Modeling Fault in Fatal Pedestrian Crashes by Using Various Data Sources

Lisa K. Spainhour and Isaac A. Wootton

Binary logistic regression was used to model fault in 318 fatal pedestrian cases that occurred in Florida in the year 2000. The results were used to classify fault and identify factors that influenced fault. An expert fault assessment served as a control for predicting fault in each crash. The expert assessment team conducted a case review of each traffic crash by using additional data sources, such as traffic homicide reports, diagrams, photographs, accident reconstructions, and site visit notes. The logistic models correctly classified fault in anywhere from 84% to 97% of the cases. The existing Florida Department of Transportation algorithm correctly classified fault in only 56% to 58% of the same cases. Improvements in classification accuracy were shown to stem from two sources: the abundance of the data and the improved accuracy of the data. The mental state of the pedestrian and the driver were shown to be important in determining fault. Exhibiting a mental aberration, such as inattention, distraction, perception or decision error, or intoxication, increased the propensity for fault. Issues such as the number of lanes attempted in a crossing, the age of an individual, being a former vehicle occupant, having limited conspicuity, receiving a citation, and wet roads were also shown to be factors significant in determining fault.

Florida's traffic fatalities in 2000 accounted for 10% of the national total, with a pedestrian fatality rate (the number of pedestrian fatalities per 100,000 resident population) that was the highest in the nation (*1*). Determination of the factors that contribute to pedestrian crashes and accurate assessment of fault are critical in evaluating the safety of a transportation network for pedestrians and for developing and selecting appropriate targeted countermeasures. When a collision between a pedestrian and a motor vehicle occurs, it is judicious to ascertain fault, whether it is that of a driver, a pedestrian, or another factor, such as a vehicle, roadway condition, or environmental condition. Although the determination of fault is valuable from a safety standpoint because it enables the development of directed safety campaigns and engineering countermeasures, the fatal crash data reported in Florida crash reports do not contain an explicit fault determination. Without a clear fault assessment, current techniques tend to place a high importance on failure to comply with state statutes in assessing fault and potentially diminish or even overlook other potential contributing factors, such as roadway or traffic factors. The research described here sought to better understand fault in fatal pedestrian crashes and the factors that influence the modeling of fault.

At present, limited Florida crash report data are available from state database extracts because they incorporate only the coded data from the Florida Traffic Crash Report and do not include data from the crash narrative or diagram. The data also contain errors introduced during the various data collection and transcription processes (*2*). Images of the original crash report narratives and diagrams can be reviewed manually; but it is a time-consuming process, and narratives are often lacking in detail, especially details on driver attitudes and actions, making it difficult to assign fault.

The Florida Department of Transportation (FDOT) currently uses a simple algorithm to assign fault. The model is based on a de facto convention by which officers are expected to place the at-fault driver or pedestrian in the first section of the crash report. FDOT thereby presumes that the individual in the first section is at fault, unless a citation was given to drivers or pedestrians in subsequent sections of the crash report, in which case fault is reassigned to the person receiving the citation. Two major flaws were found in this reasoning, especially when it is applied to the pedestrian crashes reviewed in this study. First, the often fatal injuries sustained by the pedestrian frequently defer the collection of information on the pedestrian until after the driver information has been collected, meaning that information on at-fault pedestrians is often not in the first section of the crash report. Second, in the event of a fatality, even if the pedestrian in the second section is at fault, he or she is never cited in the crash. Clearly, additional factors need to be considered when the fault in pedestrian crashes is predicted. As such, the objectives of this research were (*a*) to identify relevant data sources; (*b*) to conduct detailed case reviews of pedestrian crashes by using the most accurate data available; (*c*) to identify the true condition of fault in the crashes; (*d*) to build logistic regression models to predict the probability that the driver or pedestrian, or both, was at fault; and (*e*) to compare the fault state predicted by the regression models with the true state of fault and evaluate the predictive capability of the fault models.

Researchers have already modeled fault among motor vehicle–bicycle crashes (*3*), motor vehicle–motorcycle crashes (*4*), and motor vehicle–motor vehicle crashes (*5*); however, motor vehicle–pedestrian crashes remain unexplored. Previous studies in Hawaii relied on coded crash data collected by the police and on fault determined by the investigating officers and reported on crash report forms. In Hawaii, the indication of fault on crash reports appears to be similar to the convention used in Florida, as elaborated by Kim et al. (*5*). Although pedestrian cases were not addressed, the methods of determining fault are similar; and the prominence of human behavioral or risky factors, especially alcohol, in determining fault is common.

Department of Civil and Environmental Engineering, Florida A&M University–Florida State University College of Engineering, 2525 Pottsdamer Street, Tallahassee, FL 32310-6046. Corresponding author: L. K. Spainhour, spainhou@eng.fsu.edu.

