Examining Deficiencies in Florida Pedestrian Crash Data

Isaac A. Wootton and Lisa K. Spainhour

In Florida, the Department of Highway Safety and Motor Vehicles serves as the repository for state traffic crash data collected by law enforcement officers. These state data, which are collected on traffic crash reports, limit the type and the extent of the analysis that can be performed because of constraints and errors in the data. Florida state databases lack information from crash narratives and diagrams, which are not mined for data. In addition, crash report data are shown to contain errors. An analysis was initiated to investigate data from 318 fatal crashes involving pedestrians in which detailed traffic homicide reports and other data sources were consulted. The integrity of the state-maintained data for their accuracy and completeness was investigated. Alcohol usage, fault, speed limits, vehicle speeds, and citations were the leading data fields with errors. In the case of pedestrian alcohol test results, state database records were found to be in error more than half the time. Emphasis is given to the methods that can be used to create a quality crash data set and to highlighting the additional insights that can be gained from the homicide reports and other resources, especially for the accurate determination of fault and crash causation.

Pedestrian crash data compiled by the Florida Department of Highway Safety and Motor Vehicles (DHSMV) indicate that 5,220 pedestrians were killed in Florida during the 10-year period from 1995 to 2004 (*1*). The high incidence of pedestrian fatalities, 16.4% of all traffic fatalities in Florida during the year 2000 (*2*), demonstrates the importance of state safety studies focusing on pedestrian issues. A traditional method of analyzing Florida pedestrian fatal crash data comes from a review of data extracted from the uniform Florida traffic crash report, as is the case with the statistics presented above. Many times, to make meaningful, valuable, and useful conclusions, traffic safety investigations, engineering studies, and statistical analyses require more data than are available on a crash report alone, need those data to be of high quality, and want those data in a usable format. According to Benavente et al., accurate and comprehensive data are a requirement for the successful assessment of the consequence of crashes (*3*). By soliciting multiagency resources within the state of Florida, electronic and paper law enforcement agency records at both the local and the state levels, and the expertise of crash investigation and safety experts, much additional information and insight can be gained to aid with the analysis and study of pedestrian fatal crash data, especially since the data can be trusted to be more complete and of higher quality.

DATA AND METHODS

To evaluate the crash report data in Florida state databases, a data set that was augmented and corrected by case reviews by using a variety of data sources was created. The mitigated data served as the control for evaluation of the quality of the crash report data. Data transformations were performed on variables in each of the two data sets so that the data would be suitable for use in comparison models. Statistical analysis was performed on the two data sets; and comparative measures, including Spearman's rho and interrater agreement parameters, were obtained. These measures were used to evaluate mismatches in the data sets and to evaluate potential sources of the discrepancy. Finally, methods for improving the accuracy of crash data are discussed.

A total of 318 fatal pedestrian cases occurring on state-maintained roadways in 2000 were used for this study. This study was part of a larger study of crash causation and the factors significant in fatal crashes in Florida (*4*). Two data sets are compared here. The first, termed crash report data, came directly from Florida Department of Transportation (FDOT) databases. This data set was limited to coded data extracted from standard Florida Traffic Crash Reports, for which the data are 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 key source of information for case reviews was detailed Traffic Homicide Investigation (THI) 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 roadway photographic archive system were carefully reviewed. When necessary, site visits and accident reconstructions were conducted. As part of the case review process, errors in the coded data were corrected, additional data fields were introduced, and a manual assessment of fault was conducted. 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.

FDOT uses and maintains a state mainframe database known as the Crash Analysis Reporting (CAR) system. This database is the data source for state-level crash statistical analysis, other state uses, and federal reporting. This database contains data input from the fields of the Florida traffic crash report, the location of the crash determined in-house by FDOT, and other limited site information queried from the state-maintained Roadway Characteristics Inventory (RCI) mainframe database. To obtain information beyond the data currently made available by the FDOT CAR database, a case study analysis, or a case-by-case examination, was conducted for each pedestrian fatal crash in the study, specifically, to determine the cause of the crash and the cause of the fatality or fatalities. However, a by-product of this detailed case review process was a scrutiny of the quality of the

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: I. A. Wootton, isaac@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. 31–38.
DOI: 10.3141/2002-05

data. Reviews of case-level traffic THI reports, state-maintained roadway photographic logs, site and crash scene photographs, site visit reports, query results of additional RCI database elements, and spatial data were conducted. Extracts from the FDOT CAR database and data extracted from the additional resources were used to create a data set for comparison. Each of the data sources used in this study is discussed in greater detail in the following sections. The sources of pedestrian crash information discussed here are valuable resources that can be used to improve or remedy errors in state crash report data or whenever a case review approach is taken.

Florida Traffic Crash Reports

Law enforcement agencies in Florida report fatal traffic crashes using a uniform traffic crash report long form. The Florida DHSMV collects the paper forms, provides them to a vendor for data entry into an archival database, and then supplies the computerized crash data to FDOT. For crashes on state roads, FDOT adds location data referenced to the roadway segment and mile point of the state location reference system. FDOT stores the data elements (other than the narrative and diagram) in the CAR database and archives TIFF images of the paper forms. Copies of crash reports and electronic records were furnished by FDOT for this project. The crash reports were reviewed, the narratives were read, and the diagrams were studied to aid with the development of a sequence of events for each crash. Citations were also reviewed according to the statute number and type, and it was noted when statutes were violated but citations were not given because of a fatality.

