Comparison of Methodology Approach to Identify Causal Factors of Accident Severity

Jinn-Tsai Wong and Yi-Shih Chung

Identifying the factors that significantly affect accident severity has become one of the many ways to reduce it. While many accident database studies have reported associations between factors and severities, few of them could assert causality, primarily because of uncontrolled confounding effects. This research is an attempt to resolve the issue by comparing the difference between what happened and what would have happened in different circumstances. Data on accidents were analyzed first with rough set theory to determine whether they included complete information about the circumstances of their occurrence by an accident database. The derived circumstances were then compared with each other. For those remaining accidents without sufficient information, logistic regression models were employed to investigate possible associations. Adopting the 2005 Taiwan single-auto-vehicle accident data set, the empirical study showed that an accident could be fatal mainly because of a combination of unfavorable factors instead of a single unfavorable factor. Moreover, the accidents related to rules with high support and those with low support showed distinct features.

The severity of accidents is a special concern to researchers in traffic safety because such research is aimed not only at prevention of accidents but also at reduction of their severity. One way to accomplish the latter is to identify the factors that significantly affect it. To identify the influential factors on accident severity, various approaches have been proposed. Many studies in this domain have extensively applied regression-type models (*1–6*). Some studies developed injury severity models concentrating on traffic accident records limited to a small geographic area, a particular accident type, or certain road conditions to make the domain as specific as possible so that relatively homogeneous data were obtained for identifying the significant factors. However, because of the complexity of accidents and data limitations, some inconsistent results have been seen in the literature (*7–9*).

To avoid heterogeneity problems on data analyses, some studies adopted clustering or classification methods to group accidents before evaluation of the influential factors. For example, Karlaftis and Tarko (*10*) proposed a two-stage approach to reducing area-specific heterogeneity. Hierarchical cluster analysis was applied to partition accident data so that areas with similar size and population density were appropriately grouped. In comparing their negative binomial models, they found that the separated models described the data more efficiently than the pooled model. Recently, because of their ability to accommodate abundant factors, to proceed without prior model specification, and to find nonlinear relationships between factors and severity, some statistical and soft computing methods have become popular alternatives (*11–15*).

While the statistical and soft computing classification methods have consistently reported satisfactory performance on prediction accuracy, the possible causal relationship between factors and severity has been rarely discussed. One possible reason might be the rigorous definition of causality. For example, Pearl asserted three criteria to judge causality: correlation, time sequence, and a nonspurious relationship between cause and effect (*16*). Because of the untestability of observational accident studies, causality is difficult to judge, especially for cross-sectional studies (*17, 18*). Furthermore, the factors selected by these classification methods are those with significant classifying ability on severity, which does not necessarily imply causality.

The continuous expansion of accident databases and improvement of computing ability, however, provide the opportunity to explore causality. Through control of as many affecting factors as possible, accidents can be classified into subsets with similar conditions. Therefore, a comparison of the features of these subsets would reveal the differences between what happened and what would have happened under different circumstances (*7, 18*); the technique might imply causal relationships. In addition, because an accident database can never contain sufficient factors for characterizing the occurrence of all types of accidents, it would be unreasonable to regard all accidents in a database as having complete information. Therefore, for those accidents with insufficient information, instead of soft computing classification methods, other methods could be advantageous to analyze them.

This research had three purposes: first, a systematic approach was proposed to help screen accidents that could be suitably analyzed with statistical and soft computing methods; second, the features of screened subsets were compared to identify the possible causal factors for the accidents; and third, the accidents with incomplete information were analyzed with regression methods to explore the relationships between factors and severity.

The rest of this paper is organized as follows. The methodology is proposed in next. Then, a real data set is adopted to demonstrate

J.-T. Wong, Institute of Traffic and Transportation, National Chiao Tung University, 4F, 114 Chung Hsiao W. Rd., Section 1, Taipei 100, Taiwan. Y.-S. Chung, School of Civil and Environmental Engineering, Georgia Institute of Technology, 210 Technology Circle, Savannah, GA 31407. Corresponding author: Y.-S. Chung, ychung37@mail.gatech.edu.

Transportation Research Record: Journal of the Transportation Research Board, No. 2083, Transportation Research Board of the National Academies, Washington, D.C., 2008, pp. 190–198.
DOI: 10.3141/2083-22

the proposed approach. A discussion of results follows, and conclusions are drawn in the final section.

METHODOLOGY

The research framework consists of two stages, as shown in Figure 1. The first stage was to identify the circumstances contained in an accident database. To describe the circumstances fully, all available information should be considered, such as driver characteristics, trip characteristics, vehicle information, behavioral information, and road and environmental factors. To accommodate the numerous factors, soft computing methods such as tree- or rule-based classification methods are preferred; in particular, rough set theory was adopted in this research. Interested readers can refer to Pawlak (*19*) and Pawlak and Skowron (*20*) for a thorough introduction about rough set theory. In addition, a clear tutorial about rough set theory was presented by Walczak and Massart (*21*), and Wong and Chung (*22*) showed the connections between rules from rough set theory and accident chains.

As a classification methodology, rough set theory generates rules to identify the differences among accidents. Because each rule implies the indispensable circumstances under which accidents with specific injury levels occur, the injury level will be different if one or several indispensable circumstances are different. Therefore, a comparison of the rules with high support offers the potential to understand the causes of accidents and is the focus in this study.