Transportation Research Record: Journal of the Transportation Research Board, No. 2002, Transportation Research Board of the National Academies, Washington, D.C., 2007, pp. 64–71.
DOI: 10.3141/2002-08

DATA AND METHODS

A total of 318 fatal pedestrian cases that occurred in 2000 were used for this study. Two data sets were used for this research. The first, termed crash report data, came directly from FDOT databases. The data set was limited to coded data extracted from standard Florida Traffic Crash Reports with information collected by law enforcement agencies. The data were left in their unimproved native condition for comparative investigations. The second data set, termed case review data, stemmed from manual case reviews of multiple crash data sources. The data were collected by a diverse team of homicide investigators, researchers, traffic engineers, and safety engineers for the same 318 cases. A specific objective of the case reviews entailed examination of the underlying factors contributing to a crash, especially elements related to roadway design and traffic operations. A key source of information for the case reviews was detailed Traffic Homicide Investigation reports obtained from the Florida Highway Patrol and local law enforcement agencies. In addition, photographs of crash scenes from law enforcement agencies or from the state videolog system were carefully reviewed. When necessary, site visits and accident reconstructions were conducted. The team compared the data from these resources with the crash report and corrected any missing or erroneous data. Although no data source can be guaranteed to be accurate, the expert team used the preponderance of the evidence to determine the most likely circumstances of the crash. In many cases, data elements that were missing from the original crash reports were able to be added by using the augmented data resources. In addition, as part of the case review process, a manual assessment of fault was conducted for comparative purposes.

Binary logistic regression was the basis for predictive fault modeling in this study (*6–8*). According to Kim and Boski, logistic regression provides a powerful tool for measuring the association between fault and various demographic, vehicle, roadway, and environmental factors (*4*). Logistic regression models were used to identify the variables in a data set that were most significant for predicting fault and examine the strength of dependence. The process involved the fitting of terms associated with fault into a logistic model to predict the probability of fault. Stepwise logistic regression resulted in parsimonious models that were used to identify the significant factors influencing fault in pedestrian crashes. Driver fault and pedestrian fault were considered independently in this research, and separate models were developed for each.

Data transformations were used to encode data for logistic regression. Dichotomous (Boolean) indicator variables were found to be highly indicative and were used for many of the factors. Crash report data yielded a data set with 25 variables for investigation. A total of 27 variables were extracted from the case review data for investigation as potential predictors of fault. A concerted attempt was made to replicate as many variables as possible from the crash report data set and to use a consistent coding scheme whenever feasible. The coding definitions were kept simple to enable meaningful interpretation of logistic regression results.

As stated above, an expert assessment of fault was established during the case review. The predictions of the binary logistic regression models were evaluated against this expert or true assessment of fault. The fit between the true fault assessment and the assessment predicted by the various logistic regression models based on selected cutoff values is summarized in the results. A simple and straightforward measure of performance is the correctly classified percentage, which is computed as the number of accurately classified cases (true positives and true negatives) divided by the total number of cases classified. Two other measures of classification accuracy that are often reported in the literature are sensitivity and specificity. Sensitivity is the percentage of the target group accurately classified and is also known as the correct identification of true positives. Specificity refers to the percentage of the complementary group that is correctly classified and is also known as the correct identification of true negatives.

Model fit was explored and adjustments were made to limit and control the number of misclassifications, especially the assignment of fault to an innocent party. The predicted probabilities from the logistic models were used to assign group membership. Initially, if the predicted probability for a case was 0.50 or higher, also known as the 50% cutoff value, then the case was classified as a member of the target group. Cutoff values were adjusted on the basis of key performance measures of the classification analysis, namely, the percent classification accuracy, model sensitivity, and model specificity. The cutoff value was adjusted to ensure a minimum specificity of 95% for driver fault and 90% for pedestrian fault. This meant that no more than 5% and 10% of innocent drivers and pedestrians, respectively, were misclassified as being at fault.

MODEL DEVELOPMENT

Three techniques of fault assessment were considered in this study. The first method, the expert assessment, was designated the true fault condition and served as a control for other fault assessment schemes. The second method used the current fault assessment algorithm used by FDOT. The third method uses binary logistic regression to predict fault on the basis of the values of the various data about the crash. Several models were developed by this method by using either crash report data exclusively or case review data based on additional data resources.

Expert Assessment

The expert assessment refers to the fault determined by a multidisciplinary team that performed case study reviews. The expert assessment represents the true or actual condition of pedestrian or driver fault for this study. According to the expert assessment, as shown in Table 1, 83% of the pedestrians in fatal pedestrian accidents were at fault in the crashes. The expert fault assessment served as a control for the evaluation of pedestrian and driver fault modeling results and evaluation of the current FDOT algorithm. The use of a reliable control is essential for evaluation of the results of automated prediction or classification techniques; however, it is usually impractical and cost prohibitive to use such a control with a large number

TABLE 1 Expert Fault Assessment Results

Pedestrian at Fault	Driver at Fault		
	No	Yes	Total
No	12 (3.8%)	42 (13.2%)	54 (17.0%)
Yes	219 (68.9%)	45 (14.2%)	264 (83.0%)
Total	231 (72.6%)	87 (27.4%)	318 (100%)

of cases. Expert assessment was possible in this study because manual case reviews of the numerous data sources were conducted by a team of engineering and crash reconstruction specialists. To avoid potential biases, the team members were selected for their experience and objectivity and were trained in the evaluation of pedestrian crashes.

