USING COMPLIMENTARY SET ANALYSIS TO VALIDATE THE UNDERLYING ASSUMPTIONS OF QUASI-INDUCED EXPOSURE

Xinguo Jiang
Associate Professor, School of Transportation and Logistics, Southwest Jiaotong University, Chengdu China, e-mail: jiangxi1@msu.edu

Yanjun Qiu
Professor, School of Civil Engineering, Southwest Jiaotong University, Chengdu China, e-mail: yjqiu@home.swjtu.edu.cn

Richard W. Lyles
Professor, Department of Civil and Environmental Engineering, Michigan State University, East Lansing MI, USA, e-mail: lyles@egr.msu.edu

Haitao Zheng
Professor, School of Mathematics, Southwest Jiaotong University, Chengdu China, e-mail: htzhang@gmail.com

Lishang Liu
Research Assistant, School of Transportation and Logistics, Southwest Jiaotong University, Chengdu China, e-mail: 315140305@qq.com

Submitted to the 3rd International Conference on Road Safety and Simulation
September 14-16, 2011, Indianapolis, USA

ABSTRACT

In the recent decades quasi-induced exposure has enjoyed increasing popularity with applications in the traffic safety analysis. However, issues have been raised that the majority of the relevant studies do not particularly attempt to verify the validity of the induced exposure technique prior to its adoption. In an effort to validate the critical not-at-fault assumption at the core of the applications, complimentary set analysis (a technique to test whether a driving cohort is randomly selected by its complimentary set of drivers of the same classification) is used and tested. The paper supplements the technique with a comprehensive statistical testing framework, which will enable the validation of the assumption to be conducted for various driver-vehicle characteristics (>2) at much more finely-disaggregated levels. The main findings of the research include: 1) at the most aggregated level, statistical testing does not support the hypothesis that one innocent driver-vehicle combination in the driving population is randomly impacted by the culpable parties of the same classification, mainly due to data aggregation and exposure data irregularities; 2) statistical results demonstrate an increasing trend of p-values when data are finely disaggregated in a stepwise manner, confirming the random-selection assumption of quasi-induced exposure; and 3) an important phenomenon inherent in the exposure matrix is that a driving cohort has a higher probability to collide with the same driving type as opposed to others

of the same classification. Through the study it has been verified that complementary set analysis is a straightforward, convenient, and effective technique to check the validity of quasi-induced exposure.

Keywords: complimentary set analysis, validation, data disaggregation, quasi-induced exposure.

INTRODUCTION

In the field of safety and risk analysis, quantification of exposure is of great importance since it provides safety researchers a means to make normalized comparisons between different cohorts, considering that spatial and temporal circumstances where safety concerns occur may vary considerably. Similarly in traffic safety, although crash frequency can provide valuable insights into some highway safety problems or the effectiveness of certain traffic countermeasures (typically in before-and-after studies), the problem is that the use of crash frequency implicitly assumes there is no significant change of crash exposure during the study period. With recognition of the analytical limitations, a variety of measurements have been proposed to estimate the exposure to driving hazards, among which vehicle miles travelled (VMT) is believed to be the most prominent and widely adopted.

The review of the literature reveals that concerns have been raised regarding the widespread use of VMT in the crash rate calculation. A theoretical assumption of linear conjecture between VMT and the crash frequency has been critically challenged (Janke, 1991; Hauer, 1995) and practically it is virtually impossible to obtain the VMT data at a finely disaggregated level (Lyles et al., 1991). A surrogate exposure measurement in the family of exposure measurements, namely induced exposure, appears to circumvent the drawbacks. The concept of induced exposure was first developed by Haight (1970), and subsequently modified and supplemented with a responsibility-assignment scheme and redefined as "quasi-induced exposure." The theory of quasi-induced exposure is constructed on two fundamental premises: 1) in a two-vehicle crash there are an at-fault driver (D1) and a not-at-fault driver (D2); and 2) the characteristics of the not-at-fault drivers (D2s) in two-vehicle crashes are representative of the general driving population on the road at the time and place of the crash occurrence. An exposure matrix (Lighthizer, 1989) containing D1s as the row and D2s as the column can be constructed to calculate crash propensity for different driving cohorts under different disaggregation levels.

Due to its straightforward nature, quasi-induced exposure has enjoyed increased popularity with applications in the traffic safety analysis in the recent decades. For instance, it was implemented to study specific crash types such as rear-end crashes (Yan et al., 2005a) and red-light running crashes (Yan et al., 2005b), quantify the characteristics of crash propensity of young drivers (Kirk and Stamatiadis, 2001a; McGwin and Brown, 1999) and old drivers (Hing et al., 2003), evaluate the effectiveness of graduated driver licensing program in different states (Jiang and Lyles, 2011; Fohr et al., 2005; Rice et al., 2003), and explore risk factors in epidemiology (Lenguerrand et al., 2008; Lardelli-Claret et al., 2006). Potential issues have been raised that the majority of the relevant studies do not particularly attempt to verify the validity of the induced exposure technique prior to its adoption. Whether a given dataset satisfies the not-at-fault assumption of the quasi-induced exposure dictates the applicability of such an exposure measurement in the ongoing research exercise.

