This is to certify that the
dissertation entitled

An Investigation of the Underlying Assumptions of
Quasi-Induced Exposure

presented by

Xinguo Jiang

has been accepted towards fulfillment
of the requirements for the

Ph.D. degree in Civil & Environmental Engineering

B Hanichandran (for R. Lyles)
Major Professor's Signature

May 12, 2005
Date

MSU is an Affirmative Action/Equal Opportunity Institution

AN INVESTIGATION OF THE UNDERLYING ASSUMPTIONS OF QUASI-INDUCED EXPOSURE

By

Xinguo Jiang

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Civil and Environmental Engineering

2005

ABSTRACT

AN INVESTIGATION OF THE UNDERLYING ASSUMPTIONS OF QUASI-INDUCED EXPOSURE

By

Xinguo Jiang

Traditionally, the measures of exposure fall into two general categories: direct and indirect. The former includes vehicle miles traveled (VMT), number of entering vehicles (NEV) for an intersection, and annual daily traffic (ADT). Indirect exposure is normally referred to as induced exposure. VMT is the most commonly used exposure measure in traffic safety/crash related analyses. However, VMT as an exposure measurement has also aroused criticism among traffic researchers: the underlying assumptions of VMT-based methods have been challenged. In addition, the use of VMT confronts two fundamental drawbacks in practical applications: general availability of data and finer disaggregation of exposure.

Quasi-induced exposure, an approach to estimate *relative* exposure, is capable of overcoming those difficulties confronted by the VMT method. Quasi-induced exposure has been employed by a number of traffic researchers and has demonstrated its strength in a variety of practical applications. For example, it is relatively easy to use; incorporated accident data are generally available; it is able to measure the exposure of a specific driver cohort under specific conditions (e.g., young male on Friday night on local streets). However, the theory of quasi-induced exposure is far from being perfect. A set of issues needs to be addressed: 1) there is a lack of a systematic procedure to prepare accident data; 2) there are problems involved with the responsibility assigning scheme; 3)

more importantly, there are few attempts to validate the underlying assumption: non-responsible drivers in two-vehicle accidents constitute a random sample of driving population on the road.

Two techniques have been developed to validate the underlying assumption: 1) to compare the relative exposure derived from accident data using quasi-induced exposure with the "true" exposure collected from other sources, such as VMT data, seat belt use data, and truck volume data; 2) to compare the distributions for non-responsible drivers derived from two-vehicle and three-or-more-vehicle accidents.

By means of addressing these three stated issues, this research aims to develop a guideline with regard to when to, or not to, use the quasi-induced exposure technique, and to provide a systematic procedure to manipulate accident data and assign accident faults if the validity of quasi-induced exposure is ensured.

ACKNOWLEDGMENTS

When I am moving toward the finish line of my Ph.D research, I look back on the several years at Michigan State University and I feel grateful to a number of people who contributed their efforts in accomplishing this work.

I would like to express my gratitude to my academic advisor Dr. Richard W. Lyles for his encouragement, sternness, inspiration, and persistent tutoring. I truly feel indebted—his support to me has gone beyond the relationship between an advisor and a student.

I would like to thank Dr. William C. Taylor for his thoughts, comments, and suggestions in the course of this research. I feel thankful, because I can't remember how many times I bugged him by knocking at his door for answers. In the meanwhile, I would like to thank the other two committee members, Dr. Karim Chatti and Dr. Dennis C. Gilliland, for serving in my doctoral committee and for their helpful comments and suggestions.

Last, I would like to thank my wife, Fenghua Lin, for her constant support and for her loving, caring, and understanding, and to my parents for their unconditioned love and sacrifice from the other side of the Pacific Ocean.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

A wide range of topics has been discussed in the context of measuring roadway safety and/or accident risk. These include accident frequencies, accident rates, exposure, induced exposure, and quasi-induced exposure. The analysis of accident frequency can provide valuable insights into some highway safety problems and the effectiveness of certain traffic countermeasures (typically in before-and-after studies). However, the problem with using accident frequency in traffic safety analysis is the implicit assumption that there is no significant change of accident exposure during the analysis period. Recognizing this limitation, traffic engineers have also been interested in knowing the "exposure" of different driver-vehicle combinations to different driving hazards, in addition to the frequency of accidents involving these different driver-vehicle combinations. In this context, accident rates are generally expressed as the ratio of accident frequency to exposure. Thus, quantification of exposure is of great importance—it provides researchers an opportunity to make normalized comparisons between driver groups and to accurately represent the circumstances where and/or when accidents occur. Unfortunately, quantification of exposure, while conceptually straightforward, is difficult in practice.

Traditionally, the measures of exposure fall into two general categories: direct and indirect. The former includes vehicle miles traveled (VMT), number of entering vehicles

(NEV) for an intersection, and annual daily traffic (ADT). Indirect exposure is normally referred to as induced exposure.

VMT is the most commonly used exposure measure in traffic safety/crash related analyses. However, VMT as an exposure measurement has been criticized in some instances. For example, Steward (1960) challenged one of the assumptions that all driving involved the same exposure to accident hazards, by giving an example that "when vehicles travel in a platoon at an identical speed, the leading vehicle might experience more driving hazards than the following vehicles." In addition, the use of VMT has two fundamental drawbacks in practical applications: general availability of data and finer disaggregation of exposure (Lighthizer 1989, Lyles et al. 1991). More specifically, general availability of data refers to the fact that computation of VMT for a driving cohort requires traffic volume and travel distance data. Unfortunately, under most circumstances such data are not readily available. It is virtually impossible to calculate VMT for a specific driving cohort, disaggregated by specified spatial and temporal parameters, e.g., young drivers on local highways on Friday nights.

In light of the theoretical and operational problems involved in using VMT, researchers developed an alternative approach to estimate *relative* exposure to accidents by using the accident statistics themselves, namely, induced exposure. Since accident data are more readily available, exposure can be directly estimated without requiring additional information (e.g., volume data, travel distance), which is necessary for VMT-based methods. Another desirable attribute of induced exposure is the capability of disaggregating exposure by specific variables of interest—e.g., roadway type, on driver age. In 1964, Thorpe developed the idea of induced exposure. However, induced

exposure theory was found to have some problems, especially in the scheme for assigning responsibility for the accident. Haight (1971) modified Thorpe's original work and supplemented it with a systematic responsibility-assignment scheme. The revised method is defined as "quasi-induced exposure."

The basis of quasi-induced exposure is founded of two fundamental assumptions (Lyles 1994):

1. In at least some two-vehicle accidents there is an at-fault and a not-at-fault driver.
2. Not-at-fault drivers in two-vehicle accidents are a random sample of motorists and vehicles on the road at the time of the accident.

