

Validation of FHWA Crash Models for Rural Intersections

Lessons Learned

Jutaek Oh, Craig Lyon, Simon Washington, Bhagwant Persaud, and Joe Bared

A national-level safety analysis tool is needed to complement existing analytical tools for assessment of the safety impacts of roadway design alternatives. FHWA has sponsored the development of the Interactive Highway Safety Design Model (IHSDM), which is roadway design and redesign software that estimates the safety effects of alternative designs. Considering the importance of IHSDM in shaping the future of safety-related transportation investment decisions, FHWA justifiably sponsored research with the sole intent of independently validating some of the statistical models and algorithms in IHSDM. Statistical model validation aims to accomplish many important tasks, including (a) assessment of the logical defensibility of proposed models, (b) assessment of the transferability of models over future time periods and across different geographic locations, and (c) identification of areas in which future model improvements should be made. These three activities are reported for five proposed types of rural intersection crash prediction models. The internal validation of the model revealed that the crash models potentially suffer from omitted variables that affect safety, site selection and countermeasure selection bias, poorly measured and surrogate variables, and misspecification of model functional forms. The external validation indicated the inability of models to perform on par with model estimation performance. Recommendations for improving the state of the practice from this research include the systematic conduct of carefully designed before-and-after studies, improvements in data standardization and collection practices, and the development of analytical methods to combine the results of before-and-after studies with cross-sectional studies in a meaningful and useful way.

Effective safety management requires that engineers know the present safety performance of a roadway and how it will perform if contemplated actions are taken. In effect, a reliable method for estimation of the safety performance of roadways under a host of potential future scenarios is needed. To this end, FHWA has sponsored research and development for a new approach that combines historical accident data, regression analysis, before-and-after studies, and expert judgment to make safety performance predictions that are expected to be better than those obtained by any of the individual approaches. A recent report documents an accident prediction algorithm for implementation of the new approach for two-lane rural highway sections that includes road segments and five types of intersections (1). Ongoing efforts aim to produce similar documents for other types

of facilities. A companion paper by Lyon et al. focuses on detailed aspects of this proposed algorithm (Lyon et al., pp. 78–86, this volume).

This paper, in contrast, reports on the validation of individual crash models intended for use in the Interactive Highway Safety Design Model (IHSDM). Many articles in the literature report on the value and virtues of model validation (2–4). Widespread agreement on a precise definition of model validation is lacking in the research community, although quite a bit has been written on the subject (2, 5). Validation can also be thought of as a requirement to demonstrate that a model is appropriate, meaningful, and useful for the purpose for which it is intended. Validation exercises are often associated with assessment of the prediction ability of a statistical model (2). However, it is possible for a model to predict an underlying data-generating process adequately but fail to illuminate and explain the nature of the underlying process. For this reason, model validation is broken down into two distinct aspects: internal validity and external validity. Internal model validity, as applied in this research, is concerned with the ability of the intersection crash models to explain the underlying phenomenon, whereas external model validity is concerned with the ability of the models to predict crashes over time and space. The internal and external validation tools applied are described here.

BACKGROUND AND METHODS

Validation of Accident Prediction Models in IHSDM

A statistical model, if internally valid, should agree with theoretical expectations in a number of respects. First, it should be consistent with established knowledge on the subject. Disagreement with past research should raise concern as to the plausibility of the results and require a greater burden of proof and explanation. Second, a good model should possess the salient and important features of the underlying system or phenomenon. Finally, the model or important parts of the model should agree with fundamental information and knowledge, such as physical mechanics, properties of materials, and dynamics involved with crashes. It is the rigorous application of critical thinking about a phenomenon that can often raise doubt as to the validity of hypothesized relationships between variables and that serves to provide internal validation of statistical models.

External validation is focused on the goodness of fit (GOF) of statistical models to independent data. Several GOF measures were used to assess model performance in this research. It is important to note that an objective assessment of the predictive

J. Oh and S. Washington, Department of Civil Engineering, University of Arizona, Tucson, AZ 85721-0072. C. Lyon and B. Persaud, Department of Civil Engineering, Ryerson University, Toronto, Ontario M5B 2K3, Canada. J. Bared, Research and Development, FHWA, 6300 Georgetown Pike, McLean, VA 22101.

performance of a particular model can be made only through the evaluation of several GOF criteria. In addition, a model should be assessed on the basis of both its internal and external validities. The GOF measures used to conduct external model validation include the following:

1. Pearson product moment correlation coefficients between observed and predicted crash frequencies. Pearson's product moment correlation coefficient, usually denoted by r , is one example of a correlation coefficient. It is a measure of the linear association between two variables, Y_1 and Y_2 , that have been measured on interval or ratio scales and is given by

$$r_{12} = \frac{\sum (Y_{i1} - \bar{Y}_1)(Y_{i2} - \bar{Y}_2)}{[\sum (Y_{i1} - \bar{Y}_1)^2 \sum (Y_{i2} - \bar{Y}_2)^2]^{1/2}}$$

where \bar{Y} is the mean of Y_i observations. Theoretically, a model that predicts observed data perfectly will produce a straight-line plot between the observed and the predicted values, a correlation coefficient of exactly 1.

2. Mean prediction bias (MPB). MPB provides a measure of the magnitude and direction of the average model bias in comparison with validation data. The smaller the absolute value of average prediction bias is, the better the model does at predicting the observed data. A positive MPB indicates that a model overpredicts crashes, on average, while a negative MPB indicates systematic underprediction of crashes, on average. MPB is given by

$$\text{MPB} = \frac{\sum_{i=1}^n (\hat{Y}_i - Y_i)}{n}$$

where n is the validation data sample size, and \hat{Y} is the fitted value of Y .