Historical research related to the validation on the underlying assumptions of quasi-induced exposure typically requires substantial efforts of data collection to establish a firm understanding of (driving) exposure "truth" either through sampling, surveying, or other data sources. Kirk and Stamatiadis (2001b) utilized a trip-diary method to estimate the travel exposure in the form of VMT and compared the distributions between the VMT and the estimates through quasi-induced exposure technique in Fayette County, KY. The small sample size of investigated trips from both the trip diary survey and the crashes caused difficulties in accurately estimating VMT and the relative exposure, which did not allow for a meaningful comparison between the two estimates. Stamatiadis and Deacon (1997) compared induced exposure estimates with vehicle classification data under 18 different disaggregation levels, including two development types (rural and urban), three roadway functional classifications (principal, minor arterials, and collectors), and three time periods (day, rush, and night) in Kentucky. Using regression analysis a good correlation between the estimates from the crash data and vehicle classification data was demonstrated and used to justify the selected use of quasi-induced exposure to obtain first-order approximations of relative travel by different road users. In New Zealand, Keall and Newstead (2009) compared the quasi-induced exposure estimates with VMT collected from the odometer readings of the inspected motor vehicles. The research objective was to identify the most appropriate crash type for induced exposure estimation to approximate vehicle distance driven. The results showed that none of the considered crash types demonstrated a good agreement between quasi-induced exposure and vehicle distance traveled. Vehicle distance driven data could not be finely broken down to reflect specific conditions such as lighting and injury levels, which might attribute to the discrepancy.

Furthermore, some researchers test the validity of the not-at-fault hypothesis by means of the information readily available from the crash database. In two similar studies, Chandraratna and Stamatiadis (2009) and Jiang and Lyles (2010) both obtained the exposure estimates from multi-vehicle crashes (>2 vehicles involved in a crash) and then compared it to the relative exposure calculated by quasi-induced exposure. Although different crash databases and different statistical testing methods were employed, the consensus was that no matter how many vehicles were involved in a multi-vehicle crash, the at-fault drivers appeared to collide with the not-at-fault drivers in a non-selective manner and no statistical differences were observed for different characteristics of interest (age, gender, and vehicle type).

Using two-vehicle crash data, Lighthizer (1989) proposed a general framework named "complementary set analysis" to provide a creative avenue to test whether crash victims were randomly impacted. Specifically, the distribution of not-at-fault driver–vehicle cohorts involved in crashes caused by driver–vehicle combinations with certain characteristics is compared with the distribution of not-at-fault driver–vehicle combinations of crashes caused by the complementary set of driver–vehicle combinations. Simply put, if the proportions of a young not-at-fault age group are evenly distributed among the different at-fault age groups (i.e., young, middle-age, and old drivers), it is safe to conclude that the young age group is randomly selected by the driving population and consequently the quasi-induced exposure is applicable to the given dataset. Davis and Gao (1993) improved the technique by taking into account the random variation inherent in the crash data and developed a statistical procedure to calculate confidence bounds and induced exposure estimates for a typical 2 by 2 contingency table. In studying the

crash propensity of drivers with suspended and revoked (S/R) licenses in California, DeYoung et al. (1997) utilized the concept to test the row percentage distributions for drivers with valid, S/R, and no license and the differences were found to be statistically insignificant.

In general, using the crash data (e.g., two-vehicle crashes) to validate the underlying assumptions of quasi-induced poses a number of potential advantages as opposed to the exogenous exposure data: 1) it is readily available and there is no extra data collection effort necessary; 2) it reflects the similar environments when and where the crashes occur; and 3) since the crash records stored in the database are formatted in the same manner, data bias as a result of data inconsistency is likely to be minimized. Although the complementary set analysis is an innovative approach to verify the fundamentals of quasi-induced exposure, the framework has several theoretical limitations which require further improvements. The original idea of complementary set analysis by Lighthizer (1989) does not include statistical testing on different row percentage distributions—the row distributions are compared from the practical angle to see if the maximum percentage difference is greater than an empirical value (e.g., 4%). Davis and Gao (1993) improve the complimentary set analysis with statistical testing to consider the row and column marginal probabilities with the differences between row distributions being tested using a log cross-product ratio statistic. However, the method is only applicable to a 2 by 2 contingency table (e.g., driver gender) and can't be implemented for a table with three or more categories (e.g., young, mid-aged, and old drivers). DeYoung et al. (1997) adopt the two-tailed difference of proportion test in the D1 and D2 exposure matrix, however the method is limited to compare two individual distributions (e.g., valid versus suspended/revoked license drivers) and unable to conduct comparisons among three or more distributions concurrently. These limitations can potentially prevent the widespread application of complimentary set analysis in the effort to validate the underlying assumptions of quasi-induced exposure.

The objective of the research here is to develop a statistical approach that can be used to compare the differences among multiple (two or more) distributions in an integrated manner with the use of complimentary set analysis. Thus, the proposed method will enable the validation of the assumptions to be conducted for different driver-vehicle characteristics at much more finely-disaggregated levels.

METHODOLOGY

Data Preparation

In order to explore this validation approach, the latest crash data (year 2009) were obtained from Michigan Department of Transportation. It has been extensively reported (O'Day, 1993) that the raw crash data have a number of issues such as underreporting, incomplete observations, missing critical information, inaccurate crash data, conflicting information, or inconsistent reporting practices. The quasi-induced exposure technique has stringent requirements on a given dataset, particularly the type of crash data (two-vehicle crash data only) and responsibility assignment. In order that crash data are relatively "clean," a preliminary data screening process is developed to eliminate one-vehicle or three-or-more vehicle crashes, crashes with internally conflicting information (e.g., two-vehicle crashes with information on three vehicles), unreasonable values of key driver-vehicle characteristics (e.g., driver's age below 14), missing or uncoded crash

values, or unreasonable crash types (e.g., head-on crashes occurring on freeway segments). In order to avoid the "negative halo effects" or "crash proneness bias" (DeYoung et al., 1997), the crash responsibility is assigned solely based on the evidence of the hazardous actions instead of driver citation status since quasi-induced exposure is a driving-behavior oriented technique (Jiang and Lyles, 2010). The final dataset includes two-vehicle crash data with crash fault clearly assigned to one of the two involved drivers.

