9-2008

Development of Crash-Severity-Index Models for the Measurement of Work Zone Risk Levels

Yingfeng Li
Texas A&M University

Yong Bai
Marquette University, yong.bai@marquette.edu

Accident Analysis & Prevention, Vol. 40, No. 5 (September 2008): 1724-1731. DOI.

Development of Crash-Severity-Index Models for the Measurement of Work Zone Risk Levels

Yingfeng Li
Department of Civil, Environmental, and Architectural Engineering, The University of Kansas, Lawrence, KS, USA
Texas Transportation Institute, San Antonio, TX, USA
Yong Bai
Department of Civil, Environmental, and Architectural Engineering, The University of Kansas, Lawrence, KS, USA

Abstract

Highway work zones interrupt regular traffic flows and create safety problems. Improving safety without sacrificing the main function of highways is a challenging task that traffic engineers and researchers have to confront. In this study, the concept of using crash severity index (CSI) for work zone safety evaluation was proposed and a set of CSI models were developed through the modeling of work zone crash severity outcomes. A CSI is a numerical value between zero and one that is estimated from given work zone variables. It is interpreted as the likelihood of having fatality/fatalities when a severe crash occurs in a given work zone. The CSI models were developed using a three-step approach. First, a wide range of crash variables were examined in a comprehensive manner and the significant risk factors that had impact on crash severity were selected. Second, the CSI models were developed using logistic regression technique by incorporating the selected risk factors. Finally, the developed models were validated using the recent crash data and their ability in assessing work zone risk levels were analyzed. Results of this study showed that CSI models can provide straightforward measurements of work zone risk levels.

Keywords
Crash, Highway, Risk, Safety, Work zone

1. Introduction
As the highway system ages, government agencies have to allocate a greater percentage of their funding on preserving, expanding, and enhancing existing highway networks. Work zones on the highway system interrupt regular traffic flows and create safety problems. Improving safety without sacrificing the main function of highways has become a challenging task that traffic engineers and researchers have to confront.

Work zone safety can be affected by combinations of various risk factors and some combined effects might not be fully recognized during work zone designs. Understanding risks discovered from work zone crash data analyses is a key step towards lowering risk levels and preventing the occurrence of severe crashes. In this study, the concept of the crash severity index (CSI) was proposed for the evaluation of risk levels in work zones. A CSI is designed to be a numerical value between zero and one that can be estimated from given work zone risk factors. It is interpreted as the likelihood of having fatality/fatalities when a severe crash occurs in a given work zone. When quoted hereafter, severe crashes refer to crashes involving fatality/fatalities (i.e., fatal crashes) or injury/injuries (i.e., injury crashes) of either passengers or drivers of the involved vehicles. In this study, the CSI models were developed through the modeling of work zone crash severity outcomes based on the work zone fatal and injury crash data in Kansas.

A CSI reflects the risk level of a given work zone assuming that the work zone will have a high risk level for travelers if the likelihood of having fatality/fatalities in a severe crash is high. To develop the CSI models, chi-square statistics and Cochran–Mantel–Haenszel (CMH) statistics were first utilized to identify the significant risk factors. The logistic regression method was then deployed to develop the models. CSI models provide straightforward measurements of work zone risk levels based on a wide range of variables that may contribute to severe crashes. Traffic engineers can use the developed models to assess the risk level for either an existing work zone or a newly proposed work zone, which provides an opportunity to develop safety countermeasures to eliminate or mitigate the risks for the traveling public.

2. Literature review

The logistic regression technique was selected for the CSI model development in this study. Logistic regression models are direct probability models that have no requirements on the distributions of the explanatory variables or predictors (Harrell, 2001). This technique is more flexible and more likely to yield accurate results in traffic crash analyses where the safety impact of contributing factors needs to be quantified. In addition, logistic regression models generate outcome values between zero and one, which makes this statistical method ideal for developing models to estimate numerical outcomes with specified ranges.

The significance of logistic regression technique in the analysis of traffic safety has been recognized for years. Hill (2003) and Li and Bai (2006) utilized this technique in the analysis of work zone fatal crashes to quantify the effectiveness of traffic control devices. The technique was also used to model the relationships between crashes severity and wide ranges of crash variables. Lu et al. (2006) utilized logistic regression to develop models to predict the severity of median crossover crashes in Wisconsin. Chang and Yeh (2006) used the logistic regression in their analysis of fatality risk factors for motorcyclists in Taiwan. The logistic regression was deployed by Kim et al. (2000) in their analyses of alcohol impact on motorcycle crashes. In their analyses, a logistic regression model was developed to explain the likelihood of an alcohol-related motorcycle crash as a function of rider characteristics and environmental and temporal factors.

Other similar methods were also used in previous crash severity analyses. Dissanayake and Lu (2002) developed a set of sequential binary logistic regression models to analyze the contributing factors and predict the crash severity of single-vehicle fixed-object crashes involving young drivers. The researchers utilized the SAS software package to develop the regression models that took into account crash factors such as gender, driver impairment, and geometric conditions of crash locations. Ouyang et al. (2002) developed a simultaneous binary logit model to address the relationships between injury severity outcomes and various crash factors involved in car–truck collisions.