Current FDOT Algorithm

The current method used by FDOT to assess fault in pedestrian cases performs poorly when it is evaluated against the expert fault assessment. The FDOT algorithm does not predict fault correctly in nearly half of the cases, with correct classifications of only 56% for pedestrian fault and 58% for driver fault. For pedestrian fault, the main type of misclassification was a false negative, which occurred 138 times (43% of the cases). Error in the FDOT method of assessing fault can be attributed to a failure by the algorithm to classify a considerable number of at-fault pedestrians when a case review revealed that the pedestrian was indeed responsible. This problem is mirrored by driver fault, with the FDOT algorithm showing that drivers were too often classified as at fault when they were not, as exhibited by a high false-positive rate (112 occurrences, or 35% of the cases).

A false-positive result is considered a more grievous error than a false-negative result, in part because of the underlying commitment to presume innocence until guilt is proven and a corresponding desire not to implicate innocent parties. In the arena of law enforcement, this would be of paramount importance. Safety-related countermeasures are aimed at two different groups of people: those who are at fault and those who are not. It is desirable to distinguish clearly the characteristics of the at-fault group with a high level of discernment.

The current fault prediction generated by the FDOT algorithm performs poorly because it relies on the faulty assumption that crash data collection and reporting techniques are consistent. The low false-positive rate for pedestrian fault (three occurrences, or 1% of the cases) is commendable and dispels criticisms alleging a propensity by state agencies in Florida to hold pedestrians at fault in crashes. However, the lack of sensitivity in the current FDOT algorithm and the failure to detect at-fault pedestrians are genuine deficiencies. The main reason for the high false-negative rate is the dependence on the section number for classifying fault. As described above, some officers, unaware of the consequences, place information for a driver who is not at fault in section one of the crash report, thereby leading to a misclassification of fault by the FDOT algorithm. The information contained in the narrative of the crash report or in the homicide investigation documents, as extracted for this study, contained the case details necessary for correct fault assessment. Driver fault classification results support the same findings.

Binary Logistic Regression Models

To improve upon the fault assessment of the current FDOT algorithm, fault prediction models based on binary logistic regression were developed. Pedestrian fault was considered independently of driver fault. As such, two sets of models were created to make use of the data taken directly from crash reports. The structure of each model is as follows:

$$p(x) = \frac{e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}}{1 + e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}}$$

$$\text{logit}\left[p(x)\right] = \ln\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

where

p = probability of fault,
β_0 = constant term,
β_i = ith coefficient (logit), and
X_i = value of ith independent predictor variable.

Table 2 presents various goodness-of-fit statistics for the fault models and describes the accuracies of the predictive models compared with that of the expert assessment presented in Table 1. As summarized in Table 2, the first set of models (Models P1, P2, and P3) predicted fault among pedestrians, whereas the second set of models (Models D1, D2, and D3) predicted fault among the drivers striking pedestrians. For Models P1 and D1, crash report data were left in the native condition as obtained from FDOT database extracts, which has a number of missing and erroneous entries and does not include crash narrative information. Models P2 and D2 are based on the data from case reviews, which are augmented in both scope and accuracy, as described above. Models P3 and D3 used the exact same variables from the crash report data as initial Models P1 and D1, respectively; however, the data were taken from case reviews, meaning that missing and incorrect values were replaced whenever possible.

Originally, 30 variables were educed from the source data: five variables were used as identifiers, whereas the remaining 25 variables were investigated as potential predictors of fault. Variables that were significant (at the 90% confidence level) were identified by using forward and backward stepwise logistic regressions. As shown in Table 2, after stepwise variable elimination, the models contained nine or 10 variables, whereas the driver fault models contained six to 11 variables.

Each prediction model results in a probability that the person was at fault. A cutoff value of .5 maps probabilities above .5 to true (at fault) and those below .5 to false (not at fault). The classification cutoff value was adjusted from .5 to the value that maximized sensitivity (the percentage of the target group accurately classified) and specificity (the percentage of the other group accurately classified). Compared with the expert or true assessment of fault, Table 2 shows the number of false positives and false negatives with both the default cutoff value and the improved cutoff value (e.g., .84 for Model P1). It shows that Models P2 and D2, described in more detail below, classify fault correctly in the highest percentage of cases involving drivers and pedestrians, respectively.

The models that use crash report data only (Models P1 and D1) show that the use of supplementary FDOT data, even without any data accuracy improvements, is profitable for fault prediction. The primary benefit of the models that use crash report data is that they offer a significant increase in accuracy over that provided by the classification of the FDOT algorithm and require limited data handling. By relying on the unimproved data that are already housed in state databases, the feasibility of using more fields or data to predict fault is high and the costs and other barriers to implementation are kept low. Although dependence only on FDOT data is the primary benefit of these models, it is also a weakness because the data housed in the FDOT databases have been shown to be of reasonably poor quality (2, 9, 10). The lack of quality data leads to limitations in model capability. Although the false-positive rate was kept low by improving the model fit, the number of false negatives is large enough to warrant improvement efforts aimed at better detecting those who are actually at fault.