For the **first assumption**, quasi-induced exposure requires the utilization of only two-vehicle accident data with one at-fault or responsible driver and one not-at-fault or non-responsible driver. Accident responsibility for causation is typically assigned to one of the drivers in a two-vehicle accident based on a police accident report (e.g., Carr 1969, Hall 1970, and Carlson 1970). The driver-vehicle combination that is responsible for the accident is defined as Driver-1 or D1. Consistent with the above, the not-at-fault driver-vehicle combination is defined as Driver-2 or D2. The **second assumption** can be rephrased based on the terms defined: D2s are randomly "selected" by D1s from all vehicles existing on the system at the time of the accident and, thus, D2s constitute a random sample of driver-vehicle combinations and, inductively, a measure of exposure (Lyles 1994).

Quasi-induced exposure can't be used with confidence unless these two assumptions have been shown to be reasonable. In this context, the research effort here focuses on validating the fundamental premises of quasi-induced exposure from

empirical and theoretical perspectives. It incorporates 1) use of different sources of data, serving as accident exposure "truth," to compare with the *relative* exposure estimated using the quasi-induced approach and 2) comparison of D2 distributions between three-or-more-vehicle accidents and two-vehicle accidents. Other goals of this research are to address several issues relevant to quasi-induced exposure on the condition that the exposure technique is useful: developing systematic rules for preparing accident data, assigning responsibility, and exploring its theoretical difficulties.

The next chapter is a literature review covering topics related to quasi-induced exposure.

Chapter 2

LITERATURE REVIEW

2.1 Introduction

Under the general subject of traffic safety, considerable work has been done on the measurement of safety and/or accident risk. This includes accident frequencies, and rates as well as exposure, induced exposure, and quasi-induced exposure. The research effort here is on quasi-induced exposure. The particular concern lies in how to validate the underlying assumptions of this technique. The definition of an accident rate is the ratio of accident frequency to accident exposure. Rates can be calculated for a road section, a set of intersections of the same type, a group of vehicles having some common features, an age cohort of drivers or some other combination of driver, vehicle, and/or environmental features. The numerator of the ratio can be straightforwardly expressed as accident frequency or the number of accidents. The denominator can be expressed in several ways, including direct measures such as VMT, NEV for an intersection, vehicle registration, and ADT, and indirect measures using induced or quasi-induced exposure. Comparison of different exposure methods employed under different circumstances will illustrate the strength and weakness of quasi-induced exposure. In order to investigate quasi-induced exposure related issues, it is first necessary to understand the inherent problems with traditional accident frequencies, accident rates, and measures of exposure.

There are three main topics to be explored here: accident data, accident rates and exposure, and induced exposure.

2.2 Accident data

Accident data are utilized by traffic engineers and researchers to plan, establish, and evaluate safety programs in general. The interpretation of such data may lead to a better understanding of operational problems, be of assistance in devising countermeasures for those problems, and, in many cases, allow the evaluation of the effectiveness of countermeasure programs. Certainly, not all safety-related decisions are based solely on accident data, but high reliability in both the quality and the quantity of accident data is important. The quality of accident data refers to the accuracy, timeliness, and completeness of the data used to address traffic problems. Quality of accident data was defined by O'Day (1993, pp. 1):

- Completeness of coverage (ascertainment)—the degree to which the data collection system contains all cases defined by the data collection threshold;
- Consistency of coverage—whether the degree of ascertainment varies by jurisdiction, time, personal characteristics, weather, or other factors;
- Missing data—in addition to the problem of missing cases, there may be missing data elements for cases that are reported;
- Consistency of interpretation—whether the report elements are reported in the same manner in different states or local jurisdictions, or by different reporting officers;
- The right data—another aspect of quality is having the right data elements;
- Appropriate level of detail—this depends on the variable and on the questions asked;
- Correct entry procedures—all of the above factors may be compromised or enhanced by the treatment of the data at the point of entering it into the computer; and
- Freedom from response error—when something was measured, was it measured correctly?

The literature pertaining to these components is presented in the next section and is divided into three main sub-sections: underreporting, inaccuracy of data, and inconsistency of reporting.

2.2.1 Underreporting or incomplete observations

Underreporting refers to accidents that should have been investigated but for which no data were collected or accidents that have been investigated but for which no data were recorded. The data for these accidents are simply not available. Incomplete observation refers to accidents that have been investigated but for which incomplete data were recorded. If there are incomplete data in the dataset, the measure of interest (e.g., exposure of older drivers at night) may be biased because the proportion in the incomplete or underreported data (which are unavailable) could potentially be different from those data which are available.

In 1971, Scott and Carroll reviewed the state of completeness of accident data in several states. They indicated that reporting completeness, which excluded missing whole accidents and incomplete accidents, was 48 percent in Washington D.C., 32 percent in Maryland, and 30 percent in Virginia. Chipman (1983) tested the accuracy of reports on fatal motor vehicle crashes by comparing vital statistics for Canadian provinces and territories with police-reported traffic fatalities. She discovered that counts of police-reported deaths were larger than the vital statistics source indicated—in one year as large as 7 percent difference for the entire country (434 deaths). She concluded that 7 percent underreporting or misclassification in such a statistic was unacceptable for many research applications.

With the concern for reporting completeness recognized, researchers attempted to determine the factors that affected the inclination to report an accident. Hauer and Hakkert (1988) discussed 14 studies of underreporting of accidents published from 1971 to 1985. It seemed evident that fatal accidents were reported more fully than serious injury accidents and that the coverage of the latter was, in turn, better than that for slight injuries. They found, on average, police records missed 20 percent of injuries that required hospitalization and up to half of the injuries that did not. In addition, the probability of reporting an injury sustained in a motor vehicle accident increased with the age of the injured person. For young children it was 20 to 30 percent, and for persons over 60 it was around 70% (Hautzinger et al., 1985). Another factor is the number of vehicles involved. In a report by Smith (1966), the reporting percentage for property-damage-only (PDO) in single-vehicle accidents was 57%; in multi-vehicle accidents the corresponding percentage was 96%. It is also believed that underreporting is more serious in large cities where the police are overloaded and not able to take the time required for full reporting (O'Day 1993). In Detroit, for example, police officials announced in the early 1970s that they would investigate accidents only when they were needed at the site. In this context, PDO accidents are likely to be severely under-reported.

O'Day (1993) argued that it was improper to simply assume that missing data were not biased with respect to the acquired data. It is probably better to assume the opposite. Typical examples are inclement weather and fatal accident reporting (O'Day 1993, pp. 3):

- During periods of inclement weather it is often not possible for the available police to attend accidents, so that higher underreporting is associated with weather.

- It is difficult for police officers to get all accident participants for interview when serious crashes happen in which some occupants have been transported to a hospital. The likelihood of injury increases with age, so that age information might be more likely to be missing for older persons.

Another important factor in underreporting is the minimum threshold in accident reporting (Willis 1983). A minimum threshold is a criterion established by law for accident reporting and processing which mandates the reporting of accident only when a specified amount of property damage is reached. It is used to prevent the flooding of accident files with data on minor (often insignificant) collisions. Nevertheless, McKnight (1981) argued that accident statistics must include all the accidents that occur. The reason is: "minimum thresholds of property damage and injury are used in all accident reporting systems to keep the system from being swamped with statistics on minor accidents that would be of **no** real benefit to the practitioner or scientist." For example, traffic engineers will not know the characteristics of involved drivers in the accidents below the minimum threshold. Thus, they not only are unable to identify whether certain driver cohort will be prone to get involved in accidents below the minimum threshold, but underestimate the total number of accidents for those drivers.