On the basis of classification results, it is possible to compare the rules and find potential causal factors, especially for those types of accidents that frequently appear. However, two difficulties exist. First, the available information is unable to differentiate all accidents. Some accidents under identical circumstances may lead to different results. This problem mainly results from insufficient information. Second, even if accidents could be clearly distinguished, some rules may show extremely low frequency of occurrence (the frequency of occurrence is called support in rough set theory). These low-support accidents may occur by chance (bad luck), and strong causal relationships between factors and accident consequences may not exist. Accordingly, these accidents and the corresponding rules would be inappropriate for rule comparisons. Instead, statistical analysis such as regression models would be more appropriate to catch the features of these low-support accidents, that is, the use of error terms to represent the insufficient information and the randomness. The problem then is how to distinguish between the accidents suitable for rule comparisons and those suitable for statistical analysis. The choice of the threshold should result in a satisfactory performance on postvalidity evaluations or predictions.

EMPIRICAL STUDY

Data

Taiwan 2005 single-auto-vehicle (SAV) accident data was adopted to demonstrate the feasibility of the proposed approach; in particular, accident severity was considered as the target variable. SAV accidents are those in which only a single vehicle is involved. Because no other vehicles, as well as no pedestrians, are involved, SAV accidents are simpler than multivehicle accidents and a good start for the research. The data were collected by police departments and included all the death-involved and injury-only accidents. The total number of SAV accidents, excluding invalid cases, was 3,138. The number of invalid cases was 27, which accounted for 0.86% of the total cases. These cases were invalid mainly due to the unknown attribute values of the driver characteristics. They were directly ignored in the study on the basis of their relatively small size. The collected attributes and their corresponding categories are summarized in Table 1.

Classification with Rough Set

The Taiwan 2005 SAV accident data was first analyzed with rough set theory, with which the software Rough Set Data Explorer (ROSE2) (*23, 24*) was used to generate a minimum rule set covering all objects. This analysis consisted of two steps: variable selection and rule induction. The first step was to identify the variables that were unable to differentiate the accident severity. In the analysis, four of 25 variables were identified as redundant, including pavement material, surface deficiency, signal condition, and weather condition, which may result from the following two reasons. First, their effects could be replaced by other variables. For example, the effect of the weather variable could be substituted by that of the surface condition variable because rain would result in a wet surface. It is understood that the weather condition would affect not merely surface conditions; for example, strong wind or a large snow fall would raise the difficulty on drivers' control of their vehicles. However, these weather conditions rarely occur in Taiwan. The second reason is that these redundant variables had no significant impact on accident severity. For example, 98.6% and 98.5% of the accidents were reported on roads with an asphalt pavement and on roads without surface deficiency, respectively. Therefore, the pavement material and surface deficiency variables were reported as redundant. After excluding the four redundant variables, the remaining 21 variables were considered in generating rules.

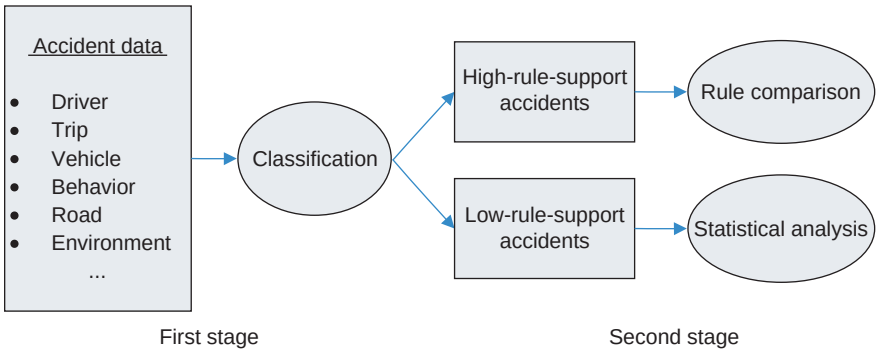

FIGURE 1 Research framework.

TABLE 1 Attribute and Category

Attribute	Category
Age	Younger (<18), young (18–35), middle-aged (36–55), elderly (>55)
Gender	Male, female
License type	Regular, occupational, other
License condition	Valid, invalid, unknown
Occupation	Student, working people, no job, unknown
Trip purpose	Necessary (working, school, business), other
Trip time	MP (07–09), DOP (09–16), AP (16–19), NOP (19–23), midnight (23–07)
Seat belt use	Fastening, not fastening, unknown
Cell phone use	Using, not using, unknown
Drinking condition	Drinking, not drinking, other
Road type	Highway, urban, rural
Speed limit (km/h)	≤ 50, 51–79, ≥ 80
Road shape	Intersection, segment, ramp or other
Pavement material	Asphalt, other, no pavement
Surface deficiency	Normal, other (e.g., holes, soft, and so on)
Surface condition	Dry, wet or other
Obstruction	Yes, no (within 15 m)
Sight distance	Good, poor (based on road design speed)
Signal type	Regular, flash, no signal
Signal condition	Normal, abnormal, no signal
Median	Island, marker, marking, none
Roadside marking	Yes, no
Weather	Sunny or cloudy, rainy, other
Illumination	With light, no light
Alignment	Straight, curved, other
Accident severity	Death involved, injury only

With 21 nonredundant explanatory variables, 315 rules were generated with rough set theory to represent the 3,138 accident cases. This study applied the most frequently used algorithm—minimum covering—to generate rules. Its aim was to generate the minimum number as well as the shortest rules to cover all accidents. Of these, 295 rules were exact rules, and 20 were approximate rules. An "exact rule" refers to a situation for which the severity of an accident could be identified under a particular circumstance. In contrast, an "approximate rule" represents a circumstance under which the accident severity could not be uniquely determined.