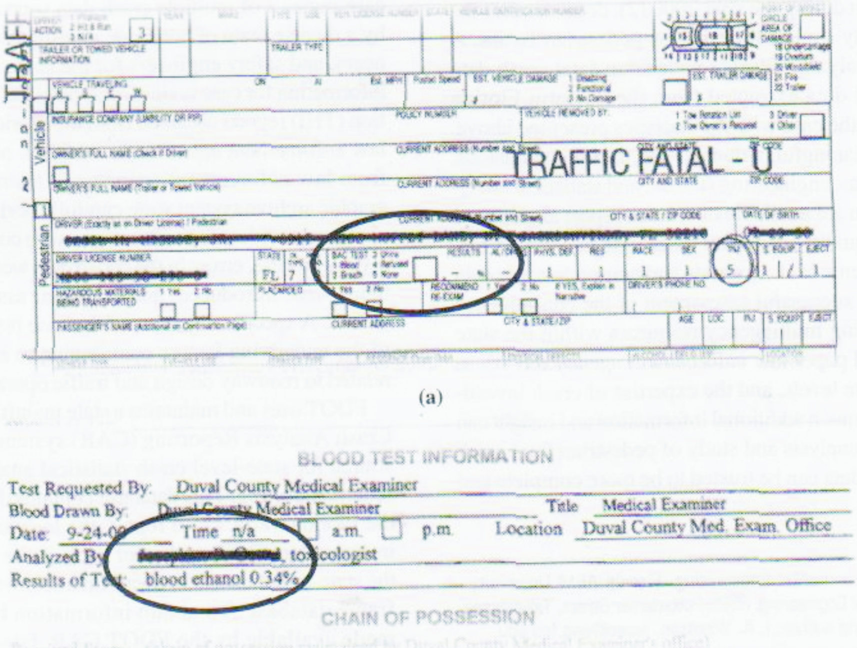
THI Reports

The Florida Highway Patrol and other local law enforcement agencies in Florida conduct a detailed traffic homicide investigation when a crash results in a fatality. THI reports are usually significantly more detailed than the crash reports, are in narrative format, and many times include a scaled crash scene diagram and sometimes even include reconstruction information. THI reports were reviewed, as they were a significant source for verifying, augmenting, or correcting the available information. The benefits of using the THI reports included the following:

- Checking the data consistency of the Florida traffic crash report. One of the largest causes of inconsistencies involved alcohol use, in which the THI report, which is based on autopsy and other medical information, notes the use of alcohol or drugs, whereas the Florida traffic crash report did not (Figure 1).
- Determining which driver or pedestrian was at fault, as indicated by the traffic homicide investigator or as implicated by the details given for the crash. In this study it was noted when there was an inconsistency with the FDOT at-fault determination and when there were questionable causes or multiple faults.
- Reviewing driver history (past crashes or citations, including adjudications) when it was provided as part of the THI report.
- Reviewing crash circumstances and categorizing potential contributing causes as environmental, roadway, vehicle, and person (driver, passenger, or nonmotorist). Issues to be investigated during site visits can be identified.

State Roadway Characteristics Inventory

FDOT maintains an electronic inventory of the state highway system known as the Roadway Characteristics Inventory (RCI). This database can be accessed and the data can be retrieved in particular cases in which additional quantitative roadway information or features are needed (Figure 2). FDOT augments all crash records in the CAR database with a snapshot of a limited number of features from the



TRAFFIC FATAL - U

DRIVER ACTION 1 Proper 2 Hit & Run 3 N/A 3

TRAILER OR TOWED VEHICLE INFORMATION | TRAILER TYPE

VEHICLE TRAVELING | ON | AT | Est. MPH | Posted Speed | EST. VEHICLE DAMAGE 1 Disabling 2 Functional 3 No Damage | EST. TRAILER DAMAGE

INSURANCE COMPANY (LIABILITY OR PIP) | POLICY NUMBER | VEHICLE REMOVED BY: 1 Tow Rotation List 2 Tow Owner's Request 3 Driver 4 Other

OWNER'S FULL NAME (Check if Driver) □ | CURRENT ADDRESS (Number and Street) | CITY AND STATE | ZIP CODE

OWNER'S FULL NAME (Trailer or Towed Vehicle) | CURRENT ADDRESS (Number and Street) | CITY AND STATE | ZIP CODE

DRIVER (Exactly as on Driver License) / Pedestrian | CURRENT ADDRESS (Number and Street) | CITY & STATE / ZIP CODE | DATE OF BIRTH

[redacted] | [redacted]

DRIVER LICENSE NUMBER [redacted] | STATE FL | DL TYPE 7 | BAC TEST 1 Blood 2 Breath 3 Urine 4 Refused 5 None □ | RESULTS . – % | AL/DRUGS – | PHYS. DEF 1 | RES 1 | RACE 1 | SEX 1 | INJ 5 | S. EQUIP 3 / | EJECT 1

HAZARDOUS MATERIALS BEING TRANSPORTED 1 Yes 2 No □ □ | PLACARDED 1 Yes 2 No □ □ | RECOMMEND RE-EXAM 1 Yes 2 No □ □ | If YES, Explain in Narrative | DRIVER'S PHONE NO.