In summary, past research has indicated that underreporting and incomplete data are two common problems in reporting and/or recording accidents. Past research has also demonstrated that a variety of factors contribute to these problems: weather conditions, severity of accidents, minimum threshold limitations, availability of police, and size limitations of the accident reporting system. Therefore, it is important to assess completeness of accident data by taking the aforementioned factors into consideration before they are used for traffic safety related analysis. For the example of availability of police, if special attention is given to the accidents occurring in the Detroit area or a

bigger region containing this area, it is necessary to know what accidents are frequently underreported through either field investigations or questionnaire surveys.

2.2.2 Inaccuracy of data

The problem of underreporting is compounded by a variety of inaccuracies and errors that creep into the eventual computerized records of the accident. Inaccuracy of data refers to elements (of a crash) that have been recorded but with unreasonable, unknown, or biased values for various parameters in the accident reports. Identification of some accident inaccuracies can be achieved by comparing the police accident reports with some external sources (like hospital reports).

In the United States, traffic accidents are usually investigated by police officers who complete a standard form designed and promulgated by a state agency. Therefore, the quality of accident data depends at least on the performance of the police investigators and those reviewing the data and entering them into computers. McKnight (1981) pointed out that few police had enough training in accident reconstruction to determine what really happened and those that had the training often lacked the time necessary to gather and analyze the available data. Shiner et al. (1983) compared police records of 124 accidents with detailed information collected by multidisciplinary accident investigation teams (assumed to be correct). It was found that the police data were the most reliable for the following six variables: location, date, day of week, and numbers of drivers, passengers and vehicles in each accident. However, the police reports analyzed provided very little information regarding the presence of driver factors, human conditions, and vehicular and environmental/roadway factors and deficiencies. Hautzinger et al. (1985) also found inaccuracies in the reporting of accident type and

vehicle maneuver in 5% to 16% of the cases. Seemingly, the most conspicuous reason for the inaccuracy in reporting is the inconsistency and insufficiency of police training.

Inaccuracy of reporting accident location is another aspect of inaccurate accident data. Zegeer (1982) estimated the accuracy of accident location in Alabama, California, Michigan, and Illinois. Based on a questionnaire, Zegeer (1982, pp. 2) concluded that:

> A sizable portion (10 to 30 percent) of accidents cannot be located due to obvious locational coding errors or omission of location referencing information. Inaccuracy and incompleteness of the location descriptions problems arise because many police agencies are understaffed and must attend to other police duties, and thus accident report accuracy may not have a high priority.

An FHWA report (1997) recommended that all states be able to identify accident locations to the nearest 0.1mile in rural areas and 100 feet in urban area. The report also mentioned that police officers did not generally have portable computing devices available to guide and facilitate data collection, thus limiting quality and productivity. This is a problem of major importance, since the location is the principal mean for linkage to other spatial data (e.g., roadway inventory). Carreker et al. (2000) attempted to assess the accuracy of the existing accident location system of the Georgia DOT (GDOT) through a comparison of the locations of crashes based on the original crash reports and the locations chosen by GDOT using the route and mile number system. Three types of errors were identified that might have prevented the route and mile system from accurately locating the crash sites (Carreker et al. 2000, pp. 2):

- Type 1: route and mile numbers are incorrect, 37%;
- Type 2: location is in the correct vicinity but not the correct location–the mile number is incorrect, 23%; and
- Type 3: invalid route number, mile number or road name, 12%.

In short, typical inaccuracy in reporting accident data includes miscoding of driver gender, age, vehicle type, road conditions, severity of accident, and accident

location. This is due to the limitation of data collection tools, capability of the investigating police officers, insufficiency of police training, and/or data inputting errors.

2.2.3 Inconsistency of reporting

Inconsistency of reporting refers to the phenomenon where accident data are recorded, reported, or investigated in an inconsistent manner. For the same or very similar accident facts, different state, county, or police officers have different perceptions and definitions and, consequently, the same facts might be reflected in different ways in accident reports. Particularly, inconsistency of reporting becomes crucial when the attempt is made to combine and compare accident datasets from different data sources (typically, different jurisdictions). It is not much value to use a database containing inconsistently reported accident data, since the results might be biased, conflicting or even erroneous.

O'Day (1993) pointed out that inconsistency existed in categorizing vehicle type between different states or local jurisdictions. For example, one state might group pickup trucks and small vans in a single category in accident data while another might use separate categories for such vehicles. In the same report, O'Day also mentioned the inconsistent identification of vehicle defects as the cause of accidents. One state might utilize specific variables in the report to identify defective vehicles, such as tires, brakes, steering, or lights, while another state records such information only in the narrative of the report.

Another inconsistency problem is in reporting injury severity (O'Day 1993). The majority of states record injury on a five-point scale often referred to as the KABC0 scale: K is "person with fatal injury," A is "person with incapacitating injury," B is

"person with non-incapacitating evident injury," C is "person with possible injury," and 0 is "no injury." Statistics from 23 states (1988-1990) showed that California reported only 4.9% ("A") injuries while Illinois reported up to 23.8%. It does not seem likely that these apparent differences in accident severity are real. Obviously, inconsistency in accident severity definition between states is a contributing factor to this variation. Therefore, the inconsistency of definitions of accident severity limits the development of meaningful studies across states and prevents integration of accident data from various states.

In summary, the literature review contains three examples of inconsistent reporting problems existing in state accident reports—different definitions of defective vehicles, injury severity, and accident location. It illustrates the point that fusion of accident data from different states will be a problem and, thus, developing traffic safety studies with the use of such data should be done with caution. Furthermore, the review of literature also revealed that no past research explored the issue of combining different accident data from different police agencies within the same state. A later section will address this concern.

2.2.4 Summary

To summarize the discussion of accident data quality, the issues are not whether accident data are useful for evaluating highway safety problem nor is it argued that the use of accident data in traffic safety analysis should be abandoned. Rather, the concerns are with identifying those aspects of accident data that should be examined during the process of manipulating and using accident data. The goal is to improve their reliability and quality in order to develop more dependable and practical solutions to real-world traffic problems.

2.3 Accident rates and exposure

The advantage of using accident rates and exposure in safety analysis versus accident frequencies lies in the ability to explicitly consider the driving exposure where accident data were collected. In this sense, high-quality accident data are necessary for a safety analysis, but not necessarily sufficient.

The definition of an accident rate is the ratio of accident frequency to some measure of accident exposure. Accident frequency generally refers to the number of accidents for a driver group, vehicle group, or driver-vehicle combination. Accident exposure refers to the measurement of total driving hazards that the particular driver-vehicle combination confronts when it is on the road. Accident frequency can be used for relatively simple safety analysis, often before-after analysis, to examine the effectiveness of certain safety countermeasures. As pointed out by Lyles et al. (1993), although accident frequency is useful in certain situations, the problem with using it is the lack of consideration of opportunity for accidents to occur, for instance, it is seldom recognized that higher traffic volume alone may lead to a difference in the frequency of accidents. A common problem associated with many before-after analyses is that it assumes that the traffic countermeasure is the only contributing factor of the change in the magnitude of accidents, or simply traffic and road conditions remain the same.