For the purpose of analysis, the accident cases were separated into two subsets. The choice of threshold of the rule support was determined by examining the average hit rate of the rules. With Monte Carlo simulations in which specific percentages of crashes were trained and tested, it was found that the average hit rates were increasing with the exclusion of accidents related to low support rules, especially for the minority class: fatal accidents. In particular, the increase was significant when the accidents related to the rules with a support level lower than six were dropped. Therefore, the support level of six was considered the threshold to differentiate between rules. However, the details of the simulation tests were beyond the scope of this study; therefore, they were omitted here. In summary, the accidents were divided into two groups: one consisted of accidents with a rule support of at least six, and the other consisted of accidents with approximate rules and with a rule support lower than six. These two subsets were analyzed separately.

Procedure of Rule Comparison

The subset with accidents of high rule support was adopted for rule comparisons. The comparisons consisted of two steps: the first was to find the most similar rules for each selected strong rule (i.e., a rule with support of at least six) from the remaining strong rules; the second was to check whether the accident severities were different between the selected rule and its most similar rules. In the following, an example of rule comparison was provided.

It is supposed that a rule, denoted as the selected rule, was chosen from the rule set. This rule described a particular circumstance for SAV accident occurrence: a female driver with a valid driver's license driving on a road with a low speed limit (less than 50 km/h) with seat belt fastened but without specific trip purposes. The SAV accidents under such circumstances were of the injury-only type. If the specified attributes were changed (e.g., from female to male), the result would be different (i.e., from injury only to death involved or to other). "Other" represents the accident severity of approximate rules, which can be injury only or death involved. Some condition attributes were specified, but others were not. The severity does not change even though those unspecified attributes change. For instance, whether a driver was young, middle aged, or old, the severity of the SAV accidents under the circumstance described by the selected rule would remain the same.

On the basis of the selected rule, its similar rules were searched. A "similar rule" is defined as one that has the greatest number of identical specified attributes to the selected rule. Two similar rules were found. Similar Rule 1 described the condition of a middle-aged driver with a valid driver's license, seat belt fastened, cell phone not used, without specific trip purposes driving on a road with a low speed limit (less than 50 km/h) that is equipped with roadside marking and illumination. Similar Rule 2 described the condition of a young male driver with a valid regular driver's license, seat belt fastened, without specific trip purposes driving at midnight on a straight road with a low speed limit (less than 50 km/h) and dry surface equipped with median marking but without signals.

Both similar rules had only one indispensable attribute value different from the selected one. This could be verified by expanding the unspecified attributes of the selected rule to match its similar rules. As shown in Figure 2, the attributes age, cell phone use, road shape, roadside, and illumination of the selected rule could be expanded to be identical to those of similar Rule 1 without affecting the accident severity of the selected rule. By comparing the expanded rule and similar Rule 1 (the upper right section of Figure 2), one can see that only the attribute gender was different where the expanded rule specified it as female but was unspecified in similar Rule 1. Similarly, the same expansion can be done to compare the selected rule and similar Rule 2: the attribute gender was also the only distinct one between these two rules (the lower right section of Figure 2).

Rule 1 pointed out that a male driver's accident severity was greatly reduced if he was mature (middle aged and driving without using a cell phone) and driving in a friendly road environment (with roadside markers and illumination). Rule 2 pointed out that young male drivers on in an unfriendly environment (a road not designed as safety oriented at midnight) could be fatal. This result implied that the com-