PASSENGER'S NAME (Additional on Continuation Page) | CURRENT ADDRESS | CITY & STATE/ZIP | AGE | LOC | INJ | S. EQUIP | EJECT

(a)

BLOOD TEST INFORMATION

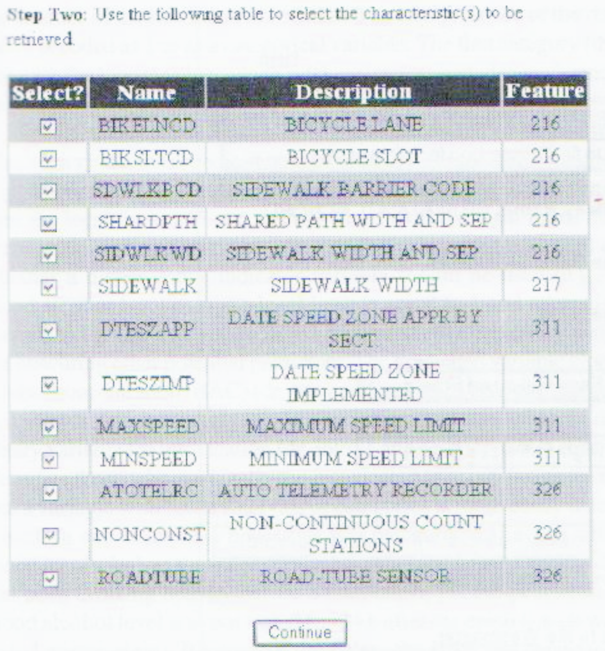
Test Requested By: Duval County Medical Examiner
Blood Drawn By: Duval County Medical Examiner Title Medical Examiner
Date: 9-24-09 Time n/a □ a.m. □ p.m. Location Duval County Med. Exam. Office
Analyzed By: [redacted] toxicologist
Results of Test: blood ethanol 0.34%

CHAIN OF POSSESSION

Received From: (chain of possession maintained by Duval County Medical Examiner's office)

(b)

FIGURE 1 Example of alcohol reporting inconsistency: (*a*) Florida traffic crash report alcohol results left blank and (*b*) THI report gives chemical test results.



Step Two: Use the following table to select the characteristic(s) to be retrieved.

Select?	Name	Description	Feature
☑	BIKELNCD	BICYCLE LANE	216
☑	BIKSLTCD	BICYCLE SLOT	216
☑	SDWLKBCD	SIDEWALK BARRIER CODE	216
☑	SHARDPTH	SHARED PATH WDTH AND SEP	216
☑	SIDWLKWD	SIDEWALK WIDTH AND SEP	216
☑	SIDEWALK	SIDEWALK WIDTH	217
☑	DTESZAPP	DATE SPEED ZONE APPR BY SECT	311
☑	DTESZIMP	DATE SPEED ZONE IMPLEMENTED	311
☑	MAXSPEED	MAXIMUM SPEED LIMIT	311
☑	MINSPEED	MINIMUM SPEED LIMIT	311
☑	ATOTELRC	AUTO TELEMETRY RECORDER	326
☑	NONCONST	NON-CONTINUOUS COUNT STATIONS	326
☑	ROADTUBE	ROAD-TUBE SENSOR	326

FIGURE 2 Sample RCI characteristics that relate to pedestrian crashes.

RCI database. It is noteworthy that the main RCI database is updated on a 3-year basis, and additional roadway characteristics at the time of the crash can be lost because of the dynamic nature of the RCI database. Furthermore, because up to a year can pass before the CAR database is augmented with the RCI data, it is possible that the time stamp for the RCI data no longer matches that of the crash at the time of the update. No attempt was made to ascertain the degree of this problem.

Crash Scene Photos

Crash scene photos can be reviewed to look for roadway features significant in pedestrian crashes (e.g., pedestrian signals, crosswalks, and shoulder warning devices). Skid marks and other evidence that can be used to make a hypothesis about the driver's action were used as part of the case review process (e.g., skid marks could show the evasive actions taken to avoid a collision, and a lack of skid marks combined with a small exit angle could indicate driver fatigue issues). Figure 3 shows an example of a crash scene photo taken as part of a traffic homicide investigation. When sufficient information was available from the crash or homicide reports and accompanying photos, vehicle speeds were reconstructed and driver perception–reaction times and steering inputs were determined.

Roadway Video Log

Video logs provided by FDOT can be accessed and reviewed. Video logs are still photos taken in both directions at 3-year intervals from the rightmost lane of all state-maintained roads (Figure 4). The video logs are useful for the investigation of potential roadway design and traffic operations issues that might contribute to a crash or a fatality. The video logs are still valuable for investigating things such as sight distances, signage, crosswalks (marked and unmarked), the presence of pedestrian signals, speed transition zones, and other information.

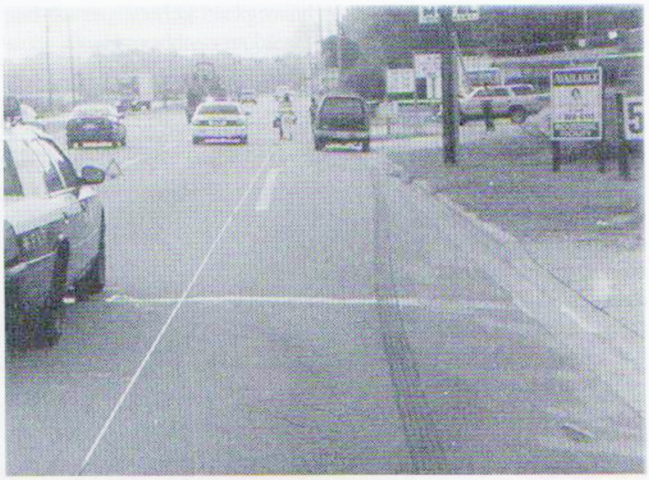


FIGURE 3 Sample crash scene photograph.