Accident rates, on the other hand, include consideration of a measure of the opportunity for accidents to occur. Comparison of accident rates can assist road safety researchers in developing safety countermeasures in ways that comparison of absolute frequencies of accidents can not. Inherent in the concept of exposure is the idea of using exposure data to determine accident rates which indicate the relative degree of risk or

danger of various road traffic situations (situations broadly include all relevant vehicle, person, and environmental characteristics). When the discussion of accident frequency is combined with accident exposure, it has more flexibility and capability for use in evaluating an operational intervention, identifying potential traffic problems, and making traffic safety policy. Seemingly, the introduction of accident exposure in safety analysis makes a fundamental difference. The distinction is recognized in a discussion by Jovanis et al. (1991, pp. 2):

> Not considering the amount of exposure means that consideration is only given to the characteristics of the accidents that have occurred, not the road and traffic conditions of driving during which accidents have not occurred.

That is, safety analysis without the inclusion of accident exposure can be incomplete. The argument is that the validity of an accident rate as a basis of comparison is based on the ability to accurately determine the measure of exposure. In a study by Chapman (1993), the exposure to risk of accidents was characterized in two ways: 1) exposure and accident rates for a vehicle or road user (referred to as group exposure) operating on the system, and 2) for particular sites or fixed objects (referred to as site exposure) existing in the system. Group exposure is defined as the number of accident opportunities a particular driver experiences as he/she drives around the road network (Hodge 1985). Site exposure is defined as the amount of opportunity for accidents occurring at a particular site or group of sites (Hodge 1985).

For group exposure, vehicle miles traveled is the most commonly used measure. In addition, the duration of travel, number of discrete trips, and number of crossings have also been used as direct measurements of exposure. For site exposure, direct traffic counts, number of traffic conflicts, sum of entering flows at intersections, cross-product of conflicting flows at intersections or the square root of the cross-product of conflicting

flows at intersections are methods/measures recommended by Chapman (1973). Most of the methods mentioned above (direct measurement) are hampered by data availability and there is a greater difficulty in calculating sub-group exposure of road users or site types (e.g., old drivers, small cars, collector streets).

The review of the literature on exposure starts with a definition, followed by discussion of group exposure, and ends with discussion of some issues regarding accident exposure and accident rates.

2.3.1 Definition of exposure

In 1942, De Silva mentioned the concept of exposure and defined it as "the number and relative danger of the hazards he (the driver) encounters." He noted that general exposure to different hazards varies in terms of where, when, and how a person drove. Dunlap (1953) stated that exposure was a measure of "the frequency of the existence of a situation which may or may not involve an accident." This definition showed that the driving exposure included all the situations no matter whether there were accidents involved in the driving, which was different from De Silva. Mathewson and Brenner (1957) recommended a general definition of exposure as a "unit of risk in motor vehicle accident rates." A similar definition was proposed by Mathewson and Jacobs (1961) who defined exposure as "the frequency of occurrence of risk situations and circumstances associated with risk situations." Goeller (1968) called exposure over a given driving distance "the number of times that danger occurs." Haight (1971) noted that "exposure to accidents" evolved as a concept by analogy to "exposure to disease," and indicated continuing difficulties in giving the concept a precise meaning. Carroll (1971) offered yet another definition "driving exposure is the frequency of traffic events which

create a risk of accidents," which is very similar to Mathewson and Brenner (1957). Briefly, these definitions of exposure are similar in nature and generally consider exposure as a measurement of driving hazards or driving risk.

In 1970, Klein and Waller considered exposure as the "population at risk (in terms of passenger or vehicle miles)" and used as a denominator in the calculation of an accident or injury rate. In a report from Operations Research, Inc. (1971), exposure was viewed as "a systematic process affecting the crash system that is essentially a function of the continual interaction of driving behavior with the ever-changing environment." The authors of this report regarded exposure as "obviously something more than the gross vehicle mileage for all drivers under all driving conditions, the usual proxy measure." The elements of exposure should include: characteristics of drivers and vehicles, characteristics of the road system and intensity of system use and environmental conditions (weather, light conditions). These definitions illustrate that exposure to accident risk is a combined effect of driver, road conditions and environmental conditions in addition to passenger or vehicle miles.

Chapman (1973) reviewed the past definitions and use of exposure measures and provided an important reference for this field. His definition is "exposure is the number of opportunities for accidents of a certain type in a given time in a given area." This exposure is defined in a more strict sense—exposure is an indicator of accident opportunities confined to a specific timeframe and location.

2.3.2 Group exposure

Group exposure is defined as the number of accident opportunities that a road user experiences on the road network. Generally, the most commonly used measure for

drivers is vehicle miles traveled, while for pedestrians, duration of travel and number of crossing are more frequent. The focus here is vehicle miles traveled.

VMT is an expression of how much traffic uses the road and how far this traffic travels. VMT is the most common procedure for approximating exposure of the road user (Carr, 1969). Carroll (1973) considered the distance traveled as the most indicative measure of exposure. However, use of VMT implicitly assumes constant risk at all sites, under all environmental conditions, and for each mile traveled.

In one of many attempts to estimate VMT, a study done at University of California (1975) started with short-term traffic count surveys at numerous locations, and continuous counts at a few sites. Then the continuous count data were adjusted for seasonal influence, and daily or weekly traffic fluctuation. The VMT was estimated by multiplying volume counts by the length of the road segments for which the characteristics were assumed to be the same. For the estimation of VMT for a specific roadway segment and a particular vehicle type, the basic formula is expressed as:

$$VMT_t = t \times \left(MP_{i+1} - MP_i\right) \times \left[\frac{\left(V(1)_i + V(1)_{i+1}\right)}{2}\right] \times 365$$

where:

t – the proportion of particular vehicle type;

$V(1)_{i+1}$ – adjusted 24-hour volume counts in "ahead" leg of count location;

$V(1)_i$ – adjusted 24-hour volume counts in "back" leg of count location;

$MP_{i+1} - MP_i$ – mileage between locations; and

$\left[\left(V(2)_i + V(1)_{i+1}\right)/2\right]$ – average daily traffic.

Summing over all roadway segments yields a system-wide VMT estimate for the particular vehicle type. This report did not provide information on the magnitude of errors associated with the calibration process for seasonal, weekly, or daily fluctuation or

the number of short-term counts used to estimate VMT. Thus, accuracy of the computed VMT was unknown.

