)	< 0.001	1.76	(1.02,3.02)	0.041
Ill, k-1	1.40	(1.08,1.83)	0.012	1.35	(0.75,2.42)	0.313
Stress, k-1	1.04	(0.83,1.30)	0.746	0.94	(0.58,1.52)	0.804
Ill, k	1.01	(0.75,1.36)	0.942	0.84	(0.42,1.68)	0.615
Stress, k	1.24	(0.99,1.55)	0.056	0.92	(0.57,1.48)	0.743
Off-season Training	1.16	(0.88,1.52)	0.301	1.19	(0.74,1.91)	0.476
Position	0.88	(0.73,1.07)	0.216	0.99	(0.64,1.55)	0.976
Current/Chronic Injury	1.18	(0.97,1.44)	0.090	1.00	(0.58,1.72)	0.993
Senior A	1.50	(0.66,3.41)	0.338			
Senior B	1.40	(0.60,3.27)	0.441			
Under 21	1.54	(0.66,3.57)	0.316			
School Boy	1.34	(0.58,3.13)	0.494			

Table 14. Logistic Regression Results for Model 2

	Males			Females		
Covariate	OR	95% CI	p-value	OR	95% CI	p-value
Injured, k-1	0.76	(0.65,0.90)	0.002	0.88	(0.61,1.27)	0.506
Ill, k-1	1.10	(0.90,1.35)	0.353	1.06	(0.73,1.56)	0.747
Stress, k-1	0.82	(0.69,0.96)	0.014	1.05	(0.78,1.41)	0.745
Ill, k	0.71	(0.58,0.87)	< 0.001	0.56	(0.39,0.82)	0.003
Stress, k	1.12	(0.95,1.32)	0.185	1.39	(1.04,1.87)	0.028
Off-season Training	1.14	(0.95,1.36)	0.166	1.36	(1.03,1.79)	0.031
Position	0.86	(0.75,0.99)	0.037	0.80	(0.61,1.04)	0.096
Current/Chronic Injury	0.76	(0.66,0.88)	< 0.001	0.62	(0.46,0.84)	0.002
Senior A	1.83	(1.13,2.97)	0.013			
Senior B	1.12	(0.68,1.84)	0.650			
Under 21	1.25	(0.76,2.05)	0.374			
School Boy	1.03	(0.63,1.68)	0.907			

Table 15. Logistic Regression Results for Model 3

Covariate	Males OR	Males 95% CI	Males p-value	Females OR	Females 95% CI	Females p-value
Injured, k-2	0.73	(0.61,0.87)	< 0.001	1.42	(0.95,2.13)	0.090
Injured, k-1	0.76	(0.64,0.92)	0.004	0.88	(0.60,1.31)	0.533
Ill, k-1	1.11	(0.88,1.38)	0.376	1.13	(0.74,1.72)	0.569
Stress, k-1	0.76	(0.64,0.91)	0.002	1.16	(0.85,1.59)	0.347
Ill, k	0.70	(0.56,0.87)	0.002	0.59	(0.40,0.89)	0.012
Stress, k	1.11	(0.93,1.33)	0.234	1.25	(0.91,1.72)	0.160
Off-season Training	1.15	(0.94,1.39)	0.168	1.37	(1.02,1.83)	0.036
Position	0.85	(0.73,0.98)	0.029	0.81	(0.61,1.07)	0.140
Current/Chronic Injury	0.79	(0.68,0.92)	0.002	0.68	(0.49,0.94)	0.021
Senior A	1.81	(1.10,2.98)	0.019			
Senior B	1.09	(0.65,1.82)	0.748			
Under 21	1.29	(0.77,2.15)	0.335			
School Boy	1.08	(0.65,1.80)	0.773			

Table 16. Logistic Regression Results for Model 4

Covariate	Males OR	Males 95% CI	Males p-value	Females OR	Females 95% CI	Females p-value
Injured, k-1	3.15	(2.55,3.89)	< 0.001	3.45	(2.16,5.52)	< 0.001
Ill, k-1	0.67	(0.49,0.93)	0.016	0.62	(0.31,1.21)	0.161
Stress, k-1	1.43	(1.14,1.80)	0.002	1.17	(0.75,1.84)	0.492
Ill, k	0.83	(0.61,1.14)	0.258	1.21	(0.66,2.23)	0.539
Stress, k	1.43	(1.14,1.79)	0.002	1.00	(0.64,1.58)	0.988
Off-season Training	0.66	(0.50,0.87)	0.003	1.80	(1.11,2.92)	0.018
Position	1.55	(1.27,1.91)	< 0.001	0.79	(0.51,1.22)	0.288
Current/Chronic Injury	1.56	(1.27,1.92)	< 0.001	3.54	(2.30,5.45)	< 0.001
Senior A	0.49	(0.27,0.88)	0.016			
Senior B	0.36	(0.19,0.68)	0.002			
Under 21	0.37	(0.20,0.69)	0.002			
School Boy	0.27	(0.15,0.51)	< 0.001			

Table 17. Logistic Regression Results for Model 5

Covariate	Males OR	Males 95% CI	Males p-value	Females OR	Females 95% CI	Females p-value
Injured, k-2	1.50	(1.18,1.91)	< 0.001	1.29	(0.74,2.26)	0.364
Injured, k-1	3.25	(2.59,4.08)	< 0.001	3.65	(2.24,5.95)	< 0.001
Ill, k-1	0.64	(0.45,0.91)	0.012	0.72	(0.36,1.44)	0.350
Stress, k-1	1.57	(1.23,2.00)	< 0.001	1.18	(0.74,1.89)	0.484
Ill, k	0.86	(0.62,1.20)	0.380	1.36	(0.73,2.52)	0.333
Stress, k	1.42	(1.11,1.80)	0.005	1.15	(0.72,1.83)	0.562
Off-season Training	0.61	(0.46,0.81)	< 0.001	1.99	(1.20,3.30)	0.008
Position	1.64	(1.32,2.04)	< 0.001	0.70	(0.44,1.11)	0.133
Current/Chronic Injury	1.59	(1.28,1.98)	< 0.001	3.30	(2.11,5.17)	< 0.001
Senior A	0.47	(0.26,0.86)	0.015			
Senior B	0.36	(0.19,0.69)	0.002			
Under 21	0.38	(0.20,0.72)	0.003			
School Boy	0.26	(0.14,0.50)	< 0.001			