In summary, literature search showed that the logistic regression has been applied to several crash severity analyses, as briefly reviewed above. However, the relationships between crash severities and multiple risk factors in highway work zones have not been fully explored. The concept of using CSI to evaluate the driving risk levels in existing or proposed highway work zones was not found in previous publications either.

3. Data description

The crash data used for CSI model development contained 85 fatal crashes between 1998 and 2004, and 604 injury crashes between 2003 and 2004 in Kansas highway work zones. The crash data were originally obtained from the Kansas Department of Transportation (KDOT) database. The KDOT database included three levels of crash severity including fatal (i.e., crashes involved fatality/fatalities), injury (crashes involved injury/injuries only), and property-damage-only (crashes without injury or fatalities). For this study, only fatal and injury crashes were analyzed. The original format of the data was that a single crash was frequently described in text in multiple data rows because of multiple vehicles, traffic control devices, or contributing factors involved. This data format could not be directly utilized for computer-aided analyses using software such as SAS. Thus, the format of crash data has to be changed using the following two steps. First, at-fault drivers were identified and their characteristics were compiled along with other crash information into spreadsheets where each crash was described in a single data row. Then, for the cases with missing or unclear information, the original crash reports, including detailed crash scene descriptions and sketches, were examined to ensure the data accuracy.

The collected crash information was organized into five categories. Each category included various crash variables with specific observations. Each observation was assigned with a number, as shown in Table 1. Some observations were combined to form more general observation groups so that the frequencies of the cross-categorized observations were increased. The increased data frequencies would minimize the errors caused by data sparseness in statistical tests and logistic regression. Some major traffic control methods and dominant driver errors associated with the crashes were also included as crash variables and their values were shown in Table 2.

Table 1. Data categories and variables

Category	**Variable**	**Observation**	**Assigned value**
Driver at fault[a]	Age	15–19	1
		20–24	2
		25–34	3
		35–44	4
		45–54	5
		55–64	6
		≥65	7
	Gender	Male	1
		Female	2
Time	Time of day (h)	6:00–10:00	1
		10:00–16:00	2
		16:00–20:00	3
		20:00–6:00	4
	Day of week	Monday	1
		Tuesday	2
		Wednesday	3
		Thursday	4
		Friday	5
		Saturday	6
		Sunday	7
Environmental conditions	Light condition	Good condition i.e., daylight	1
		Fair conditions including dawn, dusk, and dark with streetlights	2
		Poor condition i.e., dark without streetlights	3
		Other unfavorable light conditions	4
	Weather condition	Good condition i.e., no adverse conditions	1
		Poor conditions including rain, mist, drizzle, sleet, snow, fog, smoke, strong winds, blowing dust or sand, freezing	2

		rain, rain and fog, rain and wind, sleet and fog, snow and winds, and other	
	Road surface condition	Good condition i.e., dry surface	1
		Fair conditions including wet, mud, dirt, sand, and debris	2
		Poor conditions including snow, slush, ice, and snow packed	3
Road conditions	Road class	Interstates and other freeways and expressways	1
		Other principal arterials and minor arterials	2
		Low-classification roads including major collectors, minor collectors, and local roads	3
	Road character	Straight and level	1
		Straight on grade	2
		Curve and level	3
		Curve on grade	4
		Other geometric alignments	5
	Number of lanes	Actual number of the traffic lanes in two directions	–
	Speed limit (mph)	≥61	1
		51–60	2
		41–50	3
		≤40	4
	Crash location	Non-intersection areas	1
		Intersection or Intersection related areas	2
		Other areas including interchange areas, crossover areas, and other	3
	Surface type	Concrete	1
		Blacktop	2
		Other	3
	Road special feature	No special feature impact	0
		Impacted by special features including bridge, overhead bridge, railroad bridge, railroad crossing, interchange, ramp, and other	1
	Area information	Urban area	1
		Rural area	2
Crash information	Vehicle body type	Truck[b] involved	1
		Non-truck involved	2
	No. of vehicles	Actual number of the vehicles involved in a crash	–

[a]Driver at fault was the person who caused a crash according to an accident report. For a single-vehicle crash case, the driver of the crash vehicle was automatically considered as the driver at fault.
[b]Trucks include single large trucks, truck and trailers, tractor-trailers, and buses.

Table 2. Traffic control and driver error variables

Category	Variable	Variable values
Traffic control	None or inoperative	0 (not present); 1 (present)
	Officer or flagger	0 (not present); 1 (present)
	Stop sign/signal	0 (not present); 1 (present)
	Flasher	0 (not present); 1 (present)
	No-passing zone	0 (not present); 1 (present)
	Center/edge lines	0 (not present); 1 (present)
Driver error	No driver error	0 (not present); 1 (present)
	Drug or alcohol impairment	0 (not present); 1 (present)
	Disregarded traffic signs, signals, and markings	0 (not present); 1 (present)
	Exceeded posted speed limits or too fast for conditions	0 (not present); 1 (present)
	Following too closely	0 (not present); 1 (present)
	Inattentive driving[a]	0 (not present); 1 (present)

[a]Inattentive driving includes such errors on the KDOT accident reports as “fell asleep,” “inattention,” “other distraction in or on vehicle,” “distraction-cell phone,” and “distraction-other electronic devices.”