It was found that most of the video log images on file were taken during the same year as the crash study; however, this may have proved to be more of a problem had a different study period been chosen. In some cases consultation of the video logs can supplant the need for a costly and time-consuming site visit.

Site Visits

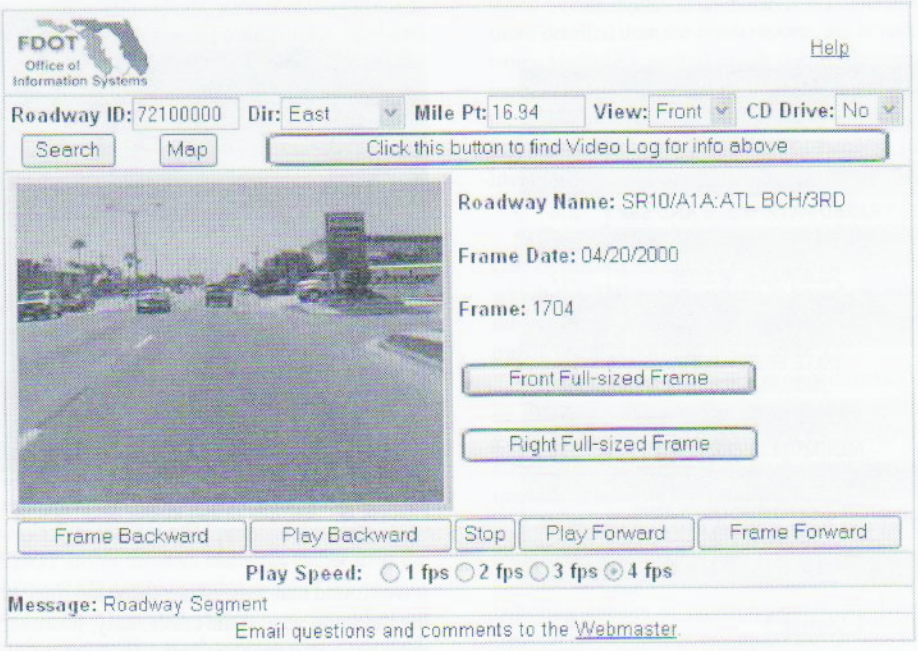
Site visits can be conducted when necessary to investigate potential roadway characteristics that contributed to the cause of a crash and when all other resources have been exhausted. Depending on the site, team members can conduct a drive through in the direction of both at-fault and not-at-fault vehicles; measure and photograph various site features, including approach features; time light or pedestrian cycles; and investigate sun glare, traffic volumes, and other time-dependent aspects. Depending on the needs of a specific case under review, the site visit data can be collected in either a qualitative or a quantitative manner.

Geographic Information System

The FDOT Planning Office maintains an inventory of data for use with geographic information systems (GISs). The layers include roadway data and traffic data. Another source of GIS data is the Florida Geographic Data Library. This online library, which is warehoused and maintained by the University of Florida, currently contains more than 350 layers of GIS data, including land use, census, traffic meter data, and U.S. Geological Survey quad maps. A GIS layer of the crash locations in this study was created by adding route events to a state base map. The GIS data enabled spatial analysis, in which plots of the study regions can be used to examine how different characteristics vary by location (Figure 5).

DATA MANIPULATION

In this study, a data set that was augmented and corrected by case reviews by using the data sources described above was created. The mitigated data served as the control for evaluating the quality of the crash report data. To obtain data suitable for use in comparison models, data transformations on variables in each of the two



Roadway ID: 72100000 **Dir:** East **Mile Pt:** 16.94 **View:** Front **CD Drive:** No

Roadway Name: SR10/A1A:ATL BCH/3RD

Frame Date: 04/20/2000

Frame: 1704

Play Speed: ○ 1 fps ○ 2 fps ○ 3 fps ● 4 fps

Message: Roadway Segment

Email questions and comments to the Webmaster.

FIGURE 4 Sample screenshot from FDOT video log viewer.

data sets were performed in a consistent manner. Variable discretization was used for many continuous variables. The term "discretization" refers to the partitioning of the range of values taken by a continuous variable into subranges that can then be treated as discrete categories. In many cases the discrete categories required a logical ordering whose consideration is important when the results are interpreted. Final coding and data manipulation techniques yielded a combination of continuous and categorical variables.

Careful consideration and even multiple iterations were needed for the development of classification schemes for certain variables. A fundamental data consideration was whether the variables selected and the coding schemes effectively captured essential features of the data source and would be useful for analytical methods. In some cases all that was needed was numerical coding of the values according to a revised classification scheme, in which the first category is 0 (e.g., lighting, where daylight is equal to 0, dusk is equal to 1, dawn is equal to 2, dark with streetlights is equal to 3, and dark with no streetlights is equal to 4). Frequently, only the Boolean outcome (the factor is present or not present) was coded. It was important to code categorical variables beginning with 0 and dichotomous variables as 0 and 1 (rather than as 1 and 2 or otherwise), because the coding affects the usefulness for modeling and statistical estimates. Usually,