TABLE 2 Summary of Logistic Regression Fault Model Results

	Pedestrian Fault Models			Driver Fault Models		
	P1	P2	P3	D1	D2	D3
Model features						
Source of data	Crash report	Case review	Case review	Crash report	Case review	Case review
Independent variable selection criteria	Significant (*p*-value ≤ 0.10)	Significant (*p*-value ≤ 0.10)	Equivalent to those in model P1	Significant (*p*-value ≤ 0.10)	Significant (*p*-value ≤ 0.10)	Equivalent to those in model D1
Logistic regression model						
Number of model variables	10	9	10	6	11	6
Sig. variables with $p \le .05$	8	8	6	5	7	4
chi^2	146.54	220.33	162.65	127.77	252.19	215.70
Prob > chi^2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Log likelihood	−71.604	−34.710	−63.551	−122.715	−60.504	−78.751
Pseudo R^2	0.506	0.760	0.561	0.342	0.676	0.578
Summary stats for fitted model: default (0.5) cutoff						
False positives	23	7	17	16	8	11
False negatives	11	4	10	35	17	24
True positives	253	260	254	52	70	63
True negatives	31	47	37	215	223	220
Sensitivity	95.8%	98.5%	96.2%	59.8%	80.5%	72.4%
Specificity	57.4%	87.0%	68.5%	93.1%	96.5%	95.2%
Correctly classified	89.3%	96.5%	91.5%	84.0%	92.1%	89.0%
Explore nature of fit: lroc						
Area under ROC curve	0.9403	0.9844	0.9521	0.8655	0.9748	0.9568
Improve fit: set cutoff						
Cutoff	0.84	0.620	0.790	0.62	0.420	0.450
ROC area	0.8817	0.9423	0.9064	0.7499	0.9001	0.8785
Std. err.	0.0227	0.0204	0.0219	0.0278	0.0208	0.0225
Summary stats for final model: adjusted cutoff						
False positives	5	5	5	12	9	11
False negatives	38	6	25	39	14	17
True positives	226	258	239	48	73	70
True negatives	49	49	49	219	222	220
Sensitivity (%)	85.6%	97.7%	90.5%	55.2%	83.9%	80.5%
Specificity (%)	90.7%	90.7%	90.7%	94.8%	96.1%	95.2%
Correctly classified (%)	86.5%	96.5%	90.6%	84.0%	92.8%	91.2%

Sig. = significant; ROC = receiver operating characteristic; std. err. = standard error.

Fault prediction models based on data obtained through case study reviews (Models P2 and P3) were created to overcome the effects of the poor and erroneous data that limited the predictive capabilities of models based only on crash report data. Not only are the accuracies of case review data improved over those of crash report data, but also the case review data contain additional information and variables stemming from the detailed case study reviews. Independent variables were drawn from a high-quality data set containing information ranging from human factors to environmental and roadway factors.

Models that use data from case reviews benefit from a sizable increase in predictive capability over that possible with the existing FDOT algorithm and even from the considerable improvements in accuracy over those for the models that use crash report data. Model results can be used to ascertain variables that are significant for the determination of fault without being limited to the information coded on a crash report. The primary source of improvement was from the increases in model sensitivity over those of the crash report data models. Pedestrian fault Model P2 was superior to all other models in terms of classification accuracy (96.5%), sensitivity (98%), and specificity (91%). By using the pedestrian fault model, only 11 of 318 pedestrians were misclassified. Driver fault Model D2 was the next best model in terms of overall performance. A major drawback to the implementation of these models is the reliance on high-quality data, which are costly and time-consuming to obtain.

Another set of models was used to examine whether the source of improvements in predictive capability between models based on crash report data (Models P1 and D1) and those based on case review data

(Models P2 and D2) was the increased amount of data or the improved data quality. Models P3 and D3 used the exact same variables from the crash report data as initial Models P1 and D1, respectively; however, the data were taken from case reviews, meaning that missing and incorrect values were replaced whenever possible. The results of the classifications from the analogous fault models proved that improvements in data quality improve the ability to predict fault. Pedestrian fault Models P1 and P3 used the same variables, yet the model that used the case review data outperformed the model that used crash review data in all measures. The performance measure used to gauge model improvement was the proportion correctly classified, which increased from 86.5% to 90.6%. In the case of the analogous driver fault model, the findings more strongly suggest that the model based on corrected (case review) data (Model D3) outperforms the model based on crash review data (Model D1) in all measures. In both cases, the increase in model sensitivity means that higher-quality data help capture the at-fault individuals. However, neither model performs as well as the models based on full case study data (Models P2 and D2), which are both accurate and abundant.

RESULTS AND DISCUSSION OF RESULTS

The logistic regression fault models are discussed individually, starting with two primary pedestrian fault models (Models P1 and P2), followed by the two primary driver fault models (Models D1 and D2). The preferred model of the two depends on the intended usage. Models based on crash report data (Models P1 and D1) suggest which variables can be used to improve fault prediction in pedestrian cases when data are limited to those that are stored in the same format and that are of the same quality as the data collected by law enforcement agencies on crash reports. The models that use case review data (Models P2 and D2) reveal the factors that influence fault prediction the most when quality and quantity can be improved by manual case reviews of additional data sources. All models provide guidance on factors that are relevant for determining fault in fatal pedestrian crashes.