4. Development of work zone crash-severity-index models

A set of CSI models were developed based on the information of severe work zone crashes involving injuries and fatalities. The procedure of model development included three steps. First, the risk factors in work zones that had impact on crash severity were determined based on the collected crash data. Second, a set of CSI models were developed by incorporating these risk factors using the logistic regression technique. Finally, the predictability of the developed models was validated using the most recent work zone crash data.

The collected crash data were divided into two groups. The dataset used for risk factor determination and model development had a total of 334 severe work zone crashes including 67 fatal crashes between 1998 and 2003 and 267 injury crashes in 2003. Adding the additional fatal crashes (1998–2002) in the model development dataset enriched the fatal crash information and thus increased model accuracy, especially for estimating CSIs at high risk level (i.e., a risk level at which fatal crashes may occur). The dataset for model verification included 355 severe crashes in year 2004 in Kansas highway work zones, among which 18 were fatal crashes and 337 were injury crashes.

4.1. Work zone risk factor determination

The determination of risk factors associated with work zone crash severity was a critical step towards developing CSI models with high accuracy and predictability. The determination process involved an examination of 29 work zone crash variables. Some of the variables may have negligible impact on the crash severity. These variables should be abandoned because incorporating them in the CSI models might not only complicate the models, but also lower their accuracies. Although most of the crash variables were mutually independent, some variables were associated with others and certain combinations of these variable pairs may interactively affect the crash severity. Thus, identifying the risk factors that both individually and interactively affect work zone crash severity became critical.

Chi-square statistics and Cochran–Mantel–Haenszel (CMH) statistics were employed to ensure the accuracy of risk factor identification. As shown in Fig. 1, the identification procedure included the following three steps and through which 18 out of 29 variables were selected as risk factors as listed in Table 3.

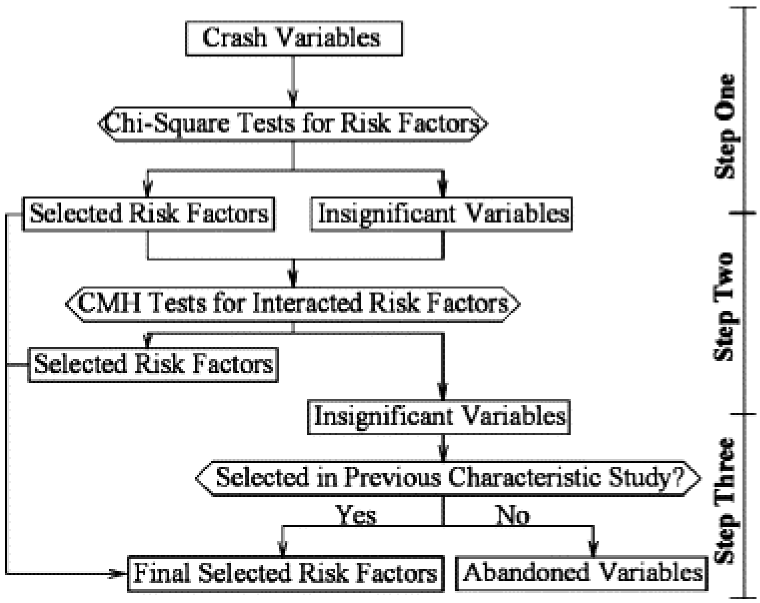

Fig. 1. Risk factor selection flowchart.

Table 3. Selected work zone risk factors

No.	**Risk factor**	**Abbr.**	**Selection step**
1	Age	AG	First step
2	Light condition	LC	First step
3	Vehicle type	VT	First step
4	Road class	RC	First step
5	Road character	RCH	First step
6	Number of lanes	LN	First step
7	Speed limit	SL	First step
8	Surface type	SUR	First step
9	None/inoperative traffic control	NTC	First step
10	Flagger	FL	First step
11	Stop sign/signal	ST	First step
12	Disregarded traffic control	DTC	First step
13	Following too close	FC	First step
14	Crash time	CT	Second step
15	Special feature	SF	Second step
16	Area information	AI	Second step
17	Alcohol/drug impairment	AL	Third step
18	Exceeded posted speed limits or too fast for conditions	SP	Third step

Step 1
The variables that are statistically associated with the crash severity were selected first as risk factors through chi-square statistics. Pearson chi-square and likelihood ratio chi-square tests were utilized in this step. A variable was selected when at least one of the two tests supported its relationship with the crash severity (i.e., a *p*-value less than or equal to the 0.1 level of significance).

Step 2
The insignificant variables from the previous step were further examined by CMH statistics at 0.1 level of significance to detect those that affect work zone crash severity interactively with certain selected risk factors. The direct impact of these variables may not strong enough to be statistically detected through chi-square tests. CMH statistics test the relationships between initially unselected variables and the crash severity variable in a three-way contingency table by controlling the selected risk factors. Some previous applications of CMH statistics in crash data exploration can be found in Chirsa-Chavala and Mak (1986) and Chen and Jovanis (2000). The significant variables supported by CMH statistics in this step were selected as risk factors. The CMH statistics used in this study included the nonzero correlation statistic, the row mean scores statistic, and the general association statistic.