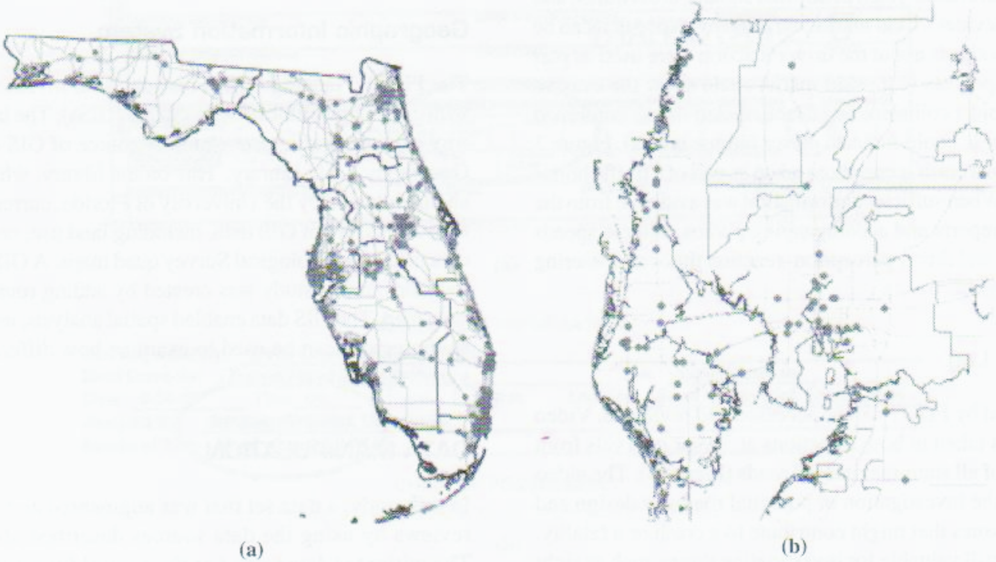


FIGURE 5 Sample GIS plots: (*a*) statewide locations of pedestrian crashes in this study and (*b*) locations of pedestrian crashes in the Tampa area.

the absence of the risk factor is coded as 0 and the presence of the risk factor is coded as 1 or as a categorical variable. The first category (the one with the lowest value, which is 0 here) is designated the reference category.

In preliminary work, much effort was expended to determine whether a graduated scheme should be used to code specific outcomes (e.g., sober, had been drinking but under the limit, one to two times over the legal limit, and two to three times over the legal limit) that occur within a broad outcome category (e.g., driver intoxication) or whether a dichotomous indicator variable should be used (e.g., a blood alcohol level over the legal limit is equal to 1 and otherwise is equal to 0). The goal was to have a high predictive capability while not obscuring other potential predictors. The graduated variable blood alcohol concentration (BAC) (dr_bac1 and ped_bac1) is more specific by indicating an increasing propensity for intoxication, whereas the binary variable for intoxication (dr_intox1 and ped_alcohol1) simply identifies those who are legally too intoxicated to drive. It was found that a variable with fewer intervals or categories is more easily determined. In one example, a pedestrian was categorized as being influenced by alcohol on the basis of witness statements that a pedestrian had been drinking heavily before a fatal crash, yet the pedestrian's blood alcohol level was not tested for 24 h after the crash (i.e., it was tested postmortem). Because of the delay, the BAC test reported a negligible blood alcohol level. The use of a dichotomous variable allowed the result to be reported as 1 as a result of the lack of additional quantitative data. Preliminary modeling attempts showed that the main advantage of using the dichotomous dr_intox1 and ped_alcohol1 variables is that they are highly predictive. However, in most cases they led to a model with fewer significant variables because the high strength of association overpowered the other predictors. Another limitation to the use of only dr_intox1 and ped_alcohol1 as predictors is in interpreting the results, especially in the case of pedestrian behavior, where there is not a clear legal distinction of what BAC constitutes intoxication. Both coding schemes were included in the final model, with the creation of both graduated and dichotomous variables, leaving the analysis to determine which was more useful.

ANALYSIS AND RESULTS

State crash report records were found to have problems that negatively affected the quality of crash data, including errors, incomplete or missing information, illegibility, and errors introduced by multiple data entries at various levels. The case-by-case review approach, the redundancy of consulting multiple data sources, consistency checks, and the engineering background of the researchers performing this study sought to address problems with crash report data and produce a case review data set that was of a high level of quality. The case review data were used as a standard for judging the deficiencies in the original crash report data. Although many variables extracted from the FDOT database of crash report data contained at least a few missing values or erroneous values, some variables were found to be particularly error prone, with a high percentage of errors compared with those for data corrected by case review. To facilitate analysis and to allow comparison, variables extracted from the FDOT database were coded to correspond to the data coding of the case review data. Seven variables were found to have large inconsistencies, which then prompted additional study. A list of these error-prone variables and their coding schemes is given in Table 1. The seven error-prone variables were also investigated to verify the magnitude of the error, the strength of the correlations, and the nature of the inconsistencies between variables on the basis of crash report data and of case review data. The results are summarized in Table 2. Note that the results for alcohol levels are presented in terms of the graduated variables (dr_bac1 and ped_bac1) rather than the binary variables (dr_intox1 and ped_alcohol1).

The results of the investigation into error-prone variables confirmed the findings that the case reviewers had seen. The direction of mismatch in Table 2 reveals the error trends noted during the case reviews. For instance, the large number of negative mismatches for pedestrian fault and the large number of positive mismatches for driver fault both resulted from crash report data that frequently assigned fault to a driver rather than a pedestrian. A negative mismatch in the ped_bac variable occurred in large part because the pedestrian was coded as "not drinking or using drug" on the crash report, a situation that resulted in a pedestrian BAC of 0. Case reviews, however, which benefited from BAC test results that were not available at the time that the initial crash report was completed, often indicated some level of alcohol consumption. The speed limit mismatch was largely due to cases in which the speed limit information was missing from the crash report records. Citations were also found to be missing or wrongly attributed to the wrong driver or pedestrian in crash report data, leading to the mismatches found. The negative mismatch in the variable speeding was predominantly the result of pedestrian accident reconstructions that yielded higher vehicle speeds than those reported on crash reports.