Pedestrian Fault Modeling by Using Data from Crash Reports

Table 3 displays the statistically significant variables in the model that uses crash report data to predict pedestrian fault (Model P1). Each row describes the variables that were found to be statistically significant after the stepwise regression. The odds ratio, which compares the odds of positive outcomes in a set of test cases with the odds of positive outcomes in a set of control cases, is an exponential function in logistic regression. For instance, an odds ratio of 12.704 ($e^{2.542}$) means that an intoxicated pedestrian is 12.704 times more likely to be at fault than a sober pedestrian. The z-statistic, p-value, and 95% confidence interval describe the confidence that can be placed in those results. For instance, a p-value of .075 (greater than .05) means that one cannot be 95% confident that driver gender is a significant variable in predicting pedestrian fault. The fact that the 95% confidence interval includes the value of 0 shows the same information. The data in Tables 4 through 6 are interpreted similarly.

Table 3 shows that several factors other than section number and citation information influence the determination of fault. Factors that lead to a higher probability of pedestrian fault are pedestrian intoxication, a male driver striking the pedestrian, low light conditions, and an increase in the speed limit. On the other hand, if a driver is intoxicated, cited, or assigned a contributing factor on the crash report, then the pedestrian is less likely to be at fault. Other factors that lead to a decrease in the chance that a pedestrian is at fault is when a crash is investigated by the Florida Highway Patrol, when there are wet roads, and when an officer assigns the pedestrian a section number higher than the one on the crash report form (the sections typically used for not-at-fault drivers or pedestrians).

Pedestrian intoxication is the most relevant variable influencing pedestrian fault. As stated above, an intoxicated pedestrian is almost 13 times more likely to be at fault than a sober pedestrian. The most relevant variable decreasing the likelihood that a pedestrian is at fault is driver contributing cause, which is a variable that indicates that a contributing cause had been assigned to the driver by the investigating officer. This shows that there is a correlation between the willingness of an officer to assign a contributing cause to the

TABLE 3 Pedestrian Fault Model Using Crash Report Data (Model P1)

Variable	Coef. (β_i)	Std. Err.	z	p	Odds Ratio	95% Conf. Interval	
						Low	High
Pedestrian intoxicated (0=N, 1=Y)	2.542	0.918	2.77	0.006	12.704	0.7428	4.3411
Driver gender (1=M, 2=F)	0.763	0.429	1.78	0.075	2.145	−0.0782	1.6047
Low lighting[a]	0.408	0.149	2.73	0.006	1.504	0.1157	0.7011
Speed limit	0.035	0.014	2.48	0.013	1.036	0.0074	0.0632
Driver intoxicated (0=N, 1=Y)	−1.217	0.730	−1.67	0.096	0.296	−2.648	0.2145
FHP reported (0=N, 1=Y)	−1.321	0.523	−2.52	0.012	0.267	−2.347	−0.2953
Wet road (0=N, 1=Y)	−1.858	0.735	−2.53	0.011	0.156	−3.298	−0.4183
Driver cited (0=N, 1=Y)	−1.905	0.592	−3.22	0.001	0.149	−3.065	−0.7454
Pedestrian section number	−1.980	0.457	−4.33	<.001	0.138	−2.875	−1.0843
Driver contributing cause (0=none, 1=any)	−2.005	0.478	−4.19	<.001	0.135	−2.943	−1.0677
Constant (β_0)	4.075	1.006	4.05	<.001	n/a	n/a	n/a

Dependent variable = ped_fault2; number of observations = 318; pseudo R^2 = .5058; log likelihood = −71.604; chi-square likelihood ratio = 146.54 (10 degrees of freedom); p-value = <.0001; coef. = coefficient; conf. = confidence; FHP = Florida Highway Patrol; N = no; Y = yes; M = male; F = female; n/a = not applicable.

[a] (0 = daylight, 1 = dusk, 2 = dawn, 3 = dark with streetlights, 4 = dark no streetlights).

TABLE 4 Pedestrian Fault Model Using Case Review Data (Model P2)

Variable	Coef. (β_i)	Std. Err.	z	p	Odds Ratio	95% Conf. Interval Low	95% Conf. Interval High
Pedestrian mental state[a]	1.334	0.263	5.07	<.001	3.798	0.8184	1.8504
Number of lanes attempted to cross	0.566	0.157	3.62	<.001	1.762	0.2593	0.8731
Driver age	0.074	0.032	2.29	0.022	1.077	0.0107	0.1368
Pedestrian age	−0.050	0.017	−2.86	0.004	0.951	−0.0838	−0.0157
Driver behavior class[b]	−0.392	0.152	−2.57	0.01	0.675	−0.691	−0.0937
Pedestrian exit vehicle (0=N, 1=Y)	−2.571	0.719	−3.58	<.001	0.076	−3.9807	−1.1618
Pedestrian section number	−2.839	0.714	−3.98	<.001	0.058	−4.2381	−1.4399
Pedestrian inconspicuous (0=N, 1=Y)	−3.259	1.913	−1.70	0.089	0.038	−7.0087	0.4917
Driver cited (0=N, 1=Y)	−3.505	0.931	−3.77	<.001	0.030	−5.3297	−1.6809
Constant (β_0)	4.870	2.005	2.43	0.015	n/a	n/a	n/a

Dependent variable = ped_fault3; number of observations = 318; pseudo R^2 = .7604; log likelihood = −34.711; chi-square likelihood ratio = 220.33 (9 degrees of freedom); p-value = <.0001.