Step 3
To identify all potential risk factors, the results of the characteristic comparisons between fatal and injury crashes were examined. Characteristic comparisons between fatal and injury were conducted in a previous project by authors and some of the results were utilized for this study directly. Risk factors that were identified based on the previous comparison study yet not detected in the steps 1 and 2 were also selected. As unveiled in the previous comparison study, factors such as alcohol/drug impairment and too fast for conditions/speeding had significant impact on crash severity outcomes but were not selected in the first two steps (Li, 2007).

4.2. Development of CSI models

Based on the selected risk factors, two groups of CSI models were developed using logistic regression including two driver-independent CSI (DI-CSI) models as one group and two driver-dependent CSI (DD-CSI) models as the other group. The DI-CSI models only included the risk factors that described the travel conditions in highway work zones. These models can be used to estimate the driving risks in work zones without knowing human factors. The estimated CSI values reflect the risk levels of proposed or existing highway work zones for traveling public. The DD-CSI models, on the other hand, are associated with particular drivers by including not only the risk factors related to work zones but also those risk factors that only certain drivers may possess such as demographic characteristics and driver errors.

4.2.1. Developed DI-CSI models

A DI-CSI model, or the comprehensive DI-CSI model, was first generated using SAS which included all driver-independent risk factors, as listed in Eq. (1). Table 4 lists the estimated variable coefficients and related statistical results for the comprehensive DI-CSI model. The Wald chi-square statistic was used to test the variable significance for the logistic regression models. SAS also outputted the values of three statistics for assessing the goodness-of-fit for the logistic regression model including the AIC statistic, the SC statistic, and the −2 log likelihood statistic. The log likelihood statistic was used to test the global null hypothesis that all the parameters associated with covariates were zero (under the null hypothesis, the −2 log likelihood statistic has a chi-square distribution). The AIC (Akaike information criterion) and SC (Schwarz criterion) statistics adjusted the −2 log likelihood statistic for the number of terms in the model and the number of observations used. These statistics are used when comparing different models for the same data and lower values of these statistics indicate a model with better goodness-of-fit (SAS, 2003):

$$\text{comprehensive DI} - \text{CSI model:DI} - \text{CSI} = \frac{\exp[g_1(x)]}{1 + \exp[g_1(x)]}$$

(1)

where

$$g_1(\boldsymbol{x}) = 7.62 - 0.11\text{CT} + 0.55\text{LC} - 0.91\text{VT} - 0.67\text{RC} + 0.13\text{RCH} - 0.86\text{LN} - 0.74\text{SL} + 0.29\text{SUR} - 0.59\text{SF} - 1.74\text{AI} - 2.69\text{NTC} - 0.48\text{FL} + 1.51\text{ST}$$

and the descriptions of the variables can be found in Table 1.

Table 4. Variables and coefficients for the comprehensive DI-CSI model

Variable	**Coeff.**	**Standard error**	**Wald chi-square**	***p*-Value**
Constant	7.62	2.20	12.00	0.001
Crash time (CT)	−0.11	0.22	0.26	0.613
Light condition (LC)	0.55	0.29	3.46	0.063
Vehicle type (VT)	−0.91	0.36	6.19	0.013
Road class (RC)	−0.67	0.53	1.57	0.210
Road character (RCH)	0.13	0.15	0.74	0.389
No. of lanes (LN)	−0.86	0.23	13.61	<0.001
Speed limit (SL)	−0.74	0.23	10.36	0.001
Surface type (SUR)	0.29	0.41	0.48	0.490
Special feature (SF)	−0.59	0.48	1.52	0.218
Area information (AI)	−1.74	0.61	8.05	0.005
None/inoperative traffic control (NTC)	−2.69	1.09	6.04	0.014
Flagger (FL)	−0.48	0.60	0.63	0.427
Stop sign/signal (ST)	1.51	0.66	5.31	0.021

AIC = 258.8; SC = 312.1; −2 log likelihood = 230.8. Testing global null hypothesis: $\beta = 0$: likelihood ratio chi-square (chi-square value, *p*-value): 104.1, <0.001; score chi-square (chi-square value, *p*-value): 89.6, <0.001; Wald chi-square (chi-square value, *p*-value): 58.3, <0.001.

In Table 4, the *p*-values of some variables, such as crash time, road character, surface type, and flagger/officer, are large (i.e., larger than the pre-set criterion of 0.3). From the statistical viewpoint, dropping these variables from the regression model does not lose much data information. Thus, a simplified DI-CSI model (Eq. (2)) was developed by including only the statistically significant variables that had relatively small *p*-values. The variables coefficients of the second DI-CSI model are presented in Table 5:

$$\text{Simplified DI} - \text{CSI model:DI} - \text{CSI} = \frac{\exp[g_2(x)]}{1 + \exp[g_2(x)]}$$

(2)

where

$$g_2(\boldsymbol{x}) = 7.64 + 0.54\text{LC} - 0.93\text{VT} - 0.59\text{RC} - 0.54\text{SF} - 0.86\text{LN} - 0.70\text{SL} - 1.62\text{AI} - 2.71\text{NTC} + 1.40\text{ST}.$$