Generally speaking, the estimated odds ratios produced by logistic regression are similar to the odds ratios produced by the GEE method. However, ignoring the correlation within subjects leads to incorrectly labelling certain covariates as significant predictors. Since logistic regression treats each observation as independent from other observations, the standard errors are underestimated, producing more significant results than the GEE method produced. In fact, if one were to run the GEE models under the assumption of independent errors and use the naive standard errors produced by the GEE method, the results would be identical to the logistic regression results.

7. DISCUSSION

I have described three approaches for the analysis of the RIPP data. The similar findings from the survey sample method and logistic regression support the application of generalized estimating equations for the RIPP data. The advantages of the GEE method, namely the ability to control for the correlation among outcomes for a given subject and the capacity to handle several covariates at a time, outweigh its disadvantages, which includes the lack of model fitting diagnostics.

To conclude, the analysis presented here could be applicable to other injury and sports injury data. Specific findings from the RIPP study could be considered for other similar contact sports, in particular American football and soccer. The methods applied to the analysis of the RIPP dataset could be appropriate for other, similarly structured, studies which are characterized by: (1) a repeated measures design, (2) binary outcomes which can reoccur multiple times for a given subject, and (3) time-dependent and time-independent covariates.

The RIPP dataset contains a wealth of information which should be considered for future research. For example, in the models presented here, the modal grade of play was used as a covariate. Most players tend to play in the same grade for the entire season . However, future research could investigate the effect of playing in a grade different from normal (especially grades higher than normal) to the risk of injury. Similarly, the modal position was used and was dichotomized into forwards and backs. Future research could investigate the risk of injury due to playing out of normal position. In addition, the inclusion of a measure of injury severity in the models could be considered. For example, the number of

weeks a player could not play due to injury would be one measure of injury severity. Another suggestion for future research would be to examine the relationship of stress to injury, perhaps by making stress an outcome variable for a model. Finally, the models presented in Section 6 attempted to quantify the relationship between the outcome of interest and covariates in the same week as well as covariates from week k-1 and week k-2. Future research could look into the relationship between the outcome and weeks k-3, k-4, etc.

Acknowledgments

I would like to thank Dr. Shrikant Bangdiwala for all of the time and support he gave me in the writing of this paper.

I would also like to thank the Injury Prevention Research Unit in Dunedin, New Zealand for the permission to use their data and the assistance from Stephen Marshall and Dr. Anna Waller in understanding and interpreting the data.

Many thanks are also due to my academic advisor, Dr. Craig Turnbull, for his support and encouragement.

Finally, I would like to thank my husband Carl for his assistance and motivation.

REFERENCES

1. The Diagram Group. (1974). *Rules of the Game: The Complete Illustrated Encyclopedia of all the Sports of the World,* New York: Paddington Press Ltd.

2. Gerrard, D.F., Waller, A.E., and Bird, Y.N. (*in press*). "The New Zealand Rugby Injury and Performance Project: II. Previous Injury Experience of a Rugby Playing Cohort," *British Journal of Sports Medicine.*

3. Hennekens, C.H. and Buring, J.E. (1987). *Epidemiology in Medicine,* Boston/Toronto: Little, Brown and Company.

4. Hosmer, D.W. and Lemeshow, S. (1989). *Applied Logistic Regression,* New York: John Wiley & Sons.

5. Karim, M.R. and Zeger, S.L. (1988). *GEE: A SAS Macro for Longitudinal Data Analysis,* (Version - 1), Technical Report #674, Department of Biostatistics, The Johns Hopkins University.

6. LaVange, L.M., Keyes, L.L., Koch, G.G., and Margolis, P.A. (1994). "Application of Sample Survey Methods for Modelling Ratios to Incidence Densities," *Statistics in Medicine,* Vol. 13, 343-355.

7. Liang, K.-Y. and Zeger, S.L. (1986). "Longitudinal Data Analysis using Generalized Linear Models," *Biometrika,* 73, 13-22.

8. Shah, B.V., Barnwell, B.G., Hunt, P.N., and LaVange, L.M. (1991). *SUDAAN User's Manual,* Release 5.50, Research Triangle Institute, Research Triangle Park, NC.

9. Waller, A.E., Freehan, M., Marshall, S.W., and Chalmers, D.J. (1994). "The New Zealand Rugby Injury and Performance Project: I. Design and Methodology of a Prospective Follow-Up Study," *British Journal of Sports Medicine,* 28(4) (*in press*).

10. Zeger, S.L. and Liang, K.-Y. (1986). "Longitudinal Data Analysis for Discrete and Continuous Outcomes," *Biometrics,* 42, 121-130.

11. Zeger, S.L. and Liang, K.-Y. (1992). "An Overview of Methods for the Analysis of Longitudinal Data," *Statistics in Medicine*, Vol. 11, 1825-1839.

12. Zeger, S.L., Liang, K.-Y., and Albert, P.S. (1988), "Models for Longitudinal Data: A Generalized Estimating Equation Approach," *Biometrics,* 44, 1049-1060.