A nonparametric correlation between two ordinal variables, Spearman's rho, was computed along with the results of a test of interrater agreement. A Spearman's rho value of 1.0 describes a perfect correlation between variables, a condition in which the crash report data

TABLE 1 Error-Prone Pedestrian Variables: Description and Definitions

Variable	Description	Code Value Definitions
ped_bac	Pedestrian blood alcohol	(1) 0.00 (2) 0.00 pres. (3) < Limit (4) > 0 (5) 1–2 X LIM (6) 2–3 X LIM (7) 3–4 X LIM (8) > 4 X LIM (9) Uk
ped_fault	Pedestrian fault	(0) Not at fault (1) At fault (2) Unknown
dr_fault	Driver fault	(0) Not at fault (1) At fault (2) Unknown
dr_bac	Driver blood alcohol	(1) 0.00 (2) 0.00 pres. (3) < Limit (4) > 0 (5) 1–2 X LIM (6) 2–3 X LIM (7) 3–4 X LIM (8) > 4 X LIM (9) Uk
speed_limit	Posted speed limit	Value in mph
dr_cited	Driver cited	(0) No/Unknown (1) Yes
speeding	Driver exceeding posted speed limit	(0) No/Unknown (1) Yes

pres. = presumed; LIM = limit; Uk = unknown; bac = blood alcohol concentration.

TABLE 2 Error-Prone Variables in Pedestrian Data Set

	Mismatches[a]					Interrater Agreement		
Variable	Tot.	Pos.[b]	Neg.[c]	% Error	Spearman's ρ	% Agreement	κ-Statistic	Strength of Agreement
ped_bac	168	38	130	52.8	0.3420	47.17	0.4063	Moderate
ped_fault	141	3	138	44.3	0.3225	55.66	0.2115	Fair
dr_fault	134	112	22	42.1	0.2354	57.86	0.1983	Poor
dr_bac	75	22	53	23.6	0.6558	76.42	0.6384	Good
speed_limit	75	0	75	23.6	0.6297	76.42	0.7228	Good
dr_cited	44	6	38	13.8	0.5016	86.16	0.4616	Moderate
speeding	39	2	37	12.3	0.4619	87.74	0.3852	Fair

[a] 318 total cases reviewed. Mismatch refers to when crash report value differs from case review value.
[b] Positive indicates a case where crash report value > case review value.
[c] Negative indicates a case where crash report value < case review value.

and the case review data match perfectly. A common way of interpreting the strength of correlation is for values under 0.5 to show a low, if any, correlation and for values between 0.5 and 0.7 to show only a moderate correlation. For example, the pedestrian alcohol results reported by FDOT show low, if any, correlation to the pedestrian alcohol results collected by case study reviews. Crash report data in state databases in Florida greatly underreport alcohol usage among pedestrians. Examination of the interrater agreement between crash report data and case review data shows similar findings. The kappa statistic measures the concordance between multiple raters on a nominal scale across cases and can also be used to find the degree of agreement between the judgments of two examiners or, in this case, two data sets (*5*). The strength of agreement reported in Table 2 interprets the kappa statistic as given by Altman (*6*).

DISCUSSION OF RESULTS

The discussion here notes and highlights the deficiencies that were found and addressed by the case study approach to improving data quality. Quality data are defined in *NCHRP Synthesis of Highway Practice 192* as data that are accurate, complete, precise, and timely (*7*). It is reasoned that the higher the quality of the data is, the more useful and robust the data are for analysis and the greater the confidence that may be placed in the findings and results.

Accuracy

With crash records, there are generally two fundamental accuracy problems. The first is related to location accuracy, and the other is related to data accuracy. Although the reporting officer fills out the fields on the crash form to furnish crash location, FDOT personnel are responsible for determining and assigning the exact crash location on the state roadway system. The location accuracy is assumed to be improved because of FDOT's involvement. The second accuracy problem typically associated with a crash report is the accuracy of the recorded data. The entry of information in a crash report involves coding, writing or typing, and drawing. Some information about the crash, such as vehicle type, vehicle actions, injuries, damages, roadway, and environment data, is coded by selecting appropriate elements from predefined lists printed on the crash report. However, other data, such as personal and vehicle information, narrative, and location information, must be explicitly written or typed on the form because such information cannot be coded, as each individual crash has unique characteristics and locations and different people and vehicles are involved. A diagram usually depicts the occurrence and the circumstances of a crash, such as the movements of vehicles and the pedestrians involved, an arrow pointing north, and signals or other fixed objects. The diagram may also include information about street or highway names and distances.

It was found that the data accuracy, or the degree to which the crash report data are correct, could be improved by gleaning the information contained in the THI reports. Most times the traffic homicide investigation was performed by a law enforcement officer different from the one who compiled the crash report. As reflected in the THI reports, traffic homicide investigators are usually more detailed, specialized, experienced, and trained when it comes to reporting crashes.

In Florida, after the local agencies send crash reports to the central repository at DHSMV, data on crash reports are reentered to record them in the state crash database. The DHSMV outsources the crash data entry to Prison Rehabilitative Industries and Diversified Enterprises Inc., in which prisoners image the crash reports and conduct data entry. During this process, crash data are subject to arbitrary changes because of a number of factors, including unreadable handwriting, incorrect data entry, a lack of expertise, and poor quality control practices. A bias introduced by prisoners has been noticed specifically in the case of crashes involving police officers. An example of this may have been seen in one case in which the electronic records omitted the involvement of two pedestrians, both of whom were police officers, who were injured while they were arresting a suspect.