Table 5. Variables and coefficients for the simplified DI-CSI model

Variable	**Coeff.**	**Standard error**	**Wald chi-square**	***p*-Value**
Constant	7.64	2.06	13.79	<0.001
Light condition (LC)	0.54	0.20	7.40	0.007
Vehicle type (VT)	−0.93	0.36	6.67	0.010
Road class (RC)	−0.59	0.52	1.27	0.260
Special feature (SF)	−0.54	0.45	1.43	0.232
No. of lanes (LN)	−0.86	0.23	14.16	<0.001
Speed limit (SL)	−0.70	0.22	9.79	0.002
Area information (AI)	−1.62	0.60	7.25	0.007
Non/inoperative traffic control (NTC)	−2.71	1.09	6.21	0.013
Stop sign/signal (ST)	1.40	0.64	4.78	0.029

AIC = 252.9; SC = 291.0; −2 log likelihood = 232.9. Testing global null hypothesis: β = 0 likelihood ratio chi-square (chi-square value, *p*-value): 101.9, <0.001; score chi-square (chi-square value, *p*-value): 88.4, <0.001; Wald chi-square (chi-square value, *p*-value): 57.8, <0.001.

4.2.2. Developed DD-CSI models

A pair of DD-CSI models was also developed by considering both work zone variables and driver characteristics. The comprehensive DD-DSI model generated by SAS was presented in Eq. (3). This model included all risk factors that were selected from the candidate crash variables. Table 6 lists the estimated variable coefficients for the model:

$$\text{comprehensive DD} - \text{CSI model:DD} - \text{CSI} = \frac{\exp[g_3(x)]}{1 + \exp[g_3(x)]}$$

(3)

where

$$g_3(\boldsymbol{x}) = 5.25 + 0.03\text{CT} + 0.51\text{LC} - 0.80\text{VT} - 0.59\text{RC} + 0.16\text{RCH} - 0.70\text{LN} - 0.84\text{SL} + 0.40\text{SUR} - 0.37\text{SF} - 1.69\text{AI} - 2.52\text{NTC} - 0.82\text{FL} + 0.78\text{ST} + 0.32\text{AG} - 0.81\text{AL} + 1.18\text{DTC} - 0.61\text{SP} - 1.98\text{FC}.$$

Table 6. Variables and coefficients for the comprehensive DD-CSI model

Variable	**Coeff.**	**Standard error**	**Wald chi-square**	***p*-Value**
Constant	5.25	2.33	5.07	0.024
Crash time (CT)	0.03	0.24	0.01	0.917
Light condition (LC)	0.51	0.32	2.48	0.116
Vehicle type (VT)	−0.80	0.39	4.13	0.042

Road class (RC)	−0.59	0.57	1.07	0.301
Road character (RCH)	0.16	0.17	0.84	0.359
No. of lanes (LN)	−0.70	0.25	8.02	0.005
Speed limit (SL)	−0.84	0.26	10.65	0.001
Surface type (SUR)	0.40	0.45	0.79	0.375
Special feature (SF)	−0.37	0.51	0.53	0.465
Area information (AI)	−1.69	0.67	6.36	0.012
None/inoperative traffic control (NTC)	−2.52	1.13	4.94	0.026
Flagger (FL)	−0.82	0.72	1.31	0.252
Stop sign/signal (ST)	0.78	0.73	1.15	0.284
Age (AG)	0.32	0.10	10.24	0.001
Alcohol/drug impairment (AL)	−0.81	0.67	1.45	0.228
Disregarded traffic control (DTC)	1.18	0.57	4.30	0.038
Speeding/too fast for condition (SP)	−0.61	0.52	1.35	0.244
Following too close (FC)	−1.98	1.07	3.39	0.066

AIC = 244.0; SC = 316.4; −2 log likelihood = 206.0. Testing global null hypothesis: β = 0 likelihood ratio chi-square (chi-square value, *p*-value): 128.9, <0.001; score chi-square (chi-square value, *p*-value): 105.8, <0.001; Wald chi-square (chi-square value, *p*-value): 58.9, <0.001.

A simplified DD-CSI model was developed as well by eliminating the variables with large *p*-values including crash time, road class, road character, road surface type, and road spatial feature. The following is the simplified DD-CSI model (Eq. (4)) and the variable coefficients are listed in Table 7:

$$\text{simplified DD} - \text{CSI model:DD} - \text{CSI} = \frac{\exp[g_4(x)]}{1 + \exp[g_4(x)]}$$

(4)

where

$$g_4(\boldsymbol{x}) = 4.88 + 0.63\text{LC} - 0.81\text{VT} - 0.58\text{LN} - 0.87\text{SL} - 1.77\text{AI} - 2.63\text{NTC} - 0.70\text{FL} + 0.73\text{ST} + 0.33\text{AG} - 0.85\text{AL} + 1.08\text{DTC} - 0.52\text{SP} - 2.01\text{FC}.$$