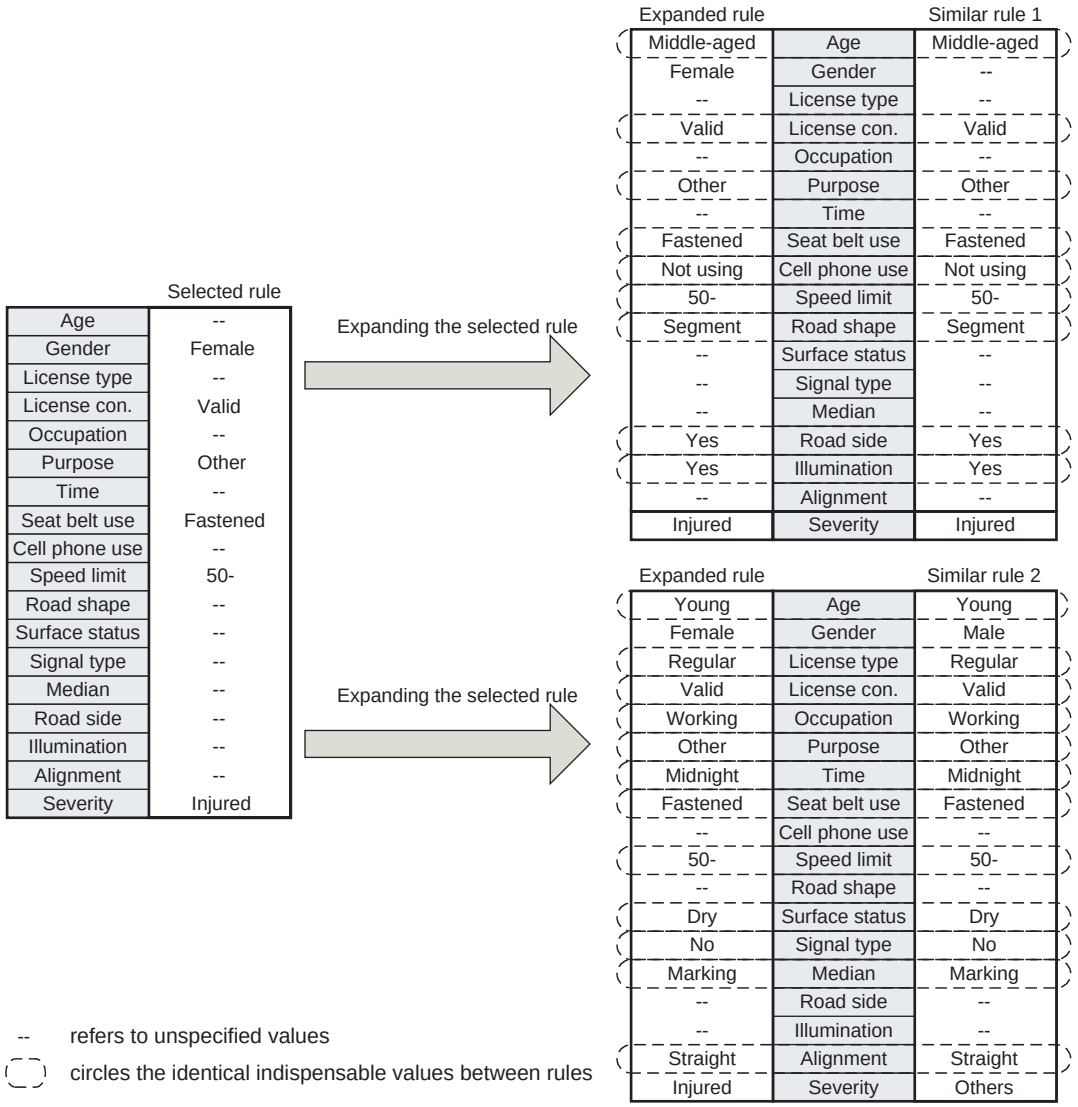

FIGURE 2 Rule comparison example.

bined attributes (age + gender + road environment) might be critical factors in changing an injury-only case to a death-involved case under a circumstance described by the selected rule.

Rule Comparison Results

Of the 315 rules, 164 of them were strong rules: 19 of those strong rules led to death-involved or other accidents, and the remaining 145 strong rules led to injury-only accidents. The following comparisons focused on the differences between (*a*) death-involved or other accidents and (*b*) injury-only accidents. In other words, the possible causal factors diverting an injury-only accident to a death-involved or other accident were examined. The rules having no similarity to injury-only rules are discussed in the following section on dissimilar death-involved or other rules, and the remaining 16 strong rules are discussed in the subsequent section on similar death-involved or other rules.

Dissimilar Death-Involved or Other Rules

There were three death-involved or other rules having no similarity to injury-only rules, as listed in Table 2. The first dissimilar rule, D1, describes the young working drivers who were drinking and might have been using cell phones driving on a curved road with poor sight distance but with lighting. While normal drivers would lower their speeds to pass a curve safely, the leading-to-death rule suggests that the corresponding driving speeds would not be low. Moreover, the curved road with poor sight distance raised the difficulty of driving. Although there were another 10 strong rules relating to curved roads and leading to injury-only cases, none of them was specified as including young drinking drivers. This might suggest that these drivers can easily misjudge the safe driving speed and cannot properly maneuver the vehicle while passing a curve with a poor sight distance.

As Table 2 shows, Rules D2 and D3 describe the corresponding death-involved accidents occurring under the conditions that the

TABLE 2 Dissimilar Strong Rules Leading to Death or Other

	Rule		
Attribute[a]	D1	D2	D3
Age	Young	—	—
Occupation	Working	—	—
Seat belt use	—	Not using	Not using
Cell	Unknown	Unknown	—
Drink	Drinking	Unknown	Unknown
Road type	—	—	Rural
Sight distance	Poor	—	—
Illumination	Yes	—	Yes
Alignment	Curved	—	—
Severity	Death	Death	Death

[a]The attributes where all the three rules were unspecified were not represented to reduce the space.

drivers were not wearing seatbelts and were possibly drinking. Fastening a seatbelt and drinking and driving have long been critical policy issues for the government of Taiwan. Violating either one, especially the latter, leads to a substantial fine. Therefore, it is expected that these two unlawful behaviors occurring at the same time, as described in Rules D2 and D3, will be rare. However, with commission of both these violations, whether combined with an unfriendly road environment or not, a death-involved case would likely occur.