Transportation engineering agencies also show interest in measuring VMT of specific road users at regional or national levels and on an ongoing basis. Ferlis et al. (1981) developed a procedure to estimate regional VMT through the use of sampling techniques. They elaborated that an efficient way of estimating regional VMT with a sample of traffic counts was to estimate the average volume in each sample stratum, multiply the average volume by the total mileage in the same stratum, and average all the stratum-specific VMT estimates to produce an estimate for the region. The estimate is shown as:

$$\overline{VMT} = \sum_{h}^{H} VMT_h$$

$$\overline{VMT}_h = M_h \times \overline{VOL_h}$$

$$\overline{VOL_h} = (1/N_h) \times \sum_{i}^{N_h} VOL_{h_i}$$

where:

H – the number of sample strata;
M_h – the mileage of stratum h;
N_h – the number of volume counts made; and
$\overline{VMT}$ – the estimated average regional VMT during the time of interest;
$\overline{VOL_h}$ – the estimated average volume in sample stratum h;
VOL_{h_i} – the volume measured on count i in sample stratum h.
$\overline{VMT_h}$ – the estimated average VMT in sample stratum h during the time of interest;

The advantage of this model lies in the straightforward computational steps and the disadvantage is that data errors emerging in a previous step will accumulate and propagate to a next step.

In an effort to forecast personal daily VMT, Kuzmyak (1981) developed a model incorporating individual and household composition factors that most directly affected the individual's travel decision making, including characteristics of the individual, the characteristics of the individual's household, travel-related considerations, purpose of travel, travel destination, and residence location. The strength of Kuzmyak's model was its ability to forecast personal VMT by means of existing travel data sets (if available) while the weakness was that it didn't consider the importance of fuel price and availability, transportation level of service, temporal consistency, and the effects of competing modes. On the other hand, this model is unable to perform fine-grained analysis of behavior by trip purpose or land use, since the model was derived and calibrated based on 1977 National Personal Travel Survey data which did not report sample location and trip details. These factors in combination substantially weaken the applicability of the model.

Although VMT as a measure of exposure has been widely used in traffic safety analysis, it has also aroused disagreement among traffic researchers. Basically, the underlying assumptions of VMT-based methods have been challenged.

The first assumption is that all driving involves the same exposure to accident hazards. This assumption was challenged by Steward (1960), who stated that "the concept of exposure has a more narrow meaning, one which takes account of probable facts in one's present, and/or immediate past environment...an individual has been exposed to a disease after he has direct contact with some carriers or has had opportunity for contact." That is, not all driving (vehicle miles) is subject to similar driving hazards under certain circumstances, such as vehicles traveling in a platoon at an identical speed.

In reality, the leading vehicles might experience more driving hazards or differently than the following vehicles.

The next assumption is that exposure to accident hazards is always proportional to miles driven. This assumption is really a matter of which level of accident data are considered. Generally speaking, when the use of VMT is an estimate of exposure at the system level, the assumption is valid. Data at this level are more aggregated and the exposure estimation is relatively gross. When exposure needs to be estimated in a disaggregated manner by variables of interest (e.g., roadway type, driver age), the assumption becomes less appropriate. Exposure to accident hazards in these situations becomes a function of traffic and personal behavior characteristics in addition to miles driven. For instance, the same driver might confront less accident hazards on a freeway than on a local street with identical length.

Another assumption is that the degree to which exposure is associated with miles driven is the same for all drivers. Substantial differences exist between different drivers in terms of driving knowledge and experience and thus drivers might respond differently to the same type of driving hazard. Given this setting, for the same miles driven, experienced drivers might "feel" being exposed to fewer hazards than inexperienced drivers do, even though the hazards themselves are the same. The point is that the objectively similar situations are not equally hazardous for different drivers.

Finally, the last assumption is that the traveling speed for groups under scrutiny is inherently assumed to be equal. In order to illustrate the point, imagine a potential at-fault driver waiting somewhere to encounter or interact with two innocent drivers, one traveling for 10 miles at 30 MPH and the other for 20 miles at 60 MPH. The at-fault

driver will have two equal opportunities (time duration) to interact with an innocent driver. The probability of each of these innocent vehicles being impacted by an at-fault driver would be the same. Using the traditional VMT method would result in the vehicle that is traveling twice as far having twice the probability of being in an accident.

Another operational-oriented criticism is leveled by Lyles et al. (1991). While VMT appears to be widely used and acceptable on a system-wide basis, it becomes "virtually impossible" to use VMT to measure the exposure of different types of motorists in different types of vehicles on different roadways (and so forth). This is related to the data availability issue of VMT. In order to compute VMT disaggregated by the variables of interest (e.g., road type, driver age, time period), it is necessary to know the traffic volume and average of mileage traveled under a specific circumstance (e.g., VMT of young drivers on local streets on Friday night). Although, theoretically, this is possible, practically there is very limited (or even no) availability of such specific information.

The point here is to illustrate that although VMT is the most frequently used exposure measurement, there are theoretical and practical difficulties stemming from problematic assumptions and data unavailability. The quality of predicted or calculated VMT data, especially at a finely-disaggregated level, has been seriously questioned. Consequently, there is a need for an alternative approach to estimate exposure. Seemingly, quasi-induced exposure poses a promising solution to this dilemma. Not only are the problematic assumptions of VMT avoided, but the *relative* exposure is estimated solely from the more readily available accident data. In contrast to VMT, its desirable features include fine disaggregation of exposure by variables of interest, a simple

calculation process, and availability of data. The details of quasi-induced exposure will be explored in a later section.

2.3.3 Related issues

There is also a controversy regarding the relationship between accident frequency and accident exposure, as advanced by Hauer (1995)—accident frequency is not linearly proportionate to accident exposure and thus the accident rate is not constant. His argument is based on the discussion of a nonlinearity relationship between accident frequency and exposure for a road user or site group. Figure 2.1 is quoted from his study. Note that in figure 2.1, the author states that "the shape of the function is immaterial, only its essence." The purpose of this figure is to show how the average number of accidents in a specified period of time would be changing if exposure changed, while all other conditions affecting accident occurrence remained fixed.

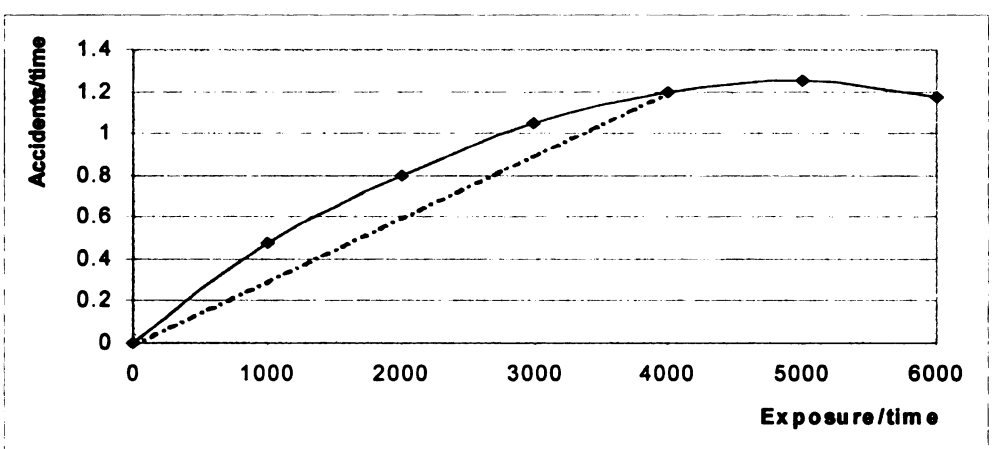

Figure 2.1. A nonlinearity relationship between accident frequency and exposure

Based on figure 2.1, the accident rate is the slope of the line joining the origin and a point in the curve. Along the curve, the accident rate varies with the change in exposure. In practice, traffic engineers and researchers expect that the number of accidents for a certain road user to increase proportionally with VMT. For instance, one

truck accident per 10,000 vehicle miles is reasonably used to predict two truck accidents per 20,000 vehicle miles. This type of prediction is called the "linearity conjecture." This assumption contradicts the tendency indicated in the curve. Hauer concluded that simple use of the accident rate would cause an over-representation problem with increasing exposure.