[a](0 = not in categories 1–5, 1 = inattentive/distracted, 2 = error in perception, 3 = decision error, 4 = alcohol/drug impairment, 5 = suicide, Alzheimer's or other mental disorder).

[b](0 = not in categories 1–7, 1 = inattentive/distracted, 2 = error in perception, 3 = decision error, 4 = overcorrected, 5 = speed, 6 = alcohol/drug intoxication, 7 = incapacitation).

TABLE 5 Driver Fault Model Using Crash Report Data (Model D1)

Variable	Coef. (β_i)	Std. Err.	z	p	Odds Ratio	95% Conf. Interval Low	95% Conf. Interval High
Driver contributing cause (0=none, 1=any)	2.475	0.360	6.88	<.001	11.884	1.7705	3.18
Driver speeding (0=N, 1=Y)	2.334	0.578	4.04	<.001	10.316	1.2006	3.4669
Driver intoxicated (0=N, 1=Y)	1.887	0.583	3.24	0.001	6.601	0.7442	3.0302
Driver cited (0=N, 1=Y)	1.402	0.508	2.76	0.006	4.064	0.4061	2.3983
Driver gender (1=M, 2=F)	−0.568	0.292	−1.94	0.052	0.566	−1.1413	0.0048
Pedestrian intoxicated (0=N, 1=Y)	−1.049	0.440	−2.38	0.017	0.350	−1.911	−0.1868
Constant (β_0)	−1.385	0.387	−3.58	<.001	n/a	n/a	n/a

Dependent variable = dr_fault2; number of observations = 318; pseudo R^2 = .3424; log likelihood = −122.715; chi-square likelihood ratio = 127.77 (6 degrees of freedom); p-value = <.0001.

TABLE 6 Driver Fault Model Using Case Review Data (Model D2)

Variable	Coef. (β_i)	Std. Err.	z	p	Odds Ratio	95% Conf. Interval Low	95% Conf. Interval High
Driver cited (0=N, 1=Y)	2.179	0.574	3.8	<.001	8.841	1.0551	3.3038
Driver intoxicated (0=N, 1=Y)	1.956	1.198	1.63	0.102	7.073	−0.3919	4.3044
Driver mental state[a]	1.402	0.202	6.95	<.001	4.065	1.007	1.7977
Distance to signal[b]	0.412	0.169	2.44	0.015	1.510	0.0806	0.7442
Roadway ADT	<.001	<.001	2.12	0.034	1.000	1.22e-06	0.00003
Speed limit	−0.049	0.029	−1.69	0.091	0.952	−0.1055	0.0078
Number of lanes crossed	−0.232	0.142	−1.63	0.103	0.793	−0.5103	0.0467
Pedestrian mental state[c]	−0.712	0.164	−4.33	<.001	0.491	−1.0337	−0.3895
Driver gender (1=M, 2=F)	−0.971	0.463	−2.1	0.036	0.379	−1.8796	−0.0631
Driver section number	−1.713	0.559	−3.07	0.002	0.180	−2.8085	−0.6184
Wet road (0=N, 1=Y)	−2.299	1.422	−1.62	0.106	0.100	−5.0859	0.4879
Constant (β_0)	3.083	1.426	2.16	0.031	n/a	n/a	n/a

Dependent variable = dr_fault3; number of observations = 318; pseudo R^2 = .6758; log likelihood = −60.504; chi-square likelihood ratio = 252.19 (11 degrees of freedom); p-value = <.0001; ADT = average daily traffic.

[a](0 = not in categories 1–7, 1 = inattentive/distracted, 2 = error in perception, 3 = decision error, 4 = overcorrected, 5 = speeding, 6 = alcohol/drug intoxication, 7 = incapacitation).

[b](0 = not a factor/blank, 1 = less than 200 ft, 2 = 200 to 600 ft, 3 = 600 ft to 0.25 mi, 4 = 0.25 mi to 0.5 mi, 5 = 0.5 to 1.0 mi, 6 = greater than 1 mi).

[c](0 = not in categories 1–5, 1 = inattentive/distracted, 2 = error in perception, 3 = decision error, 4 = alcohol/drug impairment, 5 = suicide, Alzheimer's or other mental disorder).

driver, thereby increasing the propensity for the driver to be at fault and decreasing the chances that the pedestrian is at fault. The variables included in the model support the notion that driver fault and pedestrian fault are interrelated. For example, driver gender indicates that when a driver is a male, the pedestrian is less likely to be at fault. Male drivers have an increased propensity for risky behavior (*11*) and are thereby more likely to be assigned fault. By implication, if the driver is at fault, then the pedestrian is less likely to be at fault. The model also exposes the influence of an environmental condition that is common in Florida, wet roads. When the road was wet, the pedestrian was six times less likely to be found at fault than when the road was dry. By modeling pedestrian fault separately from driver fault, a particular human party is not required to be at fault, thereby allowing a roadway or environmental condition to have been the primary cause of a crash. In some cases, wet roads or poor lighting could have caused a crash, relieving the need to assign fault to either the pedestrian or the driver.