Similar Death-Involved or Other Rules

There were 16 deaths involved or other rules similar to injury-only rules as listed in Table 3. Rules S1 and S2 were the ones most similar to injury-only rules; these two rules had been cited as similar rules by injury-only rules 47 and 46 times, respectively. Rule S1 illustrated the condition that young male working drivers with regular valid licenses driving with unspecified purposes and wearing seatbelts had been drinking alcohol and were driving around midnight on straight rural roads at low speed limits, on a dry surface with median markings and no signals. Although this condition describes drinking and driving behaviors, drinking itself cannot fully represent the cause shifting the accident to a fatal one. Examination of the strong rules shows some of them also related to drinking and driving behavior; however, as long as the drivers were not young people, it was not midnight, the quality of the corresponding road environment was not poor (i.e., it was an urban road, a road with a median, or a road with a higher speed limit), or the surface was not dry, the accident severity was shown to be injury only. When the driver is young, the corresponding behavior could be somewhat risky, and a more risky driving environment is usually associated with midnight driving (*25*). Moreover, a road of poor quality could not mitigate the bumping impact of an accident; and a dry surface might encourage fast driving, especially under low traffic

TABLE 3 Strong Rules Leading to Death or Other

	Rule							
Attribute	S1	S2	S3	S4	S5	S6	S7	S8
Age	Young	Young	Young	Young	Young	Young	Young	Young
Gender	Male	Male	—	Male	Male	Male	—	—
License type	Regular	Regular	Regular	Regular	Regular	—	Regular	—
License condition	Valid	—	—	—	—	—	—	—
Occupation	Working	Working	—	Working	—	Working	Working	Working
Purpose	Other	Other	Other	—	—	—	—	—
Time	Midnight	Midnight	Midnight	Midnight	—	Midnight	—	DOP
Protection	Using	Using	—	Using	Unknown	Unknown	Unknown	Using
Cell	—	—	Not using	Not using	—	Unknown	—	—
Drink	Drinking	—	Drinking	Drinking	Unknown	—	Unknown	—
Road type	Rural	Rural	Urban	Highway	Rural	—	Rural	Highway
Speed (km/h)	≤ 50	≤ 50	—	—	≤ 50	—	—	≥ 80+
Road shape	—	Segment	Segment	Segment	Segment	Segment	Segment	Segment
Surface status	Dry	Dry	—	Dry	—	—	—	Dry
Obstruction	—	—	No	—	—	No	—	—
Sight distance	—	—	—	—	Good	Good	Good	—
Signal type	No	No	—	—	No	—	No	—
Median	Marking	Marking	Island	—	—	Marking	—	Island
Roadside	—	Yes	Yes	—	—	Yes	Yes	—
Illumination	—	Yes	Yes	—	Yes	—	—	—
Alignment	Straight	Straight	Straight	—	—	—	—	—
Severity	Other	Other	Other	Other	Death	Death	Death	Other
Similarity[a]	47	46	18	16	11	7	7	7

[a]Similarity referred to the number of rules which were similar to this rule but led to injury-only crashes.

(midnight on rural roads). Therefore, the combined unfavorable factors led to death-involved accidents.

As stated earlier, Rule S2 illustrated a condition very similar to that for S1. These two rules were almost identical except that Rule S2 did not specify the drinking behavior but specified that the corresponding road environment may encourage fast driving with low traffic and good sight distance (driving around midnight along a straight rural road with illumination and roadside markings). Though the corresponding driver was not specified as drinking, the possibly faster driving behavior also led to death-involved accidents.

In contrast to the first two rules, Rules S3 and S4 illustrate accidents occurring on high-quality roads (highways or urban roads with medians). The driving speeds on these roads are usually high, especially on highways with a minimum speed of 80 km/h. Under conditions that combine high driving speeds with the impaired maneuvering skills, as well as lower situational awareness due to drinking, once an accident occurs, a death-involved case is expected. When compared with their similar rules, these death-involved cases could be merely injury only if (*a*) the driver was not a young male (that is, middle-aged, elderly, or female instead), (*b*) the road was narrower (an urban road without roadside markings), or (*c*) the road did not mislead drivers to drive at an inappropriately high speed. Having either one of the factors could reduce driving speeds or make the drivers drive more carefully.

Rules S5, S6, and S7 describe conditions for accidents that occurred on rural roads with low speed limits or in a low-traffic environment (midnight), except that the trip purposes were unspecified and the drinking conditions and seatbelt usages were unknown. A review of their similar rules shows that, all else being equal, accidents under Rules S5, S6, and S7 became injury only if the driver did wear a seatbelt or if the driver was not drinking. This situation addresses the effect of injury prevention through the wearing of a seatbelt and the avoidance of both deteriorated maneuvering skills and lower situational awareness due to drinking.

Rule S8 describes young working people wearing seatbelts and driving on highway segments with a dry surface during day off-peak periods. When compared with the similar rules, all else being equal, the accidents became injury-only cases if (*a*) the driver was not drinking, (*b*) the driver owned an occupational or military driver's license, or (*c*) the trip time was during afternoon peak hours. Only soldiers in charge of driving can obtain a military driver's license. Therefore, in an environment of high-speed driving, drivers with occupational or military licenses are expected to be more capable of avoiding fatal accidents than normal drivers once an accident occurs. Moreover, the traffic flow during peak hours is more dense than that during off-peak hours; consequently, the corresponding driving speed is expected to be lower. Once an accident occurs, the severity should be lower.