Statistical Testing Framework

As stated, the purpose of the research is to validate the underlying assumptions of quasi-induced exposure with the use of a technique called "complementary set analysis," which was originally introduced by Lighthizer (1989). Central to the technique is that in an exposure matrix with D1s as the row and D2s as the column, the row percentage distributions are compared to identify whether the differences among different rows are statistically significant. The null hypothesis is that there is no difference between row percentage distributions, that is, a specific driving cohort is randomly impacted by different driving cohorts on the road at the time of crash occurrence.

The point here is to develop a general statistical testing framework/approach which can be used in examining the exposure differences for various driving cohorts at multiple (≥2) disaggregation levels. A *K* row (at-fault drivers) by *K* column (not-at-fault drivers) symmetric exposure matrix is constructed (Table 1). Each cell in the matrix indicates the frequency (in numbers) that a certain driver-vehicle characteristic with *K* categories is selected or impacted by the same driving cohorts with *K* classifications. The goal of the statistical testing is to verify whether any type of not-at-fault drivers (e.g., young drivers) is randomly selected by different at-fault drivers within the same grouping categories (e.g., young drivers, mid-aged drivers, and old drivers). When the calculated *p*-value is smaller than $\alpha = 0.05$ (significant level), the null hypothesis is rejected; otherwise the null hypothesis is accepted. The chi-square test extends naturally to the above-described situation:

$$\chi^2 = \sum (f - F)^2 / F \tag{1}$$

where f is the observed frequency in each cell and F is the frequency expected if the null hypothesis of independence holds. The F value is calculated as follows.

Table 1 D1 and D2 matrix for certain driver-vehicle characteristic with K classifications

	Not-at-fault drivers (D2s)					Total
	Categories	1	2	…	K	
At-fault drivers (D1s)	1	n_{11}	n_{12}	...	n_{1K}	n_{1+}
	2	n_{21}	n_{22}	...	n_{2K}	n_{2+}
	⋮	⋮	⋮	⋮	⋮	⋮
	K	n_{K1}	n_{K2}	...	n_{KK}	n_{K+}
Total		n_{+1}	n_{+2}	...	n_{+K}	n

In the population, let p_r be the probability that a crash victim falls in row R and p_c the probability that it falls in column C. Consistent with the hypothesis of independence, the expected number of crash victims F in row R and column C will be np_rp_c, where n is the total number crashes in

Table 1. Take the ratio (row total, n_{i+})/n as the estimate of p_r and the ratio (column total, n_{+j})/n as the estimate of p_c. Let n_{ij} denote the count in the $(i,j)^{th}$ cell of a $K \times K$ contingency table ($i=1,\ldots,K$; $j=1,\ldots,K$), so $n_{i+} = \sum_j n_{ij}$, $n_{+j} = \sum_i n_{ij}$ and $n = \sum_{i,j} n_{ij}$. Then in a symmetric K×K exposure matrix, the expected frequency F in the $(i,j)^{th}$ cell is expressed as:

$$F = \left(n_{i+} \times n_{+j}\right)/n \tag{2}$$

The sum of the deviations (f - F) in each row and column is zero, which dictates the number of degrees of freedom in χ^2. Since there are $(K-1)\times(K-1)$ deviations in the $K\times K$ exposure matrix, the degree of freedom of the chi-square is $(K-1)\times(K-1)$. The hypothesis that the innocent drivers are randomly selected or impacted by the culpable drivers H_0: $\boldsymbol{\pi_{ij} = \pi_{i+} \cdot \pi_{+j}}$, where $\boldsymbol{\pi_{i+}} = \Sigma_{j=1}^{K}\pi_{ij}$, $\hat{\pi}_{ij} = \frac{n_{ij}}{n}$, , and $\boldsymbol{\pi_{+j}} = \Sigma_{i=1}^{K}\pi_{ij}, i, j = 1, \cdots, K$, can be tested, using

$$X^2 = \sum_i^K \sum_j^K \frac{\left[n_{ij} - (n_{i+}\times n_{+j})/n\right]^2}{[(n_{i+}\times n_{+j})/n]} \dot{\sim} \chi^2([K-1][K-1]) \tag{3}$$

The above structure is used in analyzing the patterns associated with the distributions for three key driver-vehicle characteristics, that is, driver gender (male and female), vehicle types (passenger cars, pickups, and heavy trucks), and driver age (young, mid-young, mid-old, and old drivers). The D2 percentage distributions of the variables of interest will be computed at the overall level (the population as a whole) and then the levels are gradually stratified to reflect more specific circumstances in the order of day of week (weekday versus weekend), development types (urban versus rural areas), roadway functional classifications (locals versus collectors or arterials), and time period (AM versus PM hours). Statistical tests are run at each individual level for three main characteristics to see if the underlying assertion of quasi-induced exposure can be effectively validated.

RESULTS

Overall Level

Tables 2, 3 and 4 are the results of complementary set analysis for three variables of interest at the overall level (without any data disaggregation). The number in each cell represents the frequency of a non-responsible driver (D2, shown in the column) collided by a corresponding responsible driver (D1, shown in the row), while the percentage corresponds to the relative proportion of the column classification (different D2 drivers) within each row. Also displayed is the relative crash involvement ratio (IR) calculated by dividing the percentage who are at-fault by the percentage who are innocent for each group. IR is an indicator of relative crash over- or under-involvement for each driver group—depending on the values of IR, one can show whether any specific driver-vehicle combination is disproportionately over-involved (IR>1) or under-involved (IR<1) in crashes relative to its proportion in the driving population.