3. Mean absolute deviation (MAD). MAD provides a measure of the average misprediction of the model. It differs from MPB in that positive and negative prediction errors do not cancel. A value close to 0 suggests that, on average, the model predicts the observed data well. MAD is given by

$$\text{MAD} = \frac{\sum_{i=1}^n |\hat{Y}_i - Y_i|}{n}$$

where n is the validation data sample size.

4. Mean squared prediction error (MSPE) and mean squared error (MSE). MSPE is the sum of the squared differences between observed and predicted crash frequencies divided by sample size. MSPE is typically used to assess the error associated with a validation or external data set and is given by

$$\text{MSPE} = \frac{\sum_{i=1}^n (\hat{Y}_i - Y_i)^2}{n_2}$$

where n_2 is the validation data sample size. MSE is the sum of the squared differences between observed and predicted crash frequencies divided by the sample size minus the number of model parameters:

$$\text{MSE} = \frac{\sum_{i=1}^n (\hat{Y}_i - Y_i)^2}{n_1 - p}$$

where n_1 is the estimation data sample size, and p is the number of degrees of freedom. A comparison of MSPE and MSE reveals potential over- or underfitting of the models to the estimation data. An MSPE that is higher than MSE may indicate that the models may have been overfit to the estimation data and that some of the observed relationships may have been spurious instead of real. This finding could also indicate that important variables were omitted from the model or the model was misspecified.

To normalize the GOF measures to compensate for the different numbers of years associated with different data sets, the research team computed GOF measures on a per-year basis. For MPB and MAD per year, MPB and MAD were divided by the number of years. However, because MSPE and MSE are the mean values of the squared errors, the research team divided MSPE and MSE by the square of the number of years to calculate MSPE and MSE per year, which resulted in a fair comparison of predictions based on different numbers of years. This correction is needed because the models developed predict crashes per unit of time. Because the variance in crashes is a function of the mean, a larger time period results in both a larger number of crashes and a larger variance in crashes, all else being equal. Computation of GOF statistics on a per year basis normalized comparisons across models with different numbers of years of observation.

Data Description

The data used in the external validation activities came from three sources: (a) the original data used for calibration of the main models for total accidents, (b) Highway Safety Information System crash data for additional years for the same intersections used in the original calibration, and (c) an independent validation data set assembled for this project for sites in Georgia. The basic statistics for these data sets are summarized in Table 1.

To validate Models I and II across time, additional years of accident data from Minnesota were used, while for Models III, IV, and V, additional years of accident data from California and Michigan were used. To validate the models across jurisdictions (space), data from Georgia were assembled and used.

Models I and II were originally developed to estimate police-reported intersection or intersection-related accidents that occurred within 250 ft of an intersection. Models III, IV, and V were each developed for two types of dependent variables. The first type was estimated by using accidents (TOTACC) occurring within 250 ft of the intersection. The second type was estimated by using only those accidents considered to be intersection related and occurring within 250 ft of the intersection (TOTACCI). For California and Georgia, special criteria were needed for the latter case because those states do not include a variable indicating whether an accident was intersection related. According to the criteria, accidents were considered intersection related if they were

1. A vehicle-pedestrian accident;
2. An accident in which one vehicle involved was making a left turn, a right turn, or a U-turn before the crash; and
3. A multivehicle accident in which the accident type was either a sideswipe, a rear end, or broadside or angled.

There was insufficient information to test whether alternative distances from the intersections would significantly influence the results.

TABLE 1 Basic Statistics for Data Sources

		Model I	Model II ^a	Model III ^b	Model IV ^b	Model V ^b
Sample Size	Original years	389	327	84	72	49
	Georgia data	121	114	52	52	51
States ^c	Original years	Mn	Mn	Ca & Mi	Ca & Mi	Ca & Mi
	Subsequent years	Mn & Wa	Mn & Wa	Ca & Mi	Ca & Mi	Ca & Mi
	Georgia data	Ga	Ga	Ga	Ga	Ga
Years Covered	Original years	5 (1985 to 1989)	5 (1985 to 1989)	3 (1993 to 1995)	3 (1993 to 1995)	3 (1993 to 1995)
	Subsequent years	9 (1990 to 1998)	9 (1990 to 1998)	2 (1996 to 1997)	2 (1996 to 1997)	2 (1996 to 1997)
	Georgia data	2 (1996 to 1997)	2 (1996 to 1997)	2 (1996 to 1997)	2 (1996 to 1997)	2 (1996 to 1997)
	Georgia data	2 (1996 to 1997)	2 (1996 to 1997)	2 (1996 to 1997)	2 (1996 to 1997)	2 (1996 to 1997)

NOTE: Models I and II are from Vogt and Bared (6), and Models III to V are from Vogt (7).
Abbreviations: Mn = Minnesota; Wa = Washington, Ca = California; Mi = Michigan; Ga = Georgia.

The definition of 250 ft is thought to capture a proximity to the intersection reasonable enough that all accidents within this distance are related, in some way, to the intersection. Of course, there are exceptions to this generalization, and this motivated the desire to select a subset of intersection-related criteria.

For consistency with the validation for Models I and II, only the TOTACCI results for intersection-related accidents are discussed. The full report contains the results for TOTACC for Models III, IV, and V, as well as the corresponding validation results for the injury accident models developed in the original research (8).