Rule S9, similar to S8, describes the accidents that occurred on highways but with drivers specified as males rather than young; moreover, the trip time was around midnight rather than at off-peak periods during the day. When compared with its similar rules, accidents under Rule S9 could become less severe if the trip time was during afternoon peak periods. The denser traffic during peak hours might restrict driving speed. Even though the drivers could be high risk (young or male drivers), the environment might limit

S9	S10	S11	S12	S13	S14	S15	S16
—	—	—	—	Young	Middle	Young	—
Male	Male	Male	—	Male	—	—	—
—	Regular	Regular	—	—	—	—	—
—	—	—	Valid	—	—	Valid	—
Working	—	Working	Working	Working	—	—	—
—	Other	—	—	—	—	—	—
Midnight	—	—	NOP	—	—	Midnight	Midnight
—	—	Unknown	—	—	—	—	Unknown
Unknown	Unknown	—	Unknown	Unknown	Unknown	Unknown	—
—	—	Unknown	Drinking	Unknown	Unknown	Unknown	—
Highway	—	—	—	—	Rural	—	—
—	—	51–79	≤ 50	—	≤ 50	51–79	51–79
—	—	Segment	—	—	Segment	—	Other
—	—	—	—	—	—	—	—
—	No	—	—	—	—	—	—
—	Poor	Good	—	Good	—	—	—
—	—	—	—	—	—	—	—
Island	—	—	—	—	—	Island	Island
—	—	—	—	No	—	—	—
—	No	—	—	Yes	—	—	—
—	—	—	—	—	—	—	—
Death	Death	Death	Death	Death	Death	Death	Death
3	3	3	2	1	1	1	1

their driving speeds, and the corresponding accidents might not be fatal.

Rule S10 describes regularly licensed male drivers driving on poorly sighted roads without any obstructions. When compared with its similar rules, all else being equal, the accidents under Rule S10 could be less severe if there were obstructions on the roads. "Obstructions" are defined as any obstacles within 15 m of a crash. This distance is much less than the defined safe sight distance, which is 45 m under normal 40 km/h driving speed, and a driver then might spot the obstacles and lower his or her driving speed. In contrast, driving at relatively high speeds by male drivers, even though the road has a poor sight distance, results in a fatal accident.

Rule S11 describes regularly licensed working people driving on a road with a medium speed limit and good sight distance. Its similar rules suggest that these accidents could be less severe if the drivers were not drinking. Similarly, accidents under Rules S12 and S13 would be less severe if the drivers were not using cell phones or not drinking and driving. The accidents under the same driving environment described by Rule S15 were less severe if the drivers were the elderly, who are usually considered to be at lower risk than young drivers. Even on a road encouraging fast driving (medium speed limit with median), elderly drivers might drive carefully and maintain a reasonable driving speed, while young drivers might not.

The information provided by the remaining rules, S14 and S16, is relatively vague because most attributes were unspecified and all the behavioral attributes were either unspecified or unknown. Moreover, the associated similar rules were different in behavioral attributes. Therefore, it is difficult to tell the differences between the selected rules and their associated similar rules.

Logistic Regression Analysis

Different from the accident cases with strong causal relationships, the 363 accidents associated with the weak support rules or the approximate rules were analyzed with regression methods to investigate possible associations between factors and to extract the variations due to insufficient information. In particular, binary logistic regression models were adopted. The model structure was revised from the one proposed by Kim et al. (*26*), in which the accident severity was affected by driver characteristics, trip characteristics, behavioral factors, environmental factors, and interactions between driver and behavioral factors. Backward elimination was applied to select variables.

The reference severity was injury only, and the estimation results are summarized in Table 4. The estimated Hosmer–Lemeshow *p*-value was .293 (>.100), which indicated that the goodness of fit was acceptable. The final variables included age, trip time, signal type, surface status, median, roadside marking, and the interaction between age and drinking. The results showed that accidents with rarely occurring patterns and those with frequently occurring patterns were different. Young drivers were less likely to be involved in a death-involved case provided that they were not drinking. Yet, under the condition that young drivers were drinking, they would be more likely to be involved in a death-involved case. Moreover, the accidents occurred around midnight (compared with other time periods) were less likely to be involved in a death-involved accident. These two results contradicted the results of the previous section that young drivers and midnight accidents were death prone, which may imply distinct features between these two types of drivers.

Furthermore, accidents occurring on roads having a dry surface (compared with wet or other surface conditions) and with roadside marking (compared with roads without roadside markings) were less likely to be death-involved accidents. In contrast, those accidents that occurred on roads with warning flash signals (compared with no signals) and with median markers (compared with no medians) were more likely to be death-involved accidents. A road with warning flash signals indicates possible traffic conflicts within the area, and the signals warn drivers to pay attention. In addition, a road with median markers implies that this section of the road is rather dangerous, and the markers warn the drivers not to drive across the centerline. These results suggested that a better road environment seems to help prevent such death-involved accidents.

TABLE 4 Logistic Regression Estimation Result[a]

Parameter	Estimate	*P*-Value	Odds		
			Odds Ratio	95% Wald Confidence Interval	
Intercept	2.841	<.0001[b]	17.124	6.426	49.844
Age (young vs. middle or old)[c]	–1.099	0.002[b]	0.333	0.164	0.662
Trip time (midnight vs. other)	–0.786	0.004[b]	0.456	0.267	0.777
Signal type (regular vs. none)	–0.548	0.216kk	0.578	0.243	1.377
Signal type (flash vs. none)	1.583	0.040[b]	4.871	1.072	22.137
Surface status (dry vs. other)	–0.942	0.009[b]	0.390	0.193	0.787
Median (island vs. none)	0.448	0.336kk	1.565	0.628	3.899
Median (marker vs. none)	1.452	0.015[b]	4.271	1.320	13.821
Median (marking vs. none)	0.186	0.690kk	1.204	0.484	2.997
Roadside marking (yes vs. no)	–1.191	0.000[b]	0.304	0.157	0.589
Age*Drink (drinking vs. not drinking)	0.716	0.036[b]	2.047	1.047	4.002
Age*Drink (unknown vs. not drinking)	1.196	0.015[b]	3.308	1.267	8.635

[a]Goodness-of-fit test: Hosmer–Lemeshow *p*-value = 0.2933.
[b]0.05 significance level.
[c]The latter term in brackets refers to the reference.