Whether the linearity conjecture can be adopted depends on whether it causes a significant effect on the results. Although it is well argued by Hauer (1995) that accident frequency is not proportional to the accident exposure in the curve, in some portion of the curve, the increase in accident frequency can still be approximately considered as proportional to accident exposure within tolerable errors. For instance, for exposure ranging from 0 to 4,000 (figure 2.1), the linearity conjecture should not be a major problem, where the linear relationship between exposure and accident frequency can be approximated. Outside the range (>4,000), the accident rate seems to be overestimated. Considering this worse scenario, in typical engineering practice an overestimated accident rate will lead to conservative countermeasures with regard to certain traffic problems and certainly it helps to promote safety.

In a very similar study by Janke (1990), the author challenged the underlying assumptions that there was a linear relationship between accidents per driver and driver mileage, which was commonly used as a basis for developing traffic safety-related studies. The author argued that for two groups of drivers who are equally competent and prudent, the same length of highway miles offers different amounts of exposure to risk. For example (Janke, 1990), high-mileage drivers typically accumulate mileage on roadways with relatively high speed limits, which generally have much lower accident

rates (per mile) than other types of roadways. Low-mileage drivers typically make short trips on local surface roads, which are comparatively more congested and thus offer more opportunity for accidents. This illustrates the point that different roadways offer different accident risk for the same mile length and consequently the relationship between accidents per vehicle mile and driver mileage is not necessarily linear.

Since strict linear proportionality of accidents to mileage does not seem to hold, Janke (1990) suggested that the induced exposure approach gives a more balanced measure of risk than does the method of using accidents per mile. It also appears to circumvent effects of the complication of the relationship between the driver mileage and driving competence in the interpretation of accident-per-mile data.

2.4 Induced exposure

In view of the difficulties of estimating exposure by means of traditional methods (e.g., VMT), researchers began to look for an alternative to estimate exposure and turned to induced exposure. The major advantage of induced exposure is that it overcomes the problem of data availability confronted by other exposure methods, as exposure is estimated directly from the accident data themselves. The original idea of induced exposure is attributable to Thorpe (1964), who suggested that by making a set of assumptions the exposure for a specific driver-vehicle combination could be estimated according to the numbers of single- and multi-vehicle accidents for that combination. The basic assumption behind this idea was that the exposure for certain classes of drivers, vehicle types, and driving environments, was proportional to the number of times that the analysis category was an "innocent victim" in collision accidents. That is, innocent drivers involved in the accidents are randomly impacted by accident initiators, and thus

innocent drivers in the accidents are a random sample of the driving population on the road at the time and place of the accident. The following section is about the development of induced exposure.

2.4.1 Definition of induced exposure

Although the concept of "induced exposure" was originally proposed by Thorpe (1964), Haight (1973) was the first author who explicitly defined the concept of induced exposure through a summary of previous work. Haight offered two definitions of induced exposure from different perspectives: "narrow-sense" and "broad-sense." The narrow-sense definition is (Haight 1973, pp. 2):

> Exposure referred exclusively to exposure to collision with other vehicles, and, consequently, in which all other types of mishap were conceived as taking place not because of exposure, but for other reasons.

More broadly, the definition of induced exposure is (Haight 1973, pp. 2):

> Exposure and proneness as mutually exclusive and exhaustive concepts (one internal and one external), which suffice to account for (expected) accident experiences.

In the narrow-sense, induced exposure is defined as a measure of driving hazards in the traffic or road environments, while in the broad-sense, induced exposure further considers the interaction between the driver and the external environments. The details of these two types of exposure are explored in the next sub-section.

2.4.1.1 Narrow definition

Thorpe's theory (1964) was categorized as the narrow definition. His theory to estimate relative exposure for a driver-vehicle combination is based on five assumptions (Thorpe 1964, pp. 1):

1. Single vehicle accidents are caused entirely by attributes of the driver-vehicle combination concerned.

2. Collision accidents are caused by the first two vehicles to hit.
3. In each collision accident there will be a "responsible" and a "not responsible" driver-vehicle combination.
4. The relative likelihood of a driver-vehicle combination being the "responsible" combination in a collision accident will be the same as the relative likelihood of that combination being involved in a single-vehicle accident.
5. The likelihood of any particular driver-vehicle combination being innocently involved in a collision accident will be the likelihood of meeting that combination anywhere on the road.

Assumption 5 has been used as the operative definition of induced exposure. The exposure, E, to an accident for a given driver/vehicle combination i is computed by the formula (Thorpe 1964, pp. 2):

$$E_i = 2T_i - S_i$$

where:

T_i – the percentage of multi-vehicle accidents for driver/vehicle combination i;

S_i – the percentage of single-vehicle accidents for the same driver/vehicle combination.

This formula is attractive to traffic researchers due to its simplistic nature in estimating exposure. In general, induced exposure has some obvious advantages as to cost, time, and convenience, since accident records are normally available.

Thorpe's induced exposure theory was not mature and drew considerable criticism either because of the underlying assumptions or the exposure formula itself. Haight (1971) pointed out that it was possible for the exposure formula to produce negative values. In Thorpe's original work (1964), he seemed to realize this problem and dismissed it as one which resulted from sampling error with small numbers, since the exposure was sensitive to errors in T_i and S_i. However, the exposure of a particular driver-vehicle combination is correlated with factors which are in turn highly correlated with the propensity for single-vehicle accidents, the exposure will be negative regardless

of the sample size. For example, suppose that young inebriated drivers drive only at night when traffic is light and therefore a high proportion of their accidents are single-vehicle. Then the accidents of young persons in two-vehicle crashes (T_i) would be small, but their involvement in single-vehicle (S_i) accidents would be large, consequently resulting in a negative exposure. The negative value of the formula results from the assumption that responsible driver-vehicle combination in single-vehicle accidents follows the same distribution as that in two-vehicle accidents.

The same issue has been explored and criticized by several traffic researchers (Carr 1969, Hall 1970, Carlson 1970, Joksch 1973, Brown 1982, Stamatiadis 1997). Most of the researchers later turn to quasi-induced exposure, which will be covered in a later section. For example, Brown (1982) argued that Thorpe's fourth assumption was very questionable, because a variety of personal characteristics (e.g., extraversion, aggression, anxiety) would tend to produce greater self-induced risk exposure when other drivers were present than when they were not. Moreover, control skill failures of drivers were largely responsible for single-vehicle accidents, whereas both control skills and roadcraft (capability of perceiving road hazards during driving) might be causally implicated in multi-vehicle accidents. Therefore, the probability of being involved in different types of accident might vary for a particular driver-vehicle combination.