Tables 2, 3, and 4 demonstrate that the involvement ratios for each individual driver-vehicle type generally conform to a *priori* expectation or known knowledge. For example, the IR for young drivers (Table 4) is 1.48, illustrating that they cause proportionally more crashes than their

existence in the driving population, while the IR for the mid-aged (young or old) drivers suggests the opposite. As for male drivers and pickup vehicles, the IRs indicate a relatively higher crash involvement compared to their counterparts. All these observations appear to be in good agreement with current research findings (Shinar 2007).

Table 2 The characteristics of different driver genders at overall level (Michigan 2009)

	Driver gender	Not-at-fault drivers (D2s), N(%)			*p*-value
		Female	Male	Marginal totals	
At-fault drivers (D1s)	Female	24969(**48.9%**)	26065(51.1%)	51034	<0.0001
	Male	27228(47.2%)	30459(**52.8%**)	57687	
	Marginal totals	52197	56524	108721	
	IRs	0.98	1.02		

Table 3 The characteristic of different vehicle types at overall level (Michigan 2009)

	Vehicle types	Not-at-fault drivers (D2s), N(%)				*p*-value
		Passenger cars	Pickups	Heavy trucks	Marginal totals	
At-fault drivers D1s	Passenger cars	80525(**86.7%**)	10377(11.2%)	1973(2.1%)	92875	<0.0001
	Pickups	11855(84.2%)	1925(**13.7%**)	294(2.1%)	14074	
	Heavy trucks	1473(83.1%)	208(11.7%)	91(**5.1%**)	1772	
	Marginal totals	93853	12510	2358	108721	
	IRs	0.99	1.13	0.75		

Table 4 The characteristics of different age groups at overall level (Michigan 2009)

	Age group	Not-at-fault drivers (D2s), N(%)					*p*-value
		Young	Mid-young	Mid-old	Old	Marginal totals	
At-fault drivers (D1s)	Young	7687(**23.0%**)	9768(29.2%)	11249(33.6%)	4751(14.2%)	33455	<0.0001
	Mid-young	5791(19.7%)	9122(**31.1**%)	10199(34.7%)	4253(14.5%)	29365	
	Mid-old	5804(19.6%)	8954(30.3%)	10396(**35.1**%)	4430(15.0%)	29584	
	Old	3272(20.1%)	4718(28.9%)	5521(33.8%)	2806(**17.2**%)	16317	
	Marginal totals	22554	32562	37365	16240	108721	
	IRs	1.48	0.90	0.79	1.00		

Important information exhibited in Table 2 is that the percentages along the diagonal of the exposure matrix are consistently greater than other values within the same column, which are noted as bold in each table. It tells that the drivers of the same characteristic are inclined to collide with each other rather than with others. For the example of not-at-fault young drivers, they have a higher probability of being selected by the same young drivers than by the mid-aged or old drivers. The phenomenon can be partially attributed to the similar driving behaviors and experiences demonstrated and roadway circumstances (e.g., roadways, time of day) traveled by the same driving cohort, relative to other types in the same classification.

Comparison of the percentage distributions of not-at-fault drivers among different rows suggests that the maximum difference for three driver-vehicle characteristics is consistently smaller than 4 percentage points. Although the threshold value of 4 percentage points is somewhat empirical, it reflects most of the natural data variations for the driver–vehicle characteristics of interest (Jiang and Lyles, 2010). Thus, from the practical point of view it is safe to argue that the differences of the row percentage distributions are insignificant and for each variable of interest the non-responsible drivers are randomly impacted by other vehicles of the same classification. Although

operationally the row percentage distributions vary within a reasonable and acceptable range, the chi-square p-values are shown to be smaller than 0.05, rejecting the hypothesis that the differences between rows are statistically insignificant. Considering that the proposed statistical testing method aims to compare the disparity of all the row distributions concurrently, relatively high proportions of crash frequency between the same driving cohorts (e.g., younger drivers with younger drivers) are identified to be the principal contributing factor to the rejection of the hypothesis. Evidentially, the high probability of collisions between the same driver types creates an imbalance among the row percentage distributions within each column.

Disaggregated Levels

One of the essential merits of quasi-induced exposure is that it can be used for the analysis of crash involvement at finely disaggregated levels. An attempt is also made to analyze the exposure change for different age groups and vehicle types, the purpose of which is to identify the transition pattern under various stratification levels of circumstances when or where crashes occur. The levels are disaggregated by temporal and spatial parameters in a superimposed manner. For age groups, the levels start from the overall level to weekend, then to urban area, local streets, and PM hours; for vehicle types, the disaggregation levels include overall, weekday, rural area, and arterials.

Tables 5 and 6 are illustrations of the characteristic distributions for vehicle types and age groups, respectively under various disaggregation levels. Also shown are the involvement ratios and chi-square p-values. In general, the observations offered at the overall level can also be applicable to the data at various disaggregation levels. The cells marked as bold along the diagonal of the exposure matrix indicate that the same vehicle types or age groups tend to collide with each other compared to its counterparts. As for pickups, young and old drivers the involvement ratios consistently have higher crash propensity (IR>1), and generate relatively more crashes in contrast to their corresponding proportions in the driving population.