DISCUSSION OF RESULTS

Confounding Effects

Finding causal factors on safety in observational studies, especially in cross-section studies, is an unresolved issue (*17*). The main difficulty lies in the numerous confounding effects while comparisons are done. Consequently, if the majority of these attributes is not well controlled, the analysis results would be biased.

As an attempt to resolve this issue, this research identified the possible causal factors by comparing the differences between entire accident patterns instead of estimating the marginal effects of each attribute. From the rough set analysis, the accident data were separated into two subsets: one contained the accidents that could be fully described by the on-hand information and consisted of a certain number of accidents that represented the possible existence of causality; the other contained the remaining accidents. The rules, derived from the rough set analysis, were then compared with each other. The comparison design was used to find the most similar rules for each rule and to examine the differences. This design allowed the control of many confounding factors as possible and partially revealed the differences between what happened and what would have happened had the circumstances in question been different.

Because the factors were found by comparing the complete rules, it is obvious that the comprehensiveness of on-hand data determines the extent to which the confounding effects are controlled. In this empirical study, 25 attributes were considered. These attributes were presumed to have impact on accident severity and examined with rough set theory to determine whether some of them were redundant. Basically, more information is welcome in such research, provided that it is relevant to the decision attribute. Moreover, there is theoretically no limitation in the attributes that rough set theory can adopt as long as the computational time is tolerant. Yet, it inclusion of attributes with similar meanings could produce unnecessary rules and impede interpretations. For example, one could obtain two rules with all other things being equal except that one rule specifies the road type as a freeway and the other specifies a high speed limit, which could only occur on freeways. There is no difference between these two conditions in the real world. A careful selection of the entry attributes could avoid such redundancy.

Internal Validity of Approach

In the empirical study, 19 strong rules representing fatal accidents were found. Although these rules indicate diverse conditions, they retain a common feature that the fatal consequences usually result from the combination of unfavorable factors rather than the marginal effect of a single factor. In contrast, one or several critical factors could be found from the rules for injury-only accidents that, once they were removed from the process of accident occurrence, then the fatal accidents could be avoided. For example, all other things being equal, a young driver is replaced by an elderly driver, or a midnight trip is replaced by a peak-hour trip. This difference addresses the fact that an accident may not occur if one or more undesirable activities in the process of accident occurrence were removed (*22*). Moreover, the 19 rules tend to suggest that the drivers involved in such accident rules are high-risk drivers not only because they are young, male, or less experienced but because they are drinking, wandering on roads around midnight, overestimating their own driving abilities, and underestimating the possible dangers hidden in the environment.

In relation to the accidents described with insufficient information or with weak causality, the estimation results indicate distinct features from those with strong causality. On the basis of logistic regression analysis, the accidents became fatal when drivers were not young and when trip time was not around midnight, although drinking and driving still played a key role in the occurrence of a fatal accident. Moreover, most of the significant variables were contributed from environmental factors. These differences may suggest that the two types of accidents were from extremely different types of drivers. The young drivers associated with the rules with strong support were those who considered to have possibly risky driving behavior in past studies. However, those associated with the rules with weak support were different. The latter population might recognize themselves as novice drivers and would drive carefully.

CONCLUSION

This paper proposed a rule-based approach for identifying possible causal factors from accident databases. Through a comparison of the differences between rules leading to injury-only accidents and those leading to fatal accidents, the causal factors leading to more serious accidents could be found. Moreover, the investigation of the circumstances under which unfavorable factors would lead to a more serious crash would be helpful in understanding the causality of the fatal-accident occurrence. The empirical study demonstrates the feasibility of the proposed approach. Instead of a single factor, the combinations of unfavorable factors would be the causes leading to fatal accidents; these factors included the drivers being young, male, or less experienced and their behaviors of drinking, wandering on roads around midnight, and overestimating their own driving capabilities and underestimating the possible dangers hidden in the environment. Furthermore, distinct features were shown between the accidents related to rules with high support and those with low support. A better road environment would be helpful to preventing fatal accidents for the latter kind of drivers but not necessarily for the former kind.

Although this approach allows the control of all relevant factors, it does not mean that the findings under this approach must be the true causal factors. The primary reason is the limited information provided by accident databases. Accidents are observable only after they have occurred. Some information is thus difficult to obtain, especially for fatal accidents. Consequently, the uncontrolled confounding factors should be carefully taken into account in ascertaining the findings. Furthermore, experimental designs for exploring driving behaviors would be helpful to complement the this shortcoming. In particular, these designs could be based on the interested rules; for example, the most cited rule leads to fatal accidents. Because a rule contains rich information, the corresponding experimental design would be specific and effective.

ACKNOWLEDGMENTS

This research was supported by a grant from the National Science Council of Taiwan. The authors thank the anonymous referees for their helpful suggestions and comments.

REFERENCES

1. Al-Ghamdi, A. S. Using Logistic Regression to Estimate the Influence of Accident Factors on Accident Severity. *Accident Analysis and Prevention*, Vol. 34, No. 6, 2002, pp. 729–741.
2. Abdel-Aty, M. Analysis of Driver Injury Severity Levels at Multiple Locations Using Ordered Probit Models. *Journal of Safety Research*, Vol. 34, No. 5, 2003, pp. 597–603.