Intersection Crash Models in IHSDM

The report by Harwood et al. presented base models and accident modification factors (AMFs) for three- and four-legged intersections of two-lane rural roads with “stop”-sign control and four-legged signalized intersections of two-lane roads (1). These base models represented the best of the available accident prediction models developed in earlier FHWA projects and included only variables that were statistically significant at the 15% level. In addition to the base models, the researchers developed full models, which included additional variables, with the intention of using the variable coefficients to estimate AMFs for use in IHSDM. The full models, along with several variants, are presented in two FHWA reports. Vogt and Bared (6) present models for three- and four-legged intersections of two-lane roads, while Vogt (7) documents models for three other types of rural intersections: three- and four-legged stop-sign-controlled intersections with four lanes on the major road and two lanes on the minor road and signalized intersections of two-lane roads. In summary, five types of intersection accident prediction models pertain to the research efforts:

- Type I, three-legged stop-controlled intersections of two-lane roads;
- Type II, four-legged stop-controlled intersections of two-lane roads;
- Type III, three-legged stop-controlled intersections with two lanes on the minor road and four lanes on the major road;
- Type IV, four-legged stop-controlled intersections with two lanes on the minor road and four lanes on the major road; and
- Type V, signalized intersections of two-lane roads.

The models were developed by using data that were limited in terms of geographical diversity (nonrandom samples) and, for Type

III, IV, and V intersections, small sample sizes. Thus, comprehensive validation of these statistical models, including predictive ability across both space and time, logical defensibility, and identification of the limitations arising from small samples and nonrandom sampling, is of paramount importance.

RESULTS

Preliminary Validation

The research team undertook two separate external validation exercises, which are described next and presented in detail in the following sections. The detailed results of these two separate external validation exercises are given for each of the five intersection crash models. It should be noted that model validation, the subject of this paper, is concerned with assessing the existing proposed models. A separate activity not presented here—model recalibration—is concerned with exploring the addition of new variables and different model functional forms. To examine those results, see the work by Washington et al. (8) and the companion paper by Lyon et al. in this volume on pp. 78–86. The two separate external validation exercises undertaken here were as follows:

1. Validation of the models against additional years of accident data for the same intersections. Because the crash models were developed as direct inputs into the Crash Prediction Module of IHSDM, the models should be able to forecast crashes during different time periods. This validation activity was used to assess the models’ ability to forecast crashes across time. The temporal stability of a model suggests that it will predict crash frequencies well across time, meaning that the effect of time or covariates that are influenced by time either is not important or, if it is important, is included in the model in some relevant variable expression.

2. Validation of the models against data for Georgia. This validation exercise was used to assess the models’ ability to forecast crashes over a jurisdiction whose data were not used in model estimation. Because the primary application of the models will be to forecast the impact of design considerations and countermeasures in regions and jurisdictions whose data were not represented in the data used in model estimation, this activity represents an important test of the practicality of the model. As a result of the predictive performance during this activity, conclusions can be drawn about the collection of variables used to forecast crashes.

Descriptions and definitions of all variables used in this paper are provided in Table 2. Comparisons of the coefficients for the published models and additional years of data for the same sites and for sites in Georgia are provided in Tables 3 (Models I and II) and 4 (Models III, IV, and V). External validation statistics are presented in Table 5. The data in all tables are based on intersection-related accidents for all crash severities combined. Throughout the discussion a value of α (the probability of a Type I error) equal to 0.10 is used. GOF statistics provide an ability to objectively assess the fit of a model to the data. Comparisons between models, however, are generally subjective. In the following analyses the terms “serious,” “moderate,” and “marginal” are applied to denote a subjective evaluation of GOF comparisons between models. Serious differences in GOF are suggestive of significant model deficiencies. Moderate differences in GOF suggest cases in which models could be improved but improvements might be difficult to obtain. Marginal differences in GOF are thought to be negligible and are potentially explained by random fluctuations in the observed data.

Internal Validity

As discussed previously, this task focused on the logical defensibility of the crash models in the originally published documents. In all fair-

ness to the original modelers, some serious constraints were imposed on statistical modeling, most of which reflect the state of affairs in safety modeling rather than poor decisions, poor research designs, or poor analysis methods. The constraints included the following:

1. Limited quality and availability of data, in particular, estimates of annual average daily traffic (AADT), which in many cases are average daily traffic (ADT) estimates adjusted by using regionwide seasonal or daily correction factors instead of facility-specific correction factors to obtain AADT estimates; the quality also suffers from the frequency and the placement of traffic counters in proximity to the intersections included in the study; in addition, crash reporting practices and thresholds are beyond the control of the researchers and contribute to poorly measured response variables (crashes), especially as they are compared across jurisdictions [for a discussion on the seriousness of these issues see the work by Hauer (9)].
2. Omitted variables, including weather information, environmental conditions, and driver population characteristics.
3. Site-selection bias, resulting in biases in statistical model parameters and arising, for example, from countermeasures applied as a result of meeting accident or volume warrants and from intersections being signalized (e.g., from four-way-stop-controlled intersections) because they meet accident or volume warrants.

TABLE 2 Variable Definitions

VARIABLE	DEFINITION
ABSGRD1	Average absolute grade on major road
ADT1	Average daily traffic on major road
ADT2	Average daily traffic on minor road
HAU	Intersection angle
HI	Sum of degree of curve in degrees per hundred feet of each horizontal curve on major road any portion of which is within 250 feet of the intersection center divided by the number of such curves
LTLNS	0 if the major road has no left-turn lane, 1 if the major road has at least one left-turn lane
NODRWY1	The number of residential and commercial driveways on the major road within 250 feet of the intersection center
MEDWDTH1	The major road median width in feet
PK%LEFT1	The percentage of all incoming major road traffic during peak hours that turns left
PK%LEFT2	The percentage of all incoming minor road traffic during peak hours that turns left
PK%TRUCK	The percentage of trucks passing through the intersection
PROT_LT	1 if protected left turn lane on major road, 0 otherwise
RT	1 if right turn lane on major road, 0 otherwise
RHR	Average roadside hazard rating within 250 feet of intersection center along major road
SPD	The average posted speed on major road in vicinity of the intersection
TOTACC	Total number of crashes occurring within 250 feet of intersection center
TOTACCI	Total number of intersection-type crashes occurring within 250 feet of intersection center
VCI	Sum of absolute change of grade in percent per hundred feet for each crest curve on major road any portion of which is within 250 feet of the intersection center, divided by the number of such curves
VEI1	Average curve grade rate along major road within 800 feet of the intersection center
VEI2	Average curve grade rate along minor road within 800 feet of the intersection center
VEICOM	Average curve grade rate along the major and minor road within 800 feet of the intersection center