Table 5 The characteristic distributions of vehicle types at different disaggregated levels

Ordered levels	Vehicle types (D1s)	Not-at-fault drivers (D2s), N(%)				*p*-value
		Passenger cars	Pickups	Heavy trucks	Marginal totals	
Weekday	Passenger cars	49535**(86.4%)**	6400(11.2%)	1417(2.5%)	57352	0.021
	Pickups	7119(82.5%)	1299**(15.1%)**	207(2.4%)	8625	
	Heavy trucks	1061(80.9%)	150(11.4%)	101**(7.7%)**	1312	
	Marginal totals	57715	7849	1725	67289	
	IRs	1.00	1.10	0.76		
Rural	Passenger cars	3669**(73.6%)**	1052(21.1%)	262(5.3%)	4983	0.061
	Pickups	1080(69.9%)	394**(25.5%)**	71(4.6%)	1545	
	Heavy trucks	157(70.7%)	37(16.7%)	28**(12.6%)**	222	
	Marginal totals	4906	1483	361	6750	
	IRs	1.02	1.04	0.61		
Arterials	Passenger cars	831(69.9%)	222(18.7%)	135(11.4%)	1188	0.078
	Pickups	220(71.4%)	61**(19.8%)**	27(8.8%)	308	
	Heavy trucks	65(73.9%)	6(6.8%)	17**(19.3%)**	88	
	Marginal totals	1116	289	179	1584	
	IRs	1.06	1.07	0.49		

Table 6 The characteristic distributions for age groups at different disaggregation levels

Ordered levels	Driver age (D1s)	Not-at-fault drivers (D2s), N(%)					*p*-value
		Young (<25)	Mid-young (25-40)	Mid-old (41-59)	Old (>59)	Marginal totals	
Weekend	Young	3117(**23.7%**)	3816(29.0%)	4409(33.5%)	1830(13.9%)	13172	0.007
	Mid-young	2402(21.3%)	3544(**31.4%**)	3704(32.9%)	1623(14.4%)	11273	
	Mid-old	2269(20.8%)	3325(30.4%)	3734(**34.2%**)	1603(14.7%)	10931	
	Old	1236(20.4%)	1803(29.8%)	1993(32.9%)	1024(**16.9%**)	6056	
	Marginal totals	9024	12488	13840	6080	41432	
	IRs	1.46	0.90	0.79	1.00		
Urban	Young	2673(**23.4%**)	3379(29.6%)	3788(33.2%)	1564(13.7%)	11404	0.018
	Mid-young	2084(21.0%)	3168(**32.0%**)	3282(33.1%)	1377(13.9%)	9911	
	Mid-old	1945(20.6%)	2911(30.9%)	3213(**34.1%**)	1361(14.4%)	9430	
	Old	1140(22.2%)	1548(30.1%)	1693(33.0%)	757(**14.7%**)	5138	
	Marginal totals	7842	11006	11976	5059	35883	
	IRs	1.45	0.90	0.79	1.02		
Locals	Young	324(**28.6%**)	345(30.4%)	331(29.2%)	134(11.8%)	1134	0.191
	Mid-young	247(25.8%)	299(**31.2%**)	289(30.1%)	124(12.9%)	959	
	Mid-old	213(22.4%)	292(30.7%)	307(**32.3%**)	138(14.5%)	950	
	Old	106(20.6%)	145(28.2%)	171(33.3%)	92(**17.9%**)	514	
	Marginal totals	890	1081	1098	488	3557	
	IRs	1.27	0.89	0.87	1.05		
PM	Young	221(26.2%)	263(31.2%)	240(28.5%)	119(14.1%)	843	0.729
	Mid-young	199(29.8%)	184(27.5%)	195(29.2%)	90(13.5%)	668	
	Mid-old	178(27.1%)	191(29.1%)	215(32.7%)	73(11.1%)	657	
	Old	95(27.8%)	94(27.5%)	112(32.7%)	41(12.0%)	342	
	Marginal totals	693	732	762	323	2510	
	IRs	1.22	0.91	0.86	1.06		

The cross-comparison between the overall and disaggregated levels serves to highlight three important issues. First, the operational difference among the row percentage distributions within each column gradually increases when the exposure is calculated at a finer stratification level. The phenomenon can be accredited to the smaller sample size of crash data as a result of data stratification to reflect a specific condition. For example in Table 5, there are only six (6) crashes of pickup trucks collided by heavy trucks on rural arterials in the weekdays, while the maximum row percentage difference reaches as much as 13%. Second, for the statistical testing it appears that the p-values increase considerably with the disaggregation levels. For rural areas and arterial roadways (Table 5), no statistical differences are found among different distributions for vehicle types ($p>0.05$). For local streets and in PM hours (Table 6), the differences among row percentage distributions for various age groups are not statistically significant. Consequently under these disaggregation conditions, the not-at-fault drivers are considered to be randomly selected by the culpable drivers at the time of crash occurrence. Third, compared to other driver-vehicle types of the same classification, the row percentages and IRs for passenger cars and mid-aged drivers remain relatively stable across the diverse data disaggregation spectrum. A plausible explanation is that these drivers are not particularly specific to certain temporal or spatial circumstances when or where crashes occur and thus their exposures on the roadway network may not be as sensitive to the diversified environmental settings.

In order to further explore how the exposure for different driver-vehicle characteristics vary with the data disaggregation, Figures 1 and 2 graphically illustrate the average D2 percentages for different vehicle types and age groups under different disaggregation conditions.