3. Bedard, M., G. H. Guyatt, M. J. Stones, and J. P. Hirdes. The Independent Contribution of Driver, Crash, and Vehicle Characteristics to Driver Fatalities. *Accident Analysis and Prevention*, Vol. 34, No. 6, 2002, pp. 717–727.
4. Chandraratna, S., N. Stamatiadis, and A. Stromberg. Crash Involvement of Drivers with Multiple Crashes. *Accident Analysis and Prevention*, Vol. 38, No. 3, 2006, pp. 532–541.
5. Kim, S., and K. Kim. Personal, Temporal and Spatial Characteristics of Seriously Injured Crash-Involved Seat Belt Non-Users in Hawaii. *Accident Analysis and Prevention*, Vol. 35, No. 1, 2003, pp. 121–130.
6. Quddus, M. A., R. B. Noland, and H. C. Chin. An Analysis of Motorcycle Injury and Vehicle Damage Severity Using Ordered Probit Models. *Journal of Safety Research*, Vol. 33, No. 4, 2002, pp. 445–462.
7. Davis, G. A. Possible Aggregation Biases in Road Safety Research and a Mechanism Approach to Accident Modeling. *Accident Analysis and Prevention*, Vol. 36, No. 6, 2004, pp. 1119–1127.
8. Elvik, R. Assessing the Validity of Road Safety Evaluation Studies by Analyzing Causal Chains. *Accident Analysis and Prevention*, Vol. 35, No. 5, 2003, pp. 741–748.
9. Elvik, R. Laws of Accident Causation. *Accident Analysis and Prevention*, Vol. 38, No. 4, 2006, pp. 742–747.
10. Karlaftis, M. G., and A. P. Tarko. Heterogeneity Considerations in Accident Modeling. *Accident Analysis and Prevention*, Vol. 30, No. 4, 1998, pp. 425–433.
11. Clarke, D. D., R. Forsyth, and R. Wright. Machine Learning in Road Accident Research: Decision Trees Describing Road Accidents During Cross-Flow Turns. *Ergonomics*, Vol. 41, No. 7, 1998, pp. 1060–1079.
12. Chang, L. Y., and H. W. Wang. Analysis of Traffic Injury Severity: An Application of Non-Parametric Classification Tree Techniques. *Accident Analysis and Prevention*, Vol. 38, No. 5, 2006, pp. 1019–1027.
13. Delen, D., R. Sharda, and M. Bessonov. Identifying Significant Predictors of Injury Severity in Traffic Accidents Using a Series of Artificial Neural Networks. *Accident Analysis and Prevention*, Vol. 38, No. 3, 2006, pp. 434–444.
14. Karlaftis, M. G., and I. Golias. Effects of Road Geometry and Traffic Volumes on Rural Roadway Accident Rates. *Accident Analysis and Prevention*, Vol. 34, No. 3, 2002, pp. 357–365.
15. Sohn, S. Y., and H. Shin. Pattern Recognition for Road Traffic Accident Severity in Korea. *Ergonomics*, Vol. 44, No. 1, 2001, pp. 107–117.
16. Pearl, J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000.
17. Hauer, E. Cause and Effect in Observational Cross-Section Studies on Road Safety. Presented at 85th Annual Meeting of the Transportation Research Board, Washington, D.C., 2006.
18. Hauer, E. Observational Before–After Studies in Road Safety. In *Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety*. Pergamon Press, Oxford, United Kingdom, 1997.
19. Pawlak, Z. Rough Sets. *International Journal of Computer and Information Science*, Vol. 11, No. 5, 1982, pp. 341–356.
20. Pawlak, Z., and A. Skowron. Rudiments of Rough Sets. *Information Sciences*, Vol. 177, No. 1, 2007, pp. 3–27.
21. Walczak, B., and D. L. Massart. Rough Sets Theory. *Chemometrics and Intelligent Laboratory Systems*, Vol. 47, No. 1, 1999, pp. 1–16.
22. Wong, J.-T., and Y.-S. Chung. Rough Set Approach for Accident Chains Exploration. *Accident Analysis and Prevention*, Vol. 39, No. 3, 2007, pp. 629–637.
23. Grzymala-Busse, J. W. LERS—A System for Learning from Examples Based on Rough Sets. *Handbook of Applications and Advances of the Rough Sets Theory*, Kluwer Academic Publishers, Dordrecht, Netherlands, 1992.
24. Grzymala-Busse, J. W., and P. Werbrouck. On the Best Search Method in the LEM1 and LEM2 Algorithms. In *Incomplete Information: Rough Set Analysis* (E. Orlowska, ed.), Physica-Verlag, New York, 1998, pp. 75–91.
25. Lin, M., and K. T. Fearn. The Provisional License: Nighttime and Passenger Restrictions—A Literature Review. *Journal of Safety Research*, Vol. 34, No. 1, 2003, pp. 51–61.
26. Kim, K., L. Nitz, J. Richardson, and L. Li. Personal and Behavioral Predictors of Automobile Crash and Injury Severity. *Accident Analysis and Prevention*, Vol. 27, No. 4, 1995, pp. 469–481.

The Safety Data, Analysis, and Evaluation Committee sponsored publication of this paper.