TABLE 3 Parameter Estimates for Models I and II Total Intersection-Related Accidents

Variable	Model I			Model II		
	A	B	C	A	B	C
Constant	-12.99 (1.15, <0.001)	-12.92 (1.03, <0.001)	-6.84 (1.14, <0.001)	-10.43 (1.32, <0.001)	-10.79 (1.05, <0.001)	-7.21 (1.32, <0.001)
Log of ADT1	0.8052 (0.06, <0.001)	0.8204 (0.07, <0.001)	0.497 (0.11, <0.001)	0.6026 (0.08, <0.001)	0.6742 (0.08, <0.001)	0.627 (0.13, <0.001)
Log of ADT2	0.5037 (0.07, <0.001)	0.4844 (0.06, <0.001)	0.239 (0.127, 0.060)	0.6091 (0.07, <0.001)	0.6651 (0.06, <0.001)	0.493 (0.154, 0.001)
HI	0.0339 (0.033, 0.300)	0.0140 (0.020, 0.494)	-0.0209 (0.027, 0.443)	0.0449 (0.047, 0.343)	0.0630 (0.045, 0.159)	-0.0165 (0.040, 0.681)
VCI	0.2901 (0.294, 0.323)	-0.265 (0.272, 0.328)	0.0294 (0.048, 0.537)	0.2885 (0.258, 0.263)	0.097 (0.195, 0.617)	-0.0438 (0.078, 0.575)
SPD	0.0285 (0.018, 0.107)	0.0374 (0.013, 0.004)	0.00686 (0.009, 0.443)	0.0187 (0.018, 0.288)	0.0117 (0.016, 0.452)	-0.0177 (0.010, 0.087)
RHR	0.1726 (0.068, 0.011)	0.0761 (0.065, 0.242)	0.2048 (0.089, 0.021)			
RT	0.2671 (0.140, 0.056)	-0.078 (0.125, 0.534)	-0.158 (0.490, 0.748)			
HAU	0.0045 (0.003, 0.158)	0.00146 (0.002, 0.507)	-0.00546 (0.003, 0.088)	-0.0049 (0.003, 0.134)	-0.00011 (0.001, 0.935)	0.00000 (0.004, 1.000)
NODRWY1				0.1235 (0.052, 0.017)	0.0874 (0.041, 0.032)	0.0940 (0.071, 0.184)
Overdispersion	0.481	0.485	0.185	0.205	0.350	0.455

NOTE: A = published estimates (standard error, *p*-value); B = calibrated for additional years (standard error, *p*-value); C = calibrated for Georgia data (standard error, *p*-value).

TABLE 4 Parameter Estimates for Models III, IV, and V Total Intersection-Related Accidents

Variable	Model III			Model IV			Model V		
	A	B	C	A	B	C	A	B	C
Constant	-15.466 (3.469, <0.001)	-16.774 (3.850, <0.001)	-8.163 (4.097, 0.068)	-11.110 (3.335, 0.001)	-11.880 (3.687, 0.002)	-4.603 (5.498, 0.477)	-6.084 (3.387, 0.072)	-7.383 (4.264, 0.117)	-6.061 (3.486, 0.124)
Log of ADT1	1.433 (0.361, <0.001)	1.350 (0.410, 0.001)	0.301 (0.339, 0.375)	0.930 (0.343, 0.007)	0.798 (0.376, 0.034)	0.563 (0.497, 0.257)	0.595 (0.285, 0.037)	0.725 (0.433, 0.094)	0.644 (0.327, 0.049)
Log of ADT2	0.269 (0.099, 0.007)	0.544 (0.143, <0.001)	0.740 (0.229, 0.001)	0.354 (0.116, 0.002)	0.662 (0.167, <0.001)	-0.043 (0.326, 0.896)	0.294 (0.197, 0.137)	0.311 (0.189, 0.100)	0.204 (0.158, 0.198)
MEDWIDTH1	-0.061 (0.036, 0.089)	-0.031 (0.050, 0.507)	0.002 (0.012, 0.868)						
NODRWY1	0.056 (0.029, 0.053)	0.036 (0.049, 0.461)	0.095 (0.126, 0.450)						
PK%LEFT1 ^a				0.149 (0.059, 0.011)	0.110 (0.056, 0.051)	-			
PROT_LT							-0.471 (0.200, 0.019)	-0.738 (0.270, 0.006)	0.637 (0.441, 0.149)
PK%LEFT2 ^b							-0.017 (0.006, 0.004)	-0.012 (0.010, 0.225)	-
VEICOM							0.113 (0.037, 0.002)	-0.074 (0.069, 0.280)	0.021 (0.179, 0.908)
PK%TRUCK ^b							0.029 (0.013, 0.028)	0.023 (0.014, 0.094)	-
Overdispersion	0.512	0.722	0.272	0.710	0.626	0.857	0.131	0.243	0.708

NOTE: A = published estimates (standard error, *p*-value); B = calibrated for additional years (standard error, *p*-value); C = calibrated for Georgia data (standard error, *p*-value).