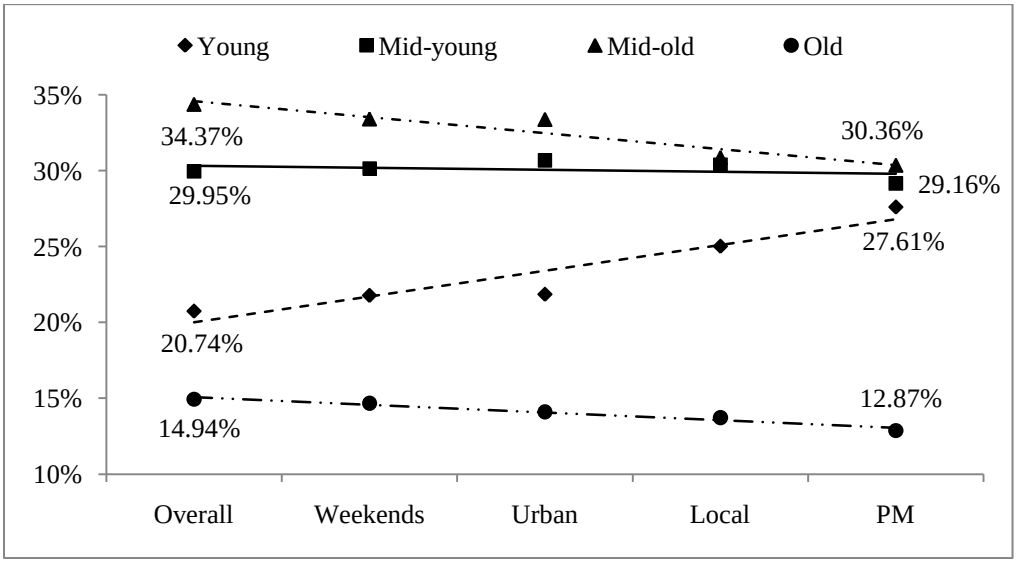

Figure 1 Average D2 percentage for different age groups under various disaggregation levels

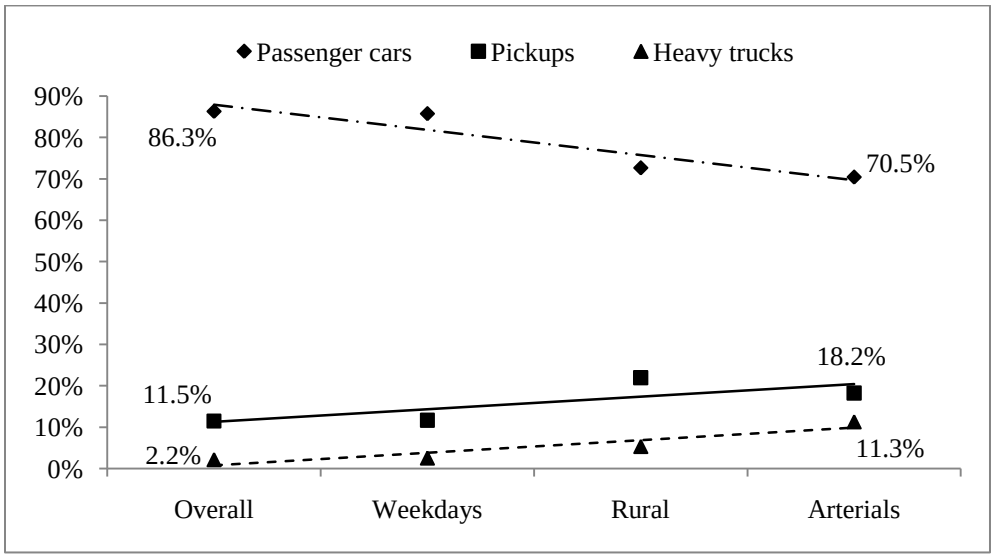

Figure 2 Average D2 percentage for different vehicle types under various disaggregation levels

Figure 1 shows that the relative proportions of mid-aged drivers approximately stay unchanged, while for young drivers the percentages increase considerably from 20.74% at the overall level to 27.61% on urban local streets in the weekend PM hours. Conversely, the percentages of old drivers (60+) decline steadily across the same levels of disaggregation. These outcomes are in accordance with the reality that young drivers are more active in the urbanized areas at the weekend night whereas the old drivers become less active. Similarly, figure 2 depicts the trend that the proportions of pickups and heavy vehicles gradually increase when crash data are

classified to specify the weekday, rural areas, and freeways in a stepwise manner, while the percentages for passenger cars show the opposite tendency. The phenomenon can be well explained by the fact that the rural arterials (mainly freeways and partially principal and minor arterials) are particularly favored by the heavy trucks compared with the local streets in the weekdays (Stamatiadis and Deacon, 1997).

DISCUSSION

The principal objective of the study was to further develop complimentary set analysis with a theoretical framework of statistical testing in the context of validating the underlying assumptions of quasi-induced exposure. The goal is to enable D2 row percentage distributions for multiple column classifications (>2) to be compared concurrently in the exposure (D1-D2) matrix. With the aid of complimentary set analysis, comparisons were conducted for three key driver-vehicle characteristics at a variety of data disaggregation levels (e.g., the overall level, time of day, hour of day) to see if the fundamental assertions of quasi-induced exposure are satisfied for a given dataset. There are three main findings from the exercise. First, at the overall level the statistical testing does not seem to support the hypothesis that one non-responsible driver-vehicle combination in the driving population is randomly impacted by the culpable parties of the same classification; in other words, the innocent drivers involved the crashes may not be reasonably representative of the driving population. From an operational standpoint, the maximum difference among the row percentage distributions appears to fall within a natural data variation or noise. Second, with the crash data finely disaggregated to reflect more constrained circumstances, the statistical testing results demonstrate an increasing trend of p-values, suggesting that the difference between the D2 row percentage distributions is becoming statistically insignificant and consequently the stratified dataset satisfies the random-selection assumption of quasi-induced exposure. Third, an important phenomenon inherent in the exposure matrix is that a driving cohort has a higher probability to collide with the same driving type as opposed to others of the same classification, due to the similarity of driving behaviors and driving environments. Through the study it has been verified that complementary set analysis is a straightforward and effectual technique to check the validity of quasi-induced exposure.