Table 7. Variables and coefficients for the simplified DD-CSI model

Variable	**Coeff.**	**Standard error**	**Wald chi-square**	***p*-Value**
Constant	4.88	1.80	7.32	0.007
Light condition (LC)	0.63	0.22	7.93	0.005
Vehicle type (VT)	−0.81	0.39	4.22	0.040
No. of lanes (LN)	−0.58	0.16	13.44	<0.001
Speed limit (SL)	−0.87	0.25	12.46	<0.001
Area information (AI)	−1.77	0.65	7.33	0.007
None/inoperative traffic control (NTC)	−2.63	1.13	5.47	0.019
Flagger (FL)	−0.70	0.70	1.02	0.313
Stop sign/signal (ST)	0.73	0.69	1.12	0.291
Age (AG)	0.33	0.10	11.12	0.001
Alcohol/drug impairment (AL)	−0.85	0.67	1.65	0.199

Disregarded traffic control (DTC)	1.08	0.55	3.88	0.049
Speeding/too fast for condition (SP)	−0.52	0.49	1.12	0.289
Following too close (FC)	−2.01	1.06	3.57	0.059

AIC = 236.9; SC = 290.2; −2 log likelihood = 208.9. Testing global null hypothesis: β = 0 likelihood ratio chi-square (chi-square value, *p*-value): 120.9, <0.001; score chi-square (chi-square value, *p*-value): 103.6, <0.001; Wald chi-square (chi-square value, *p*-value): 60.5, <0.001.

4.3. Model validation

The developed models were validated using 355 severe crash cases including 18 fatal crashes and 337 injury crashes in Kansas highway work zones in 2004. During the validation, researchers specified a CSI of one as a fatal crash and a CSI of zero as an injury crash. A CSI was calculated for each crash case based on the given crash variables. An estimated CSI number that is close to one indicated a very high risk level or a great likelihood of having a fatal crash for the given work zone travel conditions, while a CSI that is close to zero indicated a relative moderate risk level or a great likelihood of having a less severe crash such as an injury crash. The predicted CSI values were compared with the actual crash outcomes to illustrate the prediction accuracies. In addition, the four developed models were compared with each other.

Table 8 presents the comparison results between the estimated CSI numbers and the real crash severities. It shows minor differenced between the two CSI models in each category in terms of accuracy. Fig. 2, Fig. 3 graphically illustrate the estimated indices of the crashes using the two simplified models, respectively. When setting 0.5 as the criterion for the CSI (i.e., CSI ≥ 0.5 for likelihood of having a fatal crash and CSI < 0.5 for likelihood of having an injury crash), on average, the models predicted about five fatal crash cases (with CSI values greater than or equal to 0.5) out of the 18 fatal cases. On the other hand, all four models predicted about 95% of the injury cases (CSI < 0.5). Based on the 2004 injury and fatal crash data, the simplified DI- and DD-CSI models were slightly better than the comprehensive models for both accuracies in percentage and average estimated CSI values.

Table 8. Prediction accuracies of the CSI models

Model	**Accuracy**			$\sum(\mathrm{ACS}-\mathrm{CSI})^{2\mathrm{a}}$
	Fatal (%)	**Injury (%)**	**Total (%)**	
Comprehensive DI-CSI	28	95	92	22.3
Simplified DI-CSI	33	95	92	21.8
Comprehensive DD-CSI	22	95	91	22.9
Simplified DD-CSI	28	95	92	21.6

[a]Sum of squared errors, where ACS = actual crash severity (1 for fatal and 0 for injury), and CSI = estimated crash severity index.

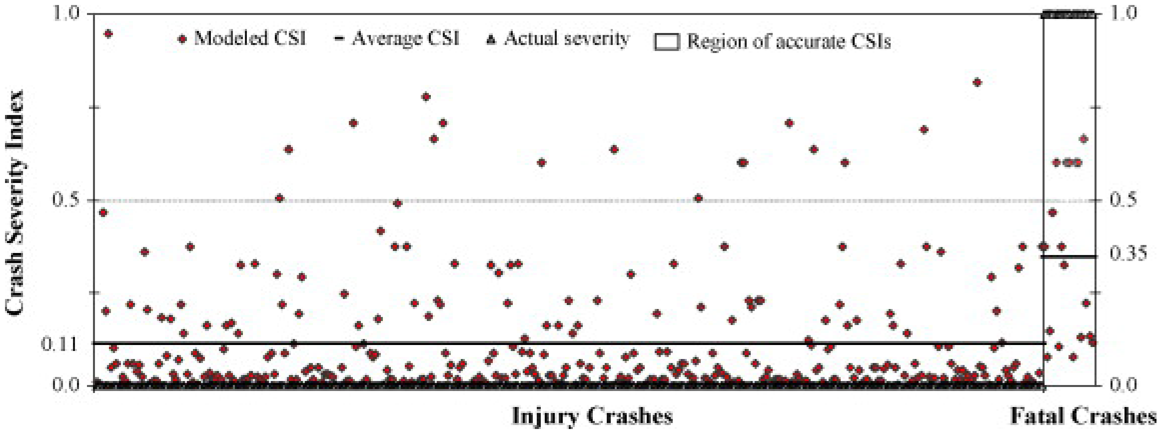

Fig. 2. CSIs estimated by the simplified DI-CSI model.