^aThe effect of PK%LEFT1 was not estimated because data for this variable were not obtainable in Georgia.

^bThe effects of PK%LEFT2 and PK%TRUCK were not estimated because data for these variables were not obtainable in Georgia.

TABLE 5 Validation Statistics for Models I to V Total Intersection-Related Accidents

Measure ^a	Model I			Model II			Model III			Model IV			Model V		
	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
No. of sites	389	367	121	389	367	121	84	84	52	72	72	51	49	49	51
Pearson Product-Moment Coefficient	0.66	0.62	0.31	0.77	0.72	0.39	0.67	0.52	0.10	0.47	0.53	0.17	0.62	0.37	0.23
MPB/Year	0.00	0.00	-0.28	0.00	-0.03	-0.36	0.00	-0.26	0.56	0.06	0.27	-0.88	0.09	1.54	1.58
MAD/Year	0.21	0.18	0.64	0.20	0.20	0.91	0.59	0.65	0.97	1.00	1.05	1.27	1.88	2.47	3.77
MSE/Year ²	0.19			0.10			0.72			2.77			6.47		
MSPE/Year ²		0.12	0.89		0.17	1.73		0.89	1.54		2.07	3.05		11.04	25.18

NOTE: A = original data (Models I and II, 1985 to 1989; Models III, IV, and V, 1993 to 1995); B = additional years of data (Models I and II, 1990 to 1998; Models III, IV, and V, 1996 and 1997); C = Georgia data (Models I to V, 1996 and 1997).

Given these constraints—which are typical of those encountered when safety assessments are conducted—it is expected, a priori, that statistical models estimated under these constraints will face internal validity concerns. By inspection of Tables 3 and 4, the following observations identify several shortcomings with respect to the logical defensibility of the crash models. Note that observations are framed with respect to the model context: aggregate statistical models that predict yearly crashes at intersections.

1. Omitted known variables. Potential omitted variables that are correlated with the variables included result in biased parameter estimates. Variables conspicuously missing from the models include weather-related variables, such as the numbers of wet pavement days (as a proportion of the number of days per year), snow and ice days, and fog or poor visibility days. These are likely to be interacted with some variables included, such as vertical and horizontal curvature variables. For instance, a vertical or horizontal curve on a wet or icy day is likely to have a different crash potential than the same road on a dry day. Other omitted variables include turning movements as percentages, as turning movements influence the potential for conflicts and crashes (the percentage of left turns during peak hours is included only in Model IV). The driving populations for sites situated near a cluster of bars, a high school, or a retirement community may be such that the crash frequencies at those sites are increased. Although they are difficult to capture, these factors are missing. Finally, there are no variables to indicate the risks associated with nighttime driving versus daytime driving or nighttime exposure versus daytime exposure. Many of the effects of these omitted variables will be picked up in regional or state effects; however, correlation of these variables with the variables included will cause bias or inefficiencies in parameter estimates.

2. Omitted unknown variables. Because these data are cross-sectional, the potential impact of unknown omitted variables is significant. That is, those factors whose influence on safety is not known or too difficult to measure are excluded from the models and may vary across sites and time. Because the set of unknown variables is likely to be different across regions or a state, their effects will enter models differentially.

3. Site-selection bias. It is difficult to determine the impact of site-selection bias on the results, but parameter estimates are often

biased negatively or positively, depending on the particular circumstances. Sites in the sample include intersections that might in the future be upgraded from a Type I or II (stop-controlled) intersection to a Type V (signalized) intersection because volume or crash warrants have been met, as well as those that were upgraded before the study. As a result, the dependent variable, crash frequency, is used to (partially) determine intersection type. If different warrants are used across jurisdictions or if jurisdictions differ in their time responses to the conversion of intersections, then comparisons across jurisdictions for the same type of intersections will be problematic.

4. Countermeasure-selection bias. Countermeasures applied when traffic volume or crash warrants are met also cause potential difficulties. For example, a left-turn lane installed because of a relatively large number of observed left-turn-related crashes, which may be elevated because of real safety deficiencies or a random upfluctuation, can result in the presence of a left-turn lane being statistically associated with a reduction or increase in the numbers of crashes. This potential bias can result in counterintuitive signs and magnitudes of effects. Another variable subject to this potential bias is the presence of a right-turn lane (Model I).

5. Poorly measured and surrogate variables. Both poorly measured and surrogate variables result in parameter estimates with bias, the direction of which is difficult to determine. AADT estimates are notoriously poorly measured because AADT is meant to measure yearly traffic but is typically estimated by applying seasonal and time factors to correct a count taken over a shorter time period (day or week). All models rely on AADT as a primary determinant of crashes. Surrogate variables used in the models include NODRWY1 and SPD. The former is meant to capture the effects of complex vehicle turning movements near intersections, but simply, the number of driveways is insufficient: a driveway serving a fast-food outlet has different movements and volumes than a driveway serving a manufacturing facility. SPD is meant to capture the effect of vehicle speeds on crash probabilities, but much research has demonstrated that prevailing speeds (e.g., 85th percentile speed) and speed differentials are better indicators of crash potential and severity.

6. Model functional forms. The models postulated in the negative binomial regressions are exponents raised to linear form in parameter functions, resulting in multiplicative factors of variables and their coefficients. Other expressions of variables may be more appropriate,

particularly for the AADT variables, which have been shown to take various multiplicative forms [see, for example, the work by Brundell-Freij and Ekman (10) and Hauer (11, 12)]. At present, crashes are assumed to be proportional to the log of AADT on both major and minor roads, which is a monotonically increasing function. Hauer (13), in recalibrating the IHSDM accident prediction models for two-lane road segments, proposes some unconventional forms in which, for example, the influence of a variable can be represented by a few regression parameters, not a function.