All errors in crash data, whether they are introduced by the reporting officer, data entry personnel, or data algorithms, and especially the miscoding of some data elements, such as the number of pedestrians involved, alcohol use, the crash location, the crash type, injury severity, surface condition, light condition, vehicle type, pedestrian action, or seat belt use, can lead to either inappropriate conclusions or an inability to use the electronic data in their native form. The most common examples of officer-related factors affecting data quality that were found, to name a few, included the tendency to select and use only a few codes from pick lists, usually those near the top of the list, even though more selections are available (e.g., contributing cases coded as "careless driving" instead of "alcohol & drugs—under influence" when alcohol is clearly involved); the miscoding of data elements or, in some cases, conflicts among the information given (e.g., the time of the crash conflicts with the light conditions); the failure to report some data elements; the lack of details in collision diagrams, such as important measurements, reference points, the identification of Vehicle 1 versus Vehicle 2, and so forth; poor handwriting; and misspellings. Mainframe algorithms applied to crash data at the state level also introduce errors, the most noticeable being the classification of the at-fault party; columns containing summary statistics (such as the number of pedestrians, vehicles, and fatalities involved in a crash), and calculated values (such as driver or pedestrian age) were also found to contain errors.

The most frequent of the errors introduced by the data algorithms in pedestrian cases was found to be the at-fault determination. According to the FDOT Safety Office, fault is defined on the basis of the citations issued in the crash. The person in the first, or lowest-numbered, section of the form cited in the crash is the one classified as at fault, unless another person in the crash was cited, and then that driver or pedestrian is identified as being at fault in the crash. This rationale is used because law enforcement officers typically code the at-fault driver or pedestrian as the first vehicle in a crash report. However, in a vehicle–pedestrian crash, this study found that officers tend to list the vehicle first, regardless of fault. Anecdotal evidence from investigating officers indicated that this occurred because the pedestrian was often being treated for injuries and was unavailable at the start of the crash investigation. Furthermore, in the case of a pedestrian fatality, a citation would not be issued to a pedestrian who had died; thus, the state database would fail to reclassify the at-fault party from the person in the first section to the one in the second section. Thus, by using FDOT's algorithm for assigning fault, in many pedestrian cases, the fault was mistakenly assigned to the vehicle driver in the first section of the form, as the review in this study found that the pedestrian was clearly at fault.

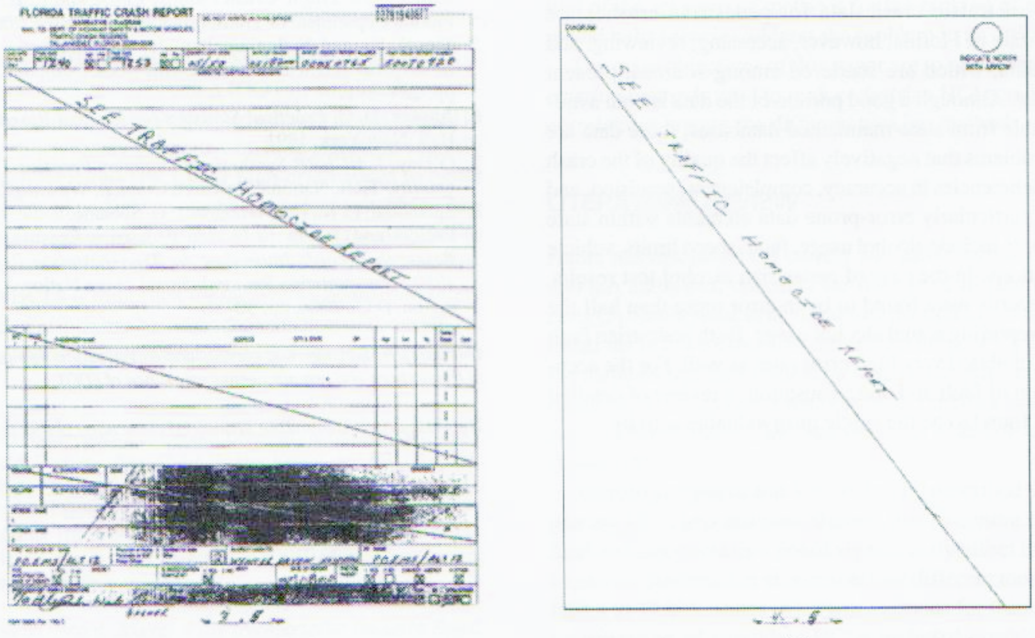
Precision

For the purposes of this study, precision can be defined as having the appropriate level of detail needed in data reporting. An example of a lack of precision, at least for the purposes of a safety investigation, could be seen in the codes used for vehicle classification on the crash report. Although vehicle type has 16 values in a predefined list, the involvement of an SUV cannot be determined from the crash report data. Vehicle identification number (VIN) decoding enabled the determination of vehicle make and model. The precision of data needed, especially those for skid mark lengths and signal timings, relied on data extracted from traffic homicide reports, VIN decoding, or, in some cases, actual site visits.