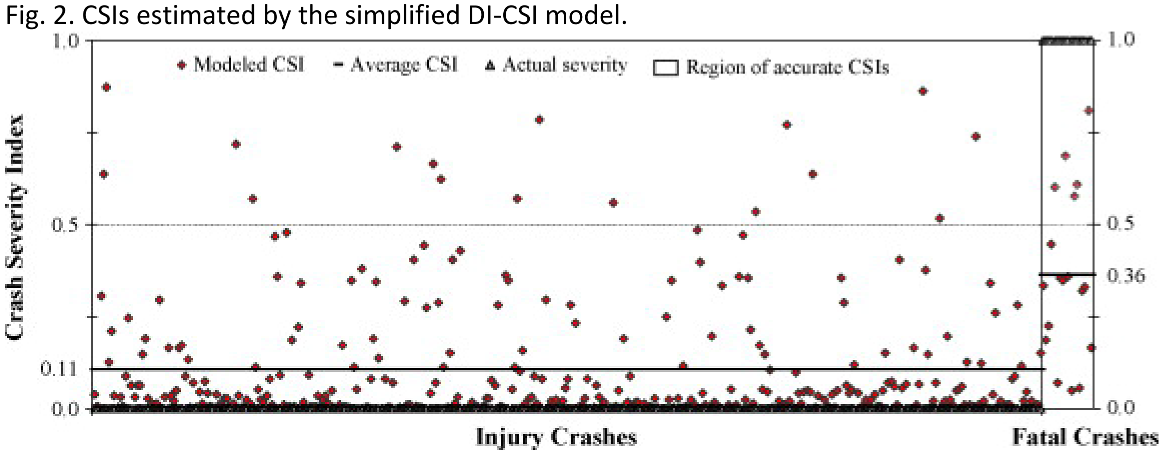

Fig. 3. CSIs estimated by the simplified DD-CSI model.

According to these four models, the average CSI for the travel conditions of injury crashes were around 0.11, while the average CSI for fatal cases fell between 0.3 and 0.36 (comprehensive DI-CSI model: 0.32; simplified DI-CSI model: 0.35; comprehensive DD-CSI model: 0.30; simplified DD-CSI model: 0.36). Generally, the models captured the differences of the input work zone travel conditions and successfully separated different traffic conditions by assigning them with different CSI values (i.e., not dramatically clustered in a certain small range). However, the accuracy of using CSI to predict the fatal crashes may be further improved through future research. For example, a larger dataset including sufficient fatal crash information may be used when available in future development.

Table 9 present some examples of work zone travel conditions with very high CSI values estimated by the comprehensive DD-CSI model. Typically, risk factors such as poor light condition, truck involvement, having only two travel lanes, and high speed limit may lead to high CSI values and equivalently, high risk levels. Note that, in the table, the travel conditions with very-high CSI values included an injury case. This indicated that a high CSI may not necessarily coincide with a fatal crash; a CSI with a high value implies that the condition is risky and it has a high likelihood of causing high-severity crashes such as fatal crashes.

Table 9. Example conditions with high CSIs

Crash variable	**High-CSI conditions**		

Empty Cell	**No. 1**	**No. 2**	**No. 3**
CSI	0.62	0.75	0.88
Actual crash severity	Fatal	Fatal	Injury
Age	65 or older	35–44	35–44
Crash time	10:00 a.m. to 4:00 p.m.	10:00 a.m. to 4:00 p.m.	8:00 p.m. to 6:00 a.m.
Light condition	Good condition	Good condition	Poor condition
Vehicle type	Non-truck involved	Truck involved	Truck involved
Road class	Other principal arterials and minor arterials	Other principal arterials and minor arterials	Other principal arterials and minor arterials
Road character	Straight and level	Other alignments	Curved and level
No. of lanes	4	2	2
Speed limit	51–60 mph	≥61 mph	≥61 mph
Surface type	Concrete	Blacktop	Blacktop
Special feature[a]	Impacted	Not present	Not present
Area information	Urban area	Rural area	Rural area
None/inoperative TC[b]	Not present	Not present	Not present
Flagger/officer	Not present	Not present	Not present
Stop sign/signal	Not present	Not present	Present
Alcohol/drug impairment	Not present	Not present	Not present
Disregarded TC	Present	Not present	Not present
Speeding/too fast for condition	Not present	Not present	Present
Following too closely	Not present	Not present	Not present

[a]Special features may include bridge, overhead bridge, railroad bridge, railroad crossing, interchange, ramp, and other.
[b]Traffic control.

5. Conclusion and recommendation

In this study, four CSI models were developed for risk level assessment in work zones based on crash severity modeling. The models incorporated the risk factors that were determined using chi-square tests, CMH statistics, and results of the previous crash characteristic study. The CSI models were designed to quantify the risk level of a work zone with a numerical value between zero and one. A CSI of one indicates a very high risk level in a given work zone, which infers that a fatal crash might take place if a crash occurs.

Two groups of models were developed, including two driver-independent CSI or DI-CSI models and two driver-dependent CSI or DD-CSI models. The DI-CSI models were developed for the work zone travel risk assessment without considering human factors or specific driving groups; the DD-CSI models, on the other hand, addressed the risks associated with travel conditions along with human errors and the characteristics of specific driving group. Thus, DD-CSI models are suitable for the driving risk assessment for given driving groups in given highway work zones.