Given the range of potential issues raised with respect to the logical defensibility of published models, there is a priori reason to expect that the models may not validate well across time and space. Validation over space is threatened by Shortcomings 1, 3, 4, 5, and 6, while validation over time is threatened by Shortcomings 2, 5, and 6. The internal validation highlights the difficulties that modelers face when estimating statistical models, especially for forecasting motor vehicle crashes. The six shortcomings identified are in no way confined to the models examined in this research. Most safety studies suffer from these general potential problems; however, carefully designed before-and-after studies can be used to minimize the impacts of Shortcomings 1, 2, and 3.

External Validity: Predictive Ability Across Time

For Model I, a comparison of the parameter estimates in Table 3 for the published model and the one based on the additional years of data showed that variables that were not statistically significant in the original model were not statistically significant in the model for the additional years, with one exception. The variable SPD became highly significant, as opposed to being marginally significant in the published model, while, conversely, RT and RHR became statistically insignificant, with the former having the opposite sign. The GOF in Table 5 for Model I suggested that the published model predicted future crashes as well as it predicted crashes from the calibration data. All of the GOF statistics showed marginal differences between the validation data and the calibration (estimation) data.

For Model II, a comparison of the parameter estimates in Table 3 for the original model and the one based on the additional years of data showed that the constant term and all variables had the same signs, but most had large differences in magnitude, in particular, those with low levels of significance in both models. In terms of the GOF in Table 5, a comparison of MSE per year² (0.10) and MSPE per year² (0.17) suggested that the variability in future crashes was not being captured as well as it was with the data on which the published model is based.

For Model III, comparison of the original model and the model based on the later data in Table 4 revealed that two variables, MEDWDTH1 and NODRWY1, had become statistically insignificant in the model on the basis of additional years of data. In terms of the GOF statistics in Table 5, the linear correlation coefficient showed a moderate decline when it was calculated on the basis of the additional years of data. A comparison of MSE per year² (0.72) and MSPE per year² (0.89) suggested a marginal increase in the lack of fit. Similarly, the MPB per year and the MAD per year showed marginal increases in the lack of fit.

For Model IV, comparison of the original model and the model based on the later data in Table 4 revealed that the model performs very well with the later data: all of the variables in the model were statistically significant in both models. The largest difference was in the

coefficient of the log of ADT2. All other coefficients were similar in magnitude and shared the same sign. In terms of the GOF statistics in Table 5, the linear correlation coefficient showed a marginal increase when it was calculated on the basis of the additional years of data. A comparison of MSE per year² (2.77) and MSPE per year² (2.07) suggested a moderate improvement in the fit to the later data. The MADs per year were similar, but the MPBs per year indicated a marginal increase in the lack of fit.

For Model V, a comparison of the parameter estimates in Table 4 revealed that there were moderate to serious differences between the models. PK%LEFT2 and VEICOM were not statistically significant in the model based on the additional years of data. The GOF statistics in Table 5 revealed a linear correlation coefficient between the observed and the predicted data with a moderate to serious decrease when calculated on the basis of the additional years of data (0.62 to 0.37). A comparison of MSE per year² (6.47) and MSPE per year² (11.04) suggested a poor fit to the additional years of data. The MADs per year showed a moderate increase in the lack of fit, while the MPBs per year indicated a serious lack of fit.

External Validity: Predictive Ability Across Space

For Model I, a comparison of the models with the original data and the Georgia data in Table 3 showed that HI, HAU, and RT had the opposite sign for the Georgia data, with RT becoming insignificant for Georgia. The validation statistics in Table 5 showed that the correlation coefficient (0.66) between observed and predicted crashes was reduced by about 50% (to 0.31) with the Georgia data, a rather serious drop. Other statistics also showed a considerable lack of fit of the Georgia data relative to the Minnesota data. The MSE per year² of 0.19 for the recalibrated model compared with an MSPE per year² of 0.89 for the model with the Georgia data represents a serious difference in fit. This difference suggests that the published model is not capturing the variability in crashes in the Georgia data.

For Model II, a comparison of the models with the original data and the Georgia data in Table 3 showed that the two ADT variables had similar magnitudes and were the only variables that the two models shared that were statistically significant. The variable SPD became statistically significant in the model with the Georgia data, whereas the variable NODRWY1 was statistically significant only in the original model. The validation statistics in Table 5 showed that the correlation coefficient between observed and predicted crashes was reduced by about 50% with the Georgia data, a rather serious drop. Other statistics showed moderate to serious increases in the lack of fit, including a moderate increase in MAD and a serious increase in MSPE per year² (1.73) compared with the MSE per year² (0.10) in the original model.

For Model III, many variables had similar magnitudes across the calibration (California and Michigan) and validation (Georgia) data sets. The one notable exception was MEDWDTH1, which was considerably larger in the Georgia data than in the original data. More than half of the Georgia sites had medians on the major road, whereas only 5.8% of the original sites had medians. Comparison of the models with the original data and the Georgia data in Table 4 showed that the models were quite different. In fact, only the variable log of ADT2 was statistically significant in the model calibrated with the Georgia data, while for the original published model, the log of ADT1, log of ADT2, MEDWDTH1, and NODRWY1 were statistically significant. Thus, a serious lack of agreement between the originally calibrated models and the models estimated with the Georgia data was

observed. All of the validation statistics in Table 5 were consistent with the lack of agreement in model specification between the original data and the Georgia data. The correlation between predicted and observed crashes was seriously different for the Georgia data. The MPB and MAD statistics showed a significant worsening of the fit. Finally, the MSE per year² for the calibration data was considerably lower than the MSPE per year² for the Georgia data. Thus, it was likely that some of the explanatory variables were inadequate for prediction with the Georgia data.