The complimentary set analysis fails to validate the not-at-fault assumptions of quasi-induced exposure with the proposed chi-square testing approach at the overall level, which is somewhat at odds with the past research findings (Jiang and Lyles, 2010). However, with the data gradually stratified over specific temporal and spatial conditions, the *p*-values given by the statistical testing start to increase and suggest that the underlying assumptions are generally met. Obviously, data disaggregation plays an important role in these validation efforts, which has been previously proven to reduce exposure data irregularities (Chandraratna and Stamatiadis, 2009). The exposure data irregularities typically occur at the overall level where the exposure estimate for a driver-vehicle combination is computed in an aggregated manner to represent various crash environments (e.g., time of day, weather conditions, day of week, roadway functional classification, and levels of land use development). For example in Figure 1, the young drivers account for approximately 20.7% of total driving population as a whole, increasing to 21.8% in the weekdays, 21.9% in the urban areas, 25.0% on the local streets, and jumping to 27.6% during the PM hours. Since for young drivers there is an inherent characteristic of changing exposure proportions under various circumstances, it can reasonably explain why the compounded

exposure distribution (e.g., at the overall level) is less accurate to represent a random sample of its driving population. Also evidenced from Figure 1 that data disaggregation has less influence on the exposure proportions of mid-aged drivers (25-40) over the same disaggregation strata: their exposures remain fairly stable over the spectrum. Consequently, the finding of the research emphasizes that the quasi-induced exposure technique can be more beneficial to those driving cohorts with varying exposures in related to data disaggregation.

Restrictions are placed on the extent of how finely the crash data can be disaggregated in order to maintain the row stability of the exposure matrix. As demonstrated from tables 5 and 6, the frequencies (representing the relative exposure) are substantially reduced with the stratification of crash data, while some practical discrepancy of row percentage distributions exceeds the threshold value (4 percentage points). When the sample size is relatively small, the data noise and/or natural data variation start to phase in and play a significant role, which can eventually dominate the testing results. In order to hold to the underlying assumptions of quasi-induced exposure, the crash sample size needs to be reasonably large to allow both statistical and practical significance to be examined. The research has identified an inherent pattern that the same driving cohort has the inclination to collide with each other especially when data are analyzed in an aggregated mode; however, for young drivers in Table 6 the pattern seems to decrease when the data are more highly disaggregated. Based on the chi-square calculation procedure, an unusual high proportion of innocent drivers within each column can skew the chi-square statistic and consequently the test may produce a small *p*-value. Comparatively, the high probability of collisions between the same driver-vehicle types affects the results of the statistical testing more at the aggregated levels than at the finely disaggregated levels (the pattern is gone at the last disaggregation level in Table 6 where the null hypothesis is accepted). This reinforces the notion that reasonable data homogeneity in the driver/vehicle population shall be ensured when quasi-induced exposure technique is implemented in the risk analysis.

Using the complimentary set analysis to validate the underlying assumptions of quasi-induced exposure has shown its promising capability and potential for widespread applications. First, the complimentary set analysis can be easily deployed on the exposure (D1-D2) matrix. Compared with using three-or-more vehicle crashes (Chandraratna and Stamatiadis, 2009; Jiang and Lyles, 2010) or external exposure "truth" (Kirk and Stamatiadis, 2001b) to accomplish the validation, the technique requires two-vehicle crash data only and there is no need for additional exposure data as the control reference. Second, the proposed testing framework for the complimentary set analysis is simplistic in nature and the resulting crash propensity from different age groups and vehicle types generally matches the priori expectations and current knowledge base. Third, the complimentary set analysis poses the ability to check the validity of quasi-induced exposure at finely disaggregated levels. This capacity matches with the strength of quasi-induced exposure to study the crash propensity of certain driver-vehicle characteristics at specified circumstances. This is important, because without a positive validation, the quasi-induced exposure method should not be used. It may be ideal to use VMT data to conduct the validation at a highly stratified level, however the validation as such can't be achieved since it is practically infeasible to obtain VMT at the same level of disaggregation.

There are also limitations associated with the proposed theoretical testing framework. The chi-square testing method appears to be sensitive to the sample size of the cells in the exposure

matrix. With identical row percentage distributions, the chi-square test is prone to reject the hypothesis when the sample size is comparatively large and accept the hypothesis when the sample size is relatively small. A methodological issue is related to the choice of threshold value in determining the operational significance. The four (4) percentage points are an empirical value and mainly depend on pragmatic practice and judgment so it may be arbitrary.

The aforementioned limitations serve to suggest directions for the future research. A modified testing method should be pursued to conduct the analysis based on cell percentages rather than actual frequencies and thus mitigate the impacts of the sample size on the test results. After all, the percentage distributions are of the most concern in the context of validating the not-at-fault conjecture of quasi-induced exposure. An alternative solution is to use statistical and operational testing methods in a combinatory manner (Jiang and Lyles, 2010). One of the pitfalls of the statistical significance test is that it is unable to assess the probability that two samples were obtained by chance or if the D2 row distributions were simply atypical. The row percentage distributions in the D1-D2 matrix can be statistically different due to a relatively large sample size, but operationally insignificant due to a small effect. Therefore, it is essential that operational significance be considered in combination with statistical significance. From this perspective, further research effort is warranted to develop a more justifiable value for the operational significance.