Pedestrian Fault Modeling by Using Data from Case Reviews

The parsimonious pedestrian fault model with accurate data of high quality extracted by case reviews (Model P2) is summarized in Table 4. The data in Table 4 are interpreted in the same manner as those in Table 3, as described above. The primary advantage of the model is that it is not limited to the coded information on the crash reports, the data in which have been shown to contain errors and lack detail. Examination of the variables relevant in Model P2 shows that human factors usually govern the prediction of fault, whereas roadway and environmental factors have little influence on fault. The highly relevant variable pedestrian mental state attempted to capture states of metal declension. The convention was to assign the lowest code possible by default; this code was overwritten by a higher code only when the case evidence revealed that the pedestrian exhibited such a condition. This means that for every increase in category for the pedestrian mental state variable, the odds that the pedestrian was at fault increases by a factor of $e^{1.334}$ (3.798). (See the footnote of Table 4 for details on the coding scheme.) For a pedestrian who was found to have an impairment caused by alcohol or drugs (Class 4), the odds of being at fault over a pedestrian in a normal mental state was 208 (3.798^4); if the pedestrian was suicidal, the odds of being at fault were 790 (3.798^5) times greater. Clearly, impairment caused by alcohol or drugs and suicidal behavior greatly increased the odds that a pedestrian was at fault.

According to the model, the greater the number of lanes that a pedestrian tried to cross before being hit, the more likely it is that he or she was at fault. However, pedestrian fault decreased greatly when a driver was cited, when a pedestrian had exited a vehicle, or when a pedestrian was assigned a higher section number on the crash report. Another finding was that a lack of pedestrian conspicuity decreased the chance that the pedestrian was at fault. If conspicuity was determined to be a factor, that is, if the pedestrian was hard to see, then the pedestrian was 26 times (1/0.038) less likely to be at fault. Typically, in this study, conspicuity in crashes was related to an environmental condition, for instance, thick fog, or to the fact that the pedestrian was wearing dark clothes at night in an area with no streetlights. The role of conspicuity in pedestrian fault prediction exemplifies the tendency to exonerate pedestrians of fault if any reasonable alternative is present.

Driver Fault Modeling by Using Data from Crash Reports

Intuitively, the variables used to predict driver fault in the driver fault model with crash report data (Model D1), as shown in Table 5, make sense. If the driver was assigned a contributing cause, was speeding, was intoxicated, was cited, or was a male, then the probability of driver fault increased. On the other hand, if the pedestrian was intoxicated, then the driver was less likely to be at fault. The most relevant variable in the model is driver contributing cause, which, like the pedestrian contributing cause variable, is a binary variable indicating whether any contributing cause was assigned to the driver. As with the pedestrian contributing cause variable, this variable shows that there is a correlation between the willingness of an officer to assign a contributing cause to the driver and the propensity of the driver to be at fault. The other two measures highly relevant in predicting fault are whether the driver was intoxicated and whether the driver was cited for a legal infraction. One possible interpretation of the odds ratios is that a driver to whom a contributing cause is assigned, who is speeding, and who is intoxicated is more than 800 times (11.88 × 10.32 × 6.60) more likely to be at fault than a driver who is coded with no improper driver action and who is neither speeding nor intoxicated.

Driver Fault Modeling by Using Data from Case Reviews

Another model for predicting driver fault is the parsimonious model that uses case review data (Model D2) that are accurate and of high quality, as summarized in Table 6. This model is useful for research or other applications in which the highest level of predictive accuracy is desirable and the resources exist to augment and improve the quality of the data.

The driver mental state variable distinguishes between different driver behaviors and actions in order of decreasing frequency. Risky, irresponsible, or other careless driver actions, such as inattention, errors in perception, decision-making errors, overcorrecting, speeding, alcohol or drug use, and even incapacitation, increased the odds that a driver was at fault. Roadway and environmental factors also influence driver fault in this model: the farther that a crash was from a signal, the more likely it was that the driver was at fault. Driver fault decreased with an increase in speed limit and the number of lanes crossed by the pedestrian before the crash. When the road was wet, the driver was 10 times (1/0.100) less likely to be found at fault as when the road was dry.

SUMMARY OF FINDINGS

The objective of this research was to compare the state of fault predicted by various logistic regression models with the true state of fault determined by an expert assessment and to evaluate the predictive capabilities of the fault models. The fault models were based on high-quality data extracted from case studies and also existing data housed by FDOT. The current FDOT algorithm for assigning fault, which relies primarily on the section number of the drivers or pedestrians in the crash and citations given, was found to be error prone in pedestrian cases. The bulk of the error was due to the failure of the FDOT algorithm to identify at-fault pedestrians and its tendency to wrongly classify drivers as at fault. The logistic regression models proved to be more accurate than the FDOT method for assigning fault in pedestrian crashes as a result of the use of additional relevant predictor variables coupled with the use of more accurate data. In both the pedestrian fault and the driver fault models, the difference in predictive capability was shown to be significant at the 95% confidence level.