For Model IV, many variables were similar across the calibration data set (California and Michigan) and validation data set (Georgia), with the exception of LTLNS, which was moderately smaller (70% had left-turn lanes) in the calibration data set (Michigan and California) than in the validation data set (Georgia), with 83% left-turn lanes. Comparison of the models with the original data and the Georgia data in Table 4 showed that the coefficients for the log of ADT1 and the log of ADT2 became insignificant in Georgia. All of the validation statistics in Table 5 were at least consistent with the lack of agreement in model specification between the original data and the Georgia data. The correlation between predicted and observed crashes was significantly reduced, going from 0.47 (original) to 0.17 (Georgia). The MPB and MAD statistics showed a moderate increase in the lack of fit. Finally, the MSE per year² (2.77) for the calibration data was moderately lower than the MSPE per year² (3.05) for the Georgia data. Thus, it was likely that some of the explanatory variables were inadequate for predictions with the Georgia data.

For Model V, Table 4 reveals moderate to serious departures between the original published model and the model based on the Georgia data. The variables PROT_LT, log of ADT2, and VEICOM were not statistically significant in the Georgia data. All of the validation statistics in Table 5 were consistent with the serious lack of agreement in model specification between the original data and the Georgia data. The correlation between predicted and observed crashes was substantially reduced, going from 0.62 (original) to 0.23 (Georgia). The MPB per year and MAD per year statistics showed a poor fit. Finally, the MSE per year² for the calibration data was significantly lower than the MSPE per year² for the Georgia data, again suggesting a serious deficiency.

Overall External Model Validation Results

Model I exhibited fairly mixed validation results. The model performed better in the same jurisdiction (Minnesota) in a future time period than in a different jurisdiction. This suggests that variables that quantify differences in facility design, traffic, or environmental conditions across states were missing from the model. Potential missing variables might include the number of wet, icy, or foggy days and the number of crashes in the dark (nighttime) versus the number in the light (daytime). The presence of right-turn lanes can pose a problem in intersection models. If right-turn lanes are installed at intersections mostly because of capacity problems, then their presence might indicate an improvement in the frequency of crashes (because of a reduction in conflicts and crashes). If they are installed because of conditions that lead to a high frequency of crashes or in conjunction with another intersection improvement prompted by conditions that lead to a high frequency of crashes, the presence of a right-turn lane might be associated with larger numbers of crashes. This phenomenon might explain the switching of the sign associated with the RT variable observed in the validation effort associated with Model I and other models. Because the model will be used for predictions in

jurisdictions other than Minnesota, improvements to the model's predictive power are desired.

Model II had fairly mixed validation results, although they were more favorable than those for Model I. The model performed better in the same jurisdiction in a future time period than in a different jurisdiction. This again suggests that variables that quantify differences in facility design, traffic, or environmental conditions across states were missing from the model. The flip-flopping of variables that were statistically significant in the published and the validation models suggests that the sample sizes, in general, were too small to detect the small effects that some of the variables have on safety. As was the case for Model I, improvements to the model's predictive power are desired.

Model III revealed some serious concerns. The model performed moderately well across time but exhibited moderate to serious model specification difficulties. This suggests that variables thought to be important may in fact not be important or were highly correlated with truly important variables. The model was seriously deficient in predictions with the Georgia data, despite an apparent similarity in the raw data. It was clear that variables in the model with the California and Michigan data are not specified appropriately for the Georgia data. In summary, Model III is in need of improvement.

Model IV had deficiencies ranging from moderate to serious. The model performed well across time but was seriously deficient in predictions with the Georgia data, as evidenced by the lack of fit and the inability of the Georgia data to fit the specified models. In summary, Model IV is in need of improvement. The observations suggested that the most important differences were jurisdictional in nature, pointing toward the consideration of design, traffic, and environmental variables that were not included in the model.

Model V revealed model deficiencies ranging from moderate to serious. The inability to re-create the original data hampered the ability to fully validate these models. The model did not appear to perform adequately over time and exhibited a serious lack of fit to the Georgia data. In summary, Model V is in need of improvement.

RECOMMENDATIONS AND FUTURE RESEARCH NEEDS

In general, the external model validations confirmed a priori expectations arising from the internal validation. The models performed sporadically at best and showed that the problems identified in the internal validation translate to the nongeneralizability of models over time and space. Specific lessons were gleaned from the statistical model validation exercise, which can guide some of the directions that future research should take to improve the intersection models and the state of the practice regarding crash model development. Specifically, the research supports the following conclusions and recommendations.

1. Difficulties were introduced by the relatively small sample sizes, small effects, and large variabilities in crash occurrence. Sample sizes are small because of the costs (collection, entry, and archiving) and quality issues associated with the compilation of crash data expected to be representative at a national level. These issues probably pose a significant hurdle to the estimation of more precise crash models. Small safety effects are associated with many proposed design changes, particularly in relation to the uncertainty associated with crash occurrence. For example, installation of lighting will reduce only nighttime crashes and may increase daytime crashes,

and, thus, overall crashes on a percent reduction (AMF) basis will have a small effect that is hard to detect statistically. Improved data quality, increased sample sizes, and random selection of sites for analysis will improve the general ability to detect small safety effects. Large variability in crashes is due, among other factors, to the human-machine interaction in the driving environment. Increasing numbers of in-vehicle distractions coupled with a greater sense of security brought about by vehicle improvements will ensure that variability in crash occurrence will remain a hindrance when crash phenomena are modeled.