In summary, the research effort was mainly focused on providing a statistical testing framework to the complimentary set analysis in terms of validating the not-at-fault assumptions of quasi-induced exposure. The method has demonstrated that at a finely disaggregated level the not-at-fault drivers are randomly selected by the drivers of the same classification. Thus, the method has manifested its great potentials to estimate the relative exposure particularly at a highly stratified temporal or spatial circumstance. Considering the simplicity, the complementary set analysis will become a useful and convenient hand-on tool to validate the fundamentals of quasi-induced theory before the exposure measurement can be implemented in the real application.

ACKNOWLEDGEMENTS

The research is supported by the Fundamental Research Funds for the Central Universities (SWJTU09CX042) and National Science Foundation of China (NSFC-50978222). The authors thank Michigan Department of Transportation (MDOT) for the crash data.

REFERENCES

Chandraratna, S. and Stamatiadis, N. (2009). “Quasi-induced exposure method: evaluation of not-at-fault assumption,” *Accident Analysis and Prevention* 41, 308-313.

Davis, G.A., and Gao, Y. (1993). “Statistical methods to support induced exposure analyses of traffic accident data,” *Transportation Research Record* 1401, 43-48.

DeYoung, D.J., Peck, R.C., and Helander, C.J. (1997). “Estimating the exposure and fatal crash rates of suspended/revoked and unlicensed drivers in California,” *Accident Analysis and Prevention* 29, 17-23.

Fohr, S.A., Layde, P.M., and Guse, C.E. (2005). "Graduated driver licensing in Wisconsin: does it create safer drivers?" *Wisconsin Medical Journal* 104, 31–36.

Haight, F.A. (1970). "A crude framework for bypassing exposure," *Journal of Safety Research* 2, 26-29.

Hauer, E. (1995). "On exposure and accident rate," *Journal of Traffic Engineering and Control* 36(3), 134-138.

Hing, J., Stamatiadis, N., and Aultman-Hall, L. (2003). "Evaluating the impact of passengers on the safety of older drivers," *Journal of Safety Research* 34, 343– 351.

Janke, M.K. (1991). "Accidents, mileages, and the exaggeration of risk," *Accident Analysis and Prevention* 23 (2), 183-188.

Jiang, X., and Lyles, R.W. (2010). "A review of the validity of the underlying assumptions of quasi-induced exposure," *Accident Analysis and Prevention* 42, 1352-1358.

Jiang, X., and Lyles, R.W. (2011). "Exposure-based assessment of the effectiveness of Michigan's graduated driver licensing nighttime driving restriction," *Safety Science* 49, 484–490.

Keall, M.D., and Newstead, S. (2009). "Selection of comparison crash types for quasi-induced exposure risk estimation," *Traffic Injury Prevention* 10, 23–29.

Kirk, A., and Stamatiadis, N. (2001a). "Crash rates and traffic maneuvers of younger drivers," *Transportation Research Record* 1779, 68-74.

Kirk, A., and Stamatiadis, N. (2001b). "Evaluation of Quasi-Induced Exposure," *Final report*, University of Kentucky, Southeastern Transportation Center.

Lardelli-Claret, P., Jimenez-Moleon, J.J., Luna-del-Castillo, J.d., Garcia-Martin, D.M., Moreno-Abril, O., and Bueno-Cavanillas, A. (2006). "Comparison between two quasi-induced exposure methods for studying risk factors for road crashes," *American Journal of Epidemiology* 163, 188–195.

Lenguerrand, E., Martin, J.L., Moskal, A., Gadegbeku, B., and Laumon, B. (2008). "Limits of the quasi-induced exposure method when compared with the standard case-control design," *Accident Analysis and Prevention* 40, 861-868.

Lighthizer, D.R. (1989). "An Empirical Validation of Quasi-Induced Exposure," Ph.D Dissertation, Department of Civil and Environmental Engineering, Michigan State University, E. Lansing, MI.

Lyles, R.W., Stamatiadis, P., and Lighthizer, D.R. (1991). "Quasi-induced exposure revisited," *Accident Analysis and Prevention* 23(4), 275-285.

McGwin, G., and Brown, D.B. (1999). "Characteristics of traffic crashes among young, middle-aged, and older drivers," *Accident Analysis and Prevention* 31,181–198.

O'Day, J. (1993). "Accident Data Quality," NCHRP Synthesis of Highway Practice, 192, 54p.

Rice, T.M., Peek-Asa, C., and Kraus, J.F. (2003). "Nighttime driving, passenger transport, and injury crash rates of young drivers," *Injury Prevention* 9, 245–250.

Shinar, D. (2007). "Traffic Safety and Human Factor," Elsevier, Amsterdam, the Netherlands.

Stamatiadis, N., and Deacon, J.A. (1997). "Quasi-induced exposure: methodology and insight," *Accident Analysis and Prevention* 29, 37-52.

Yan, X., Radwan, E., and Abdel-Aty, M. (2005a). "Characteristics of rear-end accidents at signalized intersection using multiple logistic regression model," *Accident Analysis and Prevention* 37, 983–995.

Yan, X., Radwan , E., and Birriel, E. (2005b). "Analysis of red light running crashes based on quasi-induced exposure and multiple logistic regression method," *Transportation Research Record* 1908, 70–79.