Within the logistic regression models, those that used raw crash report data only (Models P1 and D1) were able to predict fault accurately in 84% to 87% of the cases (whereas the current FDOT algorithm had only 56% to 58% accuracy); those that used corrected crash report data (Models P3 and D3) were able to predict fault 91% of the time; and those that used full case study data (Models P2 and D2) were able to predict fault in 93% to 97% of the cases. Models P1 and D1 showed that the prediction accuracy could be improved by fully using all of the data available on the crash report to predict fault rather than relying only on the section number and the citations given. Models P3 and D3 showed that the correction of errors in those data improved the accuracy even further. So, when additional data resources are not available, simply improving the quality of the crash report data would result in increases in predictive capability. This outcome shows the importance of data accuracy in both the collection and the transcription of crash records data. However, to obtain the highest predictive capability, full case studies could be used: models that used full case studies produced even more accurate fault predictions than those that used corrected crash report data, but this was at the expense of an additional effort in the collection and processing of the data. The factors found to be significant when fault in pedestrian cases was considered included mental state, alcohol consumption, and the ages of the driver and the pedestrian; pedestrian status as a former vehicle occupant; pedestrian conspicuity; the presence of driver citations; and the wetness of the road surface. It is recommended that parties interested in modeling fault collect these data in such crashes.

A limitation of this study was the narrow focus, in that the study looked only at fatal pedestrian crashes occurring on state roads in Florida. Future efforts that include nonfatal pedestrian crashes would be valuable. Fatal crashes were chosen for this study in part because of the detailed data in the Traffic Homicide Investigation report. However, because the research showed that improved predictions are possible only when the data fields from the original crash report are used, the work could be extended to nonfatal crashes. Work could be expanded to data from other states or to the prediction of fault among subgroups of pedestrian crashes. As other states (e.g., Illinois and Michigan) use a similar approach of placing the at-fault driver or pedestrian in the first section of the crash report, the effect of false positives may be of interest to practitioners elsewhere in the country; however, the overall examination of driver and pedestrian fault would be universally applicable.

Findings of high rates of pedestrian fault do not minimize the importance of providing roadway features for safe pedestrian travel. For instance, a companion study showed that in more than half of the pedestrian fatalities that involved crossing the road without a crosswalk, no protected crossings were available within 600 ft of the selected crossing location and that in almost 25% of the cases, the nearest protected crossing was more than a quarter of a mile away (9). Consideration of potential pedestrian activity in roadway design and traffic operation decisions can help ensure safe, legal pedestrian travel.

ACKNOWLEDGMENT

This research was funded in part by the State Safety Office, Florida Department of Transportation.

REFERENCES

1. *Traffic Safety Facts 2000: Pedestrians.* Publication DOT-HS-809-331. NHTSA, Washington, D.C. www-nrd.nhtsa.dot.gov/pdf/nrd-30/NCSA/TSF2000/2000pedfacts.pdf. Accessed July 31, 2006.
2. Wootton, I. A., L. K. Spainhour, and J. O. Sobanjo. Improved Methods for Reviewing Florida Pedestrian Fatal Crash Data. Presented at 30th International Forum on Traffic Records and Highway Information Systems, National Safety Council, Nashville, Tenn., 2004.
3. Kim, K., and L. Li. Modeling Fault Among Bicyclists and Drivers Involved in Collisions in Hawaii, 1986–1991. In *Transportation Research Record 1538,* TRB, National Research Council, Washington, D.C., 1996, pp. 75–80.
4. Kim, K., and J. Boski. Finding Fault in Motorcycle Crashes in Hawaii: Environmental, Temporal, Spatial, and Human Factors. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1779,* TRB, National Research Council, Washington, D.C., 2001, pp. 182–188.
5. Kim, K., L. Li, J. Richardson, and L. Nitz. Drivers At-Fault: Influences of Age, Sex, and Vehicle Type. *Journal of Safety Research,* Vol. 29, No. 3, 1998, pp. 171–179.
6. Hosmer, D., and S. Lemeshow. *Applied Logistic Regression.* John Wiley and Sons, Inc., New York, 1989.
7. Pampel, F. C. *Logistic Regression: A Primer.* Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-132. Sage, Thousand Oaks, Calif., 2000.
8. Grimm, L. G., and P. R. Yarnold. *Reading and Understanding Multivariate Statistics.* American Psychological Association, Washington, D.C., 1995.
9. Spainhour, L. K., D. Brill, J. O. Sobanjo, J. Wekezer, and P. Mtenga. *Evaluation of the Traffic Crash Fatality Cause and Effects.* Report DB-050. Florida Department of Transportation, Tallahassee, 2005.
10. Mantena, S., L. K. Spainhour, I. A. Wootton, Y. A. Owusu, R. Sotter, R. N. Mussa, and J. O. Sobanjo. Improving Accuracy and Efficiency of Florida Traffic Records Through Automated/Electronic Data Collection. Presented at 30th International Forum on Traffic Records and Highway Information Systems, National Safety Council, Nashville, Tenn., 2004.
11. Evans, L., and P. Wasielewski. Risky Driving Behaviors Related to Driver and Vehicle Characteristics. *Accident Analysis and Prevention,* Vol. 15, 1983, pp. 121–136.

The opinions, findings, and conclusions expressed in this publication are those of the authors and not necessarily those of the Florida Department of Transportation.

The Pedestrians Committee sponsored publication of this paper.