More recently, Stamatiadis et al. (1997) made use of Kentucky data to suggest that such an assumption might be erroneous. Relative involvement classified by driver age and vehicle type was significantly different for single-vehicle and multi-vehicle accidents. Their study showed that as drivers aged, they became progressively more conspicuous in multi-vehicle accidents. The ratio of single-vehicle to multi-vehicle

accident percentages decreased from 1.21 (driver's age less than 25), to 0.88 (age between 45 and 54), to 0.31 (driver's age greater than 75); as vehicle size increased, the same ratio increased from 0.97 (automobile), to 1.29 (straight truck), to 1.83 (combination truck). Obviously, large trucks were overrepresented in single-vehicle accidents. This illustrates the point that accident involvement rates in single-vehicle and multi-vehicle accidents for the same vehicle-driver combination should not be treated identically.

Another shortcoming of induced exposure stems from assumption 3—only one responsible and one not-responsible driver are involved in any accident. Based on Thorpe's theory, responsibility for accident causation can be identified and thus assigned to one of the drivers in a two-vehicle accident. Certainly, responsibility is not easily and always identifiable for all accidents. If those accident data where responsibility is not clear or shared are included in the analysis, the estimated exposure is potentially biased. For example, two aggressive young drivers chase each other in the freeway and cause a "side-swipe same" type of accident. It would be reasonable to assign accident fault to both drivers rather than one of them. Given this, to the extent possible, accidents as such are not used in induced exposure and should be eliminated from further considerations.

2.4.1.2 Broad definition

The broad definition was articulated in Koornstra's theory (1973). His model was not based directly on concepts of guilt and innocence in two driver/vehicle combinations, but rather on a separation of factors into two general types: external factors (called exposure), and internal factors (called proneness). The model was expressed as (Haight 1973):

$$D(x,y) = \phi(f(x), g(x), f(y), g(y))$$

where:

$D(x,y)$ – the number of accidents between driver-vehicle combination x and y;
$f(x)$ and $f(y)$ – the proneness of driver-vehicle combinations x and y;
$g(x)$ and $g(y)$ – the exposure of driver-vehicle combinations x and y; and
ϕ – an interaction function.

In this model, a single-vehicle accident is treated as a two-vehicle accident where the second vehicle belongs to a fictitious "dummy" category. Due to the general nature of Koornastra's model, the model is somewhat difficult to follow and much more detailed identification of the problems is needed. The potential problems are described below (Mengert 1982, pp. 2):

1. Imperfect mixing problem, which relates to the fact that two user groups may not have their exposure distributed identically over time or roadway types.
2. Accident fault assigning problem, which models accident fault as pertaining to one party or the other, or to neither party but not to both.
3. The model looks upon each accident situation as symmetric in the user groups in that the same potential accident situation would be as likely to occur if the roles of the user groups are interchanged. It is clearly incorrect for certain combination of vehicle groups (e.g., vehicles of greatly different size).
4. The same proneness distribution for a user group applies in all situations. This is similar to Thorpe's problematic assumption 4.

2.4.2 Quasi-induced exposure

Acknowledging Thorpe's (1964) problematic assumptions and theoretical difficulties, other traffic researchers (Carr 1969, Hall 1970, Carlson 1970, Joksch 1973, Cerrelli 1973) modified Thorpe's original work and supplemented it with a systematic responsibility-assigning scheme. Haight (1971) called this technique "quasi-induced exposure" and defined it as "an induced exposure method to measure the relative exposure of driver/vehicle combination to the risk of driving hazard, with a well-defined responsibility assigning system." Based on this definition, it seems that the difference

between induced exposure (generally referred to as Thorpe's and Koornastra's methods) and quasi-induced exposure lies in the responsibility assignment issue. The former uses single-vehicle accident experience to establish the "responsible" involvement of an attribute in two-vehicle accident, while the latter assigns responsibility through more reliable sources such as actual police reports or citations (Carr 1969 and Cerrelli 1973).

Other than the responsibility assignment issue, quasi-induced exposure theory is developed based on two fundamentally different underlying assumptions compared to induced exposure (Lyles 1994, pp. 2):

1. In two-vehicle accidents there is an at-fault and a not-at-fault driver.
2. Not-at-fault drivers in the two-vehicle accidents are a random sample of motorists and vehicles on the road during the study time.

For assumption one, quasi-induced exposure requires the utilization of only two-vehicle accident data with one at-fault or responsible driver and one not-at-fault or non-responsible driver. Accident responsibility for causation is assigned to one of the drivers in a two-vehicle accident based on data from police accident reports (e.g., Carr 1969, Hall 1970, and Carlson 1970). The driver-vehicle combination that is responsible for the accident is defined as Driver-1 or D1. Consistent with the above, the driver-vehicle combination who is not-at-fault is defined as Driver-2 or D2. Based on the terms defined, assumption two can be rephrased: D2s are randomly selected by D1s from all vehicles existing on the system and thus D2s constitute a random sample of driver-vehicle combinations and, inductively, a measure of exposure (Lyles 1994).

According to the assumptions, a D1-D2 matrix is constructed to calculate an involvement ratio (IR) for a particular driver-vehicle combination. Table 2.1 is a matrix showing sex of driver for both at-fault (D1s) and not-at-fault (D2s). Each row in a D1-D2

matrix is the responsible driver and each column is the non-responsible driver. The summation over each row is the total number of D1s and the summation over each column is the total number of D2s.

Table 2.1 D1-D2 matrix used in quasi-induced exposure

D1-D2 matrix		Driver-2 (not-at-fault, D2)		D1-total
		male	female	
Driver-1 (at-fault, D1)	male	A_{11}	A_{12}	$\sum_{j=1}^{2} A_{1j}$
	female	A_{21}	A_{22}	$\sum_{j=1}^{2} A_{2j}$
D2-total		$\sum_{i=1}^{2} A_{i1}$	$\sum_{i=1}^{2} A_{i2}$	$A = \sum_{i=1}^{2}\sum_{j=1}^{2} A_{ij}$

Thus, the Involvement Ratio (IR) of a driver-vehicle combination is computed as the marginal proportions of D1s divided by D2s for the particular combination. This ratio of the D1 characteristic to the D2 characteristic (e.g., sex of driver) provides a measure of relative involvement of that characteristic in accident causation. Based on the IR value, one is able to determine if a certain driver-vehicle combination causes disproportionately more (IR>1) or less (IR<1) accidents. If IR is equal to 1, the driver-vehicle combination causes accidents proportionately to their presence on the road.

As the quasi-induced exposure definition and theory indicate, it is fundamentally different from other induced exposure methods, e.g., Thorpe or Koornastra. Further comparison reveals that differentiations are not only in terms of the underlying assumptions and responsibility-assigning scheme, but in the accident data and variables used. Details are shown in table 2.2.