Completeness

The completeness of the traffic data addresses missing data. Missing data may be a result of the failure to report the data, to submit the report to a central repository, to enter the data into a system, or to find data in the system. The incompleteness or missing data limited to a particular crash could usually be corrected by consulting the THI report. A commonly found issue and a growing trend in crash reporting around the state is for officers to intentionally leave contributing causes, harmful events, narrative, or the crash diagram blank on the traffic crash report, with a comment to see the homicide investigation, as shown in Figure 6. Since homicide investigations are not submitted along with the crash reports to the central repository, the data in DHSMV and FDOT databases, as subsequently reported to the national Fatality Analysis Reporting System (FARS), are almost always incomplete for these crashes.

The consistency or uniformity of the data is closely related to the completeness of the data, at least at the state level. Inconsistency or a lack of uniformity in reporting method thresholds or in the definitions



FLORIDA TRAFFIC CRASH REPORT

SEE TRAFFIC HOMICIDE REPORT

[redacted]

(a)

SEE TRAFFIC HOMICIDE REPORT

(b)

FIGURE 6 Crash report: (*a*) narrative and (*b*) diagram are intentionally left incomplete.

of data elements would make it difficult to combine data and perform a statewide crash analysis. A good example of the lack of consistency is in the definition of pedestrians. The FDOT Safety Office, like FARS, defines a pedalcyclist as a person on a vehicle that is powered solely by pedals, whereas a pedestrian is any person not in or on a motor vehicle or other vehicle. If a person is riding a bicycle, he or she is on a vehicle and thus is a pedalcyclist and, by definition, cannot be a pedestrian. A driver who is not inside his or her vehicle at the time of a crash is a pedestrian. However, in an inconsistent manner, some officers classify a bicyclist as a pedestrian, presumably on the basis of the person's location in the right-of-way, or classify pedestrians (former drivers) as drivers. These errors can be seen in state crash report data.

Timeliness

Timeliness can refer to the availability of the data when they are needed by a user or can refer to the dynamic nature of the data relative to time. Given that the system of reporting in place is still primarily paper based, FDOT may not have crash data available for analysis until 6 months or more after a crash has occurred. This is a particular problem with data that are frequently updated after the initial crash report is filed, such as blood alcohol test reports and delayed fatalities. Time delays like this can limit the effectiveness of data analysis. Roadway data are particularly sensitive to timeliness, as a crash occurs at a particular point in time, whereas roadways are constantly changing. Data obtained from video logs and the RCI database usually do not correspond to the exact date of the crash; therefore, they may not reflect the actual conditions at the time of the crash. In addition, time discrepancies between the imaging of crash reports and the availability of update forms can result in inconsistencies between the TIFF image in the crash report and the data elements stored in the CAR system database.

SUMMARY OF FINDINGS

Large amounts of traffic crash data for pedestrian crashes are collected in the state of Florida; however, accessing, reviewing, and extracting the data, which are scattered among sources, present unique challenges. Although a good portion of the data is both available and accessible from state-maintained databases, these data are found to have problems that negatively affect the quality of the crash data, including deficiencies in accuracy, completeness, precision, and timeliness. The particularly error-prone data elements within state crash report records include alcohol usage, fault, speed limits, vehicle speeds, and citations. In the case of pedestrian alcohol test results, state database records were found to be in error more than half the time, thus underreporting actual alcohol usage. Both pedestrian fault and driver fault exhibited very high error rates as well. For the accurate determination of fault and crash causation, a review of detailed THI reports was found to be the single most valuable activity.

The data collected by using the improved methods described in this study served as the source for additional investigation and analysis of pedestrian crashes. It is noteworthy that the findings of high rates of pedestrian fault do not minimize the importance of providing roadway features for safe pedestrian travel. A companion study showed that in more than half of the pedestrian fatalities that involved crossing the road without a crosswalk (marked or unmarked), no protected crossings were available within 600 ft of the selected crossing location, and in almost 25% of the cases, the nearest protected crossing was more than a quarter of a mile away (*8*). The most prevalent crossing trip generator was a driveway or a minor side street (without sidewalks) ending in a T-intersection with no traffic control device controlling the major roadway. 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 Crash Statistics Report 2004: A Compilation of Motor Vehicle Crash Data from the Florida Crash Records Database.* Florida Department of Highway Safety and Motor Vehicles, Tallahassee. www.hsmv.state.fl.us/hsmvdocs/cf2004.pdf. Accessed July 31, 2006.
2. *Traffic Safety Facts 2000: Pedestrians.* Report DOT-HS-809-331. NHTSA, U.S. Department of Transportation. www-nrd.nhtsa.dot.gov/pdf/nrd-30/NCSA/TSF2000/2000pedfacts.pdf. Accessed July 31, 2006.
3. Benavente, M., M. A. Knodler, Jr., and H. Rothenberg. Case Study Assessment of Crash Data Challenges: Linking Databases for Analyzing Injury Specifics and Crash Compatibility Issues. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1953,* Transportation Research Board of the National Academies, Washington, D.C., 2006, pp. 180–186.
4. 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.
5. Siegel, S., and N. J. Castellan, Jr. *Nonparametric Statistics for the Behavioral Sciences.* McGraw-Hill Book Company, New York, 1988, pp. 284–291.
6. Altman, D. G. *Practical Statistics for Medical Research.* Chapman and Hall, New York, 1991.
7. O'Day, J. *NCHRP Synthesis of Highway Practice 192: Accident Data Quality.* TRB, National Research Council, Washington, D.C., 1993.
8. Spainhour, L. K., I. A. Wootton, J. O. Sobanjo, and P. A. Brady. Causative Factors and Trends in Florida Pedestrian Crashes. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1982,* Transportation Research Board of the National Academies, Washington, D.C., 2006, pp. 90–98.

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

The Pedestrians Committee sponsored publication of this paper.