2. Alternative functional forms and model specifications, some of which were recommended by the original researchers with the aim of improving the fits of the models, should be explored in future efforts. In particular, the relationship between safety and AADT needs to be carefully specified.

3. The biggest challenges may arise from the omission of variables that influence safety. These include environmental factors such as the number of wet pavement, ice, snow, and poor visibility days. Other potential omitted variables include differences in driving populations, turning movements, and nighttime versus daytime crashes and exposure.

4. Nonrandom sampling also presents potential problems that can be improved with future data quality improvements. When sites are selected for improvement because of crash outcomes or sites are created because of crash outcomes, this can bias the results. Carefully designed before-and-after studies can drastically improve the ability to cope with this problem (9).

Several general recommendations for improving the state of the practice are provided. First, a systematic, rigorous, and carefully designed series of before-and-after studies need to be conducted throughout the United States to enable the comparison of quality results across states. These results could be used to inform IHSDM, would improve the state of knowledge regarding a multitude of safety countermeasures, and would prove to be more reliable than the results of cross-sectional studies. They would also drastically improve problems resulting from the omission of unknown safety effects in statistical models.

Second, data needed to support crash models need to be of similar quality nationwide and need to be standardized. The two most important aspects are exposure estimates and crash reporting. National efforts aimed at improving data quality should continue and be initiated.

Third, analytical methods need improvement. Many countermeasures do not lend themselves well to before-and-after studies, and, thus, cross-sectional studies must be used to estimate safety effects. Examples include horizontal and vertical alignment improvements. At present, however, there are many problems with trying to compare, contrast, and combine results from before-and-after and cross-sectional studies. For example, how could before-and-after study results be used to supplement the data used to estimate the statistical models discussed in this research? Answers to this and similar questions need to be sought.

Finally, this exercise illustrates the critical importance of model validation. Models estimated on the basis of calibration data will undoubtedly perform well, but given the host of potential problems in the conduct of road safety assessments, a true measure of a

model's ability is to perform similarly well with validation data. For a number of reasons enumerated in this paper, crash models that rely solely on multivariate statistical associations will not perform adequately well. Fortunately, the developers of IHSDM acknowledge these problems and have proposed a forecasting approach that is less susceptible to, although not exempt from, these deficiencies. Details of this approach are described by Lyon et al. (pp. 78–86, this volume) and Washington et al. (8).

ACKNOWLEDGMENTS

The authors wish to express their gratitude to Andrew Vogt of Pragmatics, Inc., Yusuf Mohamedshah and Forrest Council of the Landis Corporation, and the Georgia Department of Transportation for their assistance in acquiring data. The authors also express their thanks to Ray Krammes, Michael Griffith, Mohamed Banihashemi, Carol Tan Esse, and Mike Dimaiuta of FHWA for their valuable comments. The authors also acknowledge the Natural Sciences and Engineering Research Council of Canada for its support of some of the fundamental research on which this work is based.

REFERENCES

1. Harwood, D. W., F. M. Council, E. Hauer, W. E. Hughes, and A. Vogt. *Prediction of the Expected Safety Performance of Rural Two-Lane Highways*. Report FHWA-RD-99-207. FHWA, U.S. Department of Transportation, 2000.
2. Pedhazur, E. J., and L. Pedhazur Schmelkin. *Measurement, Design, and Analysis: An Integrated Approach*. Lawrence Erlbaum Associates, Hillsdale, N.J., 1991.
3. Myers, R. H. *Classical and Modern Regression with Applications*, 2nd ed. Duxbury Press, Belmont, Calif., 1990.
4. Neter, J., M. Kutner, C. Nachtsheim, and W. Wasserman. *Applied Linear Statistical Models*, 4th ed. Irwin, Boston, Mass., 1996.
5. Cronbach, L. J. Test Validation. In *Educational Measurement*, 2nd ed. (R. L. Thorndike, ed.), American Council on Education, Washington, D.C., 1971, pp. 443–507.
6. Vogt, A., and J. Bared. *Accident Prediction Models for Two-Lane Rural Roads: Segments and Intersections*. Report FHWA-RD-98-133. FHWA, U.S. Department of Transportation, 1998.
7. Vogt, A. *Crash Models for Rural Intersections: Four-Lane by Two-Lane Stop-Controlled and Two-Lane by Two-Lane Signalized*. Report FHWA-RD-99-128. FHWA, U.S. Department of Transportation, 1999.
8. Washington, S., B. Persaud, J. Oh, and C. Lyon. *Validation and Recalibration of Accident Prediction Models for Rural Intersections*. Final Report. FHWA, U.S. Department of Transportation, 2002.
9. Hauer, E. *Observational Before-After Studies in Road Safety*. Pergamon, Tarrytown, N.Y., 1997.
10. Brundell-Freij, K., and L. Ekman. Flow and Safety—Some Aspects of the Relationship with Special Respect to Unprotected Road Users. Presented at the 70th Annual Meeting of the Transportation Research Board, Washington, D.C., 1991.
11. Hauer, E. Overview. In *ITE, the Traffic Safety Toolbox. A Primer on Traffic Safety*. Institute of Transportation Engineers, Washington, D.C., 1993.
12. Hauer, E. Exposure and Accident Rate. *Traffic Engineering & Control*, Vol. 36, No. 3, 1995, pp. 134–138.
13. Hauer, E. Re-Estimation of Models for Two-Lane Rural Road Segments. Draft working paper. FHWA, U.S. Department of Transportation, 2001.

Publication of this paper sponsored by Committee on Statistical Methodology and Statistical Computer Software in Transportation Research.