Generally, the CSI models captured the differences between the work zone conditions with fatal and injury crashes. Model validation showed that the CSIs for most work zones with severe crashes were consistent with the actual crash severity outcomes. The researchers recommend that the CSI models should be used in work zone planning or work zone safety inspection so that work zone risk factors could be identified and safety countermeasures could be developed accordingly to mitigate risk. Utilization of CSI models will help engineers to reveal work zone risks that are created by subtle combinations of a wide range of variables which otherwise may be not detectable solely based on engineering experience. Model validation showed minor accuracy differences between the comprehensive models and the simplified models. Therefore, the researcher could not reach the conclusion on which models were credibly superior. Additional validations with large datasets are needed. When there is sufficient information, it is recommended that the comprehensive models be used since they include all risk factors identified based on both statistical tests and crash characteristic studies.

While the predicted CSI values for most of the travel conditions for injury crashes were consistent with the actual crash severity observed, the predicted CSI values for some of the fatal crash cases were not consistent with the actual severity outcomes. Reasons for these inconsistencies may include:

The covariate pattern examination showed that both fatal and injury crashes were observed for some work zone conditions. A covariate pattern is a certain combination of crash variables with certain values. This suggests that a minor fraction of fatal and injury crashes could not be separated by travel conditions shown in the KDOT crash reports. The CSI numbers for these risk conditions would be either biased to a low value (if the conditions were dominated by injury crashes) or to a high value (if the conditions were dominated by fatal crashes).

In both model development and model validation datasets, the existence of very severe injury crashes (e.g., near-fatal injury crashes) and some fatal crashes, whose fatalities were due to reasons other than work zone risk factors such as physical vulnerability or not wearing a seat belt, would reduce the accuracy of the models. Using more detailed crash severity classification may eliminate or mitigate this type of error.

The crash data used for model validation had only 18 fatal crash cases. The size of the fatal crash sample might not be large enough to validate the developed models under typical fatal conditions.

Future research is recommended for the improvement of the CSI models. When available, a larger dataset should be used for the future development and validation of the CSI models. The CSI models can also be improved by taking into consideration the crashes of other severities such as property-damage-only crashes. In addition, more detailed classification of crash severities should be used during future development of CSI models so that the CSIs with intermediate values can be interpreted with corresponding severities. Information on work zone configurations, if available, should also be included in the CSI models to improve their accuracies.

Acknowledgements

The authors would like to thank Mr. Anthony Alrobaire and Mr. Rex McCommon from KDOT for their valuable help and advice during the course of this study. Heartfelt thanks also go to Dr. Yaozhong Hu in the Department of Mathematics at the University of Kansas for his valuable advice during the course of this study. The cooperation and assistance from other KDOT staff during the process of crash data collection are also greatly appreciated.

References

H. Chang, T. Yeh. Risk factors to driver fatalities in single-vehicle crashes: comparisons between non-motorcycle drivers and motorcyclists. Journal of Transportation Engineering, 132 (3) (2006), pp. 227-236

W. Chen, P.P. Jovanis. Method for identifying factors contributing to driver-injury severity in traffic crashes. Journal of the Transportation Research Board, 1707 (2000), pp. 1-9

T. Chirsa-Chavala, K.K. Mak. Identification of accident factors on highway segments: a method and applications. Journal of the Transportation Research Board, 1068 (1986), pp. 52-58

S. Dissanayake, J. Lu. Analysis of severity of young driver crashes, sequential binary logistic regression modeling. Journal of the Transportation Research Board, 1784 (2002), pp. 108-114

F.E. Harrell Jr.. Regression Modeling Strategies-with Application to Linear Models, Logistic Regression, and Survival Analysis. Springer-Verlag, New York Inc. (2001), pp. 215–221

R.W. Hill. Master Thesis: Statistical Analysis of Fatal Traffic Accident Data. Texas Tech University, Lubbock, Texas (2003)

K. Kim, S. Kim, E. Yamashita. Alcohol-impaired motorcycle crashes in Hawaii, 1986–1995. Journal of the Transportation Research Board, 1704 (2000), pp. 77-85

Li, Y., 2007. Analyzing highway work zone crashes and traffic control effectiveness. Ph.D. Dissertation. The University of Kansas, Lawrence, Kansas.

Y. Li, Y. Bai. Investigating the characteristics of fatal crashes in the highway construction zones. CIB W99 International Conference on Global Unity for Safety & Health in Construction, Beijing, China, June 28–30 (2006), pp. 301-309

G. Lu, D.A. Noyce, R.J. McKendry. Analysis of the magnitude and predictability of median crossover crashes utilizing logistic regression. Proceedings of the TRB 85th Annual Meeting, CD-ROM, January 22–26 (2006)

Y. Ouyang, V. Shankar, T. Yamamoto. Modeling the simultaneity in injury causation in multi-vehicle collisions. Journal of the Transportation Research Board, 1704 (2002), pp. 143-152

SAS Help and Documentation (SAS), 2003. The Logistic Procedure—Model Fitting Information. SAS Institute Inc., Cary, NC.