Table 2.2. Comparison of induced exposure and quasi-induced exposure

issue	induced exposure	quasi-induced exposure
responsibility assignment scheme	one-vehicle accident	police accident report, etc.
one-vehicle accident data used?	yes	no
two-vehicle accident data used?	yes	yes, only with a D1 and a D2
three or more accident data used?	yes, first two vehicles	no
measurement of exposure	$2T_i - S_i$*	D2

* T_i is the percentage of multiple-vehicle accidents for drivers i; S_i is the percentage of single-vehicle accident for the same of driver/vehicle combination.

With the general theory of quasi-induced exposure introduced, the following section is an exploration of the research on and using quasi-induced exposure. It is divided into three main topics: validation of underlying assumptions, responsibility assignment, and applications.

2.4.2.1 Validation of assumptions

The validation of underlying assumptions of quasi-induced exposure method is an essential step in determining whether it is an appropriate exposure measurement or a useful analysis tool. Only when the assumptions are satisfied by accident data sets, can quasi-induced exposure be used with confidence, otherwise it should be used with caution or not at all. As stated earlier, quasi-induced exposure's first assumption in terms of accident data type, is relatively easily met.

Assumption two requires that the innocent driver-vehicle combinations in the accident be a random sample of the driving population on the road at the time of the

crash. Surprisingly, the literature review indicates that few traffic researchers actually validate this assumption before using quasi-induced exposure to calculate the relative accident involvement ratio. Most of the researchers simply *make* the assumption and develop the analysis. For example, in examining Ontario accident data for different driver age groups, driver gender, driving experience and alcohol use, Carr (1969) simply made the assumptions and compared the results with those from Thorpe's method and concluded that quasi-induced exposure had significant advantages over Thorpe's, without specific investigations of whether the assumed postulations were valid. In a similar study using National Accident Summary Files as a data source, Cerrelli (1973) employed quasi-induced exposure to obtain a numerical estimate of the liability and driving hazard associated with each class of driver. Again, in this study there was no validation of assumptions. The same problem exists in more recent studies as well (DeYoung 1997, Stamatiadis 1998, Aldridge et al. 1999, Kirk et al. 2001a, Chandraratna et al. 2003, and Hing et al. 2003).

Lighthizer (1989) seemed to be the first author who explicitly dealt with validating the underlying assumptions of quasi-induced exposure. He proposed two techniques. The first technique was direct observation of the values of key variables in the field. The variables included fleet-gender combination (e.g., male driving auto or station wagon, male driving pickup or van) and fleet mix (auto and station wagon, pickup and van, semi-truck and truck and utility vehicles). The author employed Michigan DOT accident data (1982-1988) and on-site collected data in southwest Michigan. He compared the distributions of driver-vehicle combination from field observations with those of non-responsible drivers derived from Michigan DOT accident data for three

levels of analysis: overall, by county, and by day. The results indicated overwhelmingly good agreement for driver gender at the overall, county, and the daily levels, but not as positive for the fleet mix. Lighthizer argued that problems were due to the inadequate accident data for the road segment of interest. Actually, there is a fundamental problem associated with this technique, which implicitly assumes that the exposure "truth" collected from field observations are identical to the exposure at the time when the accident data were collected or the accident occurred. This assumption is questionable. Exposure computed by quasi-induced exposure is specific to driving population at the location where and at the time when accidents occurred; exposure "truth" collected by field observations is specific to the driving population at the location where and at the time when traffic is observed. From the perspective of location, field observation sites might not be representative of those where the accident occurred, even though both are on the same route; from the perspective of time, the time when the accident occurred is before the field observation and thus the characteristics of driving population might have changed. Lighthizer ignored these facts and did not take appropriate actions to adjust exposure "truth" and make it comparable with accident exposure. He aggregated six years (1982-1987) of two-vehicle accidents occurring on I-94 in the six counties in Michigan (Berrien, Calhoun, Jackson, Kalamazoo, Van Buren, and Washtenaw) and compared the data directly to the field data observed in the summer of 1998. The author did explore the issue of seasonal and annual fluctuations among the accident data by means of analysis of variance (ANOVA) procedure, but the concern of whether the field data and the accident data were comparable was not addressed. Factors such as a

differentiated growth rate of traffic volume and change of driving population should be taken into account.

The second technique was complementary sets analysis. That is, the distribution of non-responsible driver-vehicle combinations involved in accidents caused by driver-vehicle combinations with certain characteristics are compared with the distribution of non-responsible driver-vehicle combinations of accidents caused by the complement set of driver-vehicle combinations. If they are the same, D2s represent a random sample of the driving population; if not, D2s are not a random sample and thus the assumption is not satisfied. The author utilized Michigan DOT accident data in the analysis of driver-vehicle's distribution disaggregated by gender, roadway type, and fleet mix. Based on the results, this technique produced very encouraging results which were consistent with the assumption for several variables examined. For example, in the following D1-D2 matrix pickups and standard automobiles were examined (pp. 152).

Table 2.3 Actual distributions of D1s and D2s for pickups and standard auto on I-94

D1-D2 matrix		D2		D2-total
		auto	pickup	
D1	auto	983 (92.0)	78 (8.0)	971 (89.3)
	pickup	105 (90.5)	11 (9.5)	116 (10.7)
D1-total		998 (91.8)	89 (8.2)	1087

In table 2.3, the row distributions (as well as the marginal total) appear to be similar—within 1.5 percentage points. This suggests the similarity of the D2 distributions, which supports the assumption that D2 vehicles constitute a random sample of the vehicles on the road.

Most recently, Kirk and Stamatiadis (2001b) utilized a trip-diary approach to measure travel exposure (VMT) and compared these estimates to those derived through

quasi-induced exposure within the urban boundaries of Fayette County, KY. Since trip diaries provided data for the specific trips taken such as time of day, day of week, trip purpose and roadways that the individual drivers selected, it allowed for the comparison to be developed in a disaggregated manner. The exposure estimates from the trip diary and the quasi-induced exposure were compared for three age groups (18-34, 35-64 and 64+) disaggregated by roadway class, time of day, and day of week. The results showed that age group (35-64) had the smallest differences (less than 4 percentage points) between the two exposure estimates although there were a few instances where the differences were large (as many as 15 percentage points). The authors identified that for age group (35-64) the large sample size both in participants and in number of routes contributed to the similarity with quasi-induced exposure. However, for the age group (64+), VMT exposure was consistently 2 to 3 times higher than the exposure from the quasi-induced approach. With the exposure data in more disaggregated conditions (age, roadway class, or time of day), the differences between the two exposure estimates became conspicuous and significant. Although the validation was not successful at any of the disaggregation levels, the study nonetheless produced some interesting results: 1) there were cases where exposure estimated by each method produced similar results; and 2) differences between two exposure estimates for various age groups were significant. An essential point is made in this study that the underlying assumptions of quasi-induced exposure can be tested through the comparison between the VMT calculated from eternally available data (a trip diary, in this case) with the exposure given by quasi-induced exposure.

In 2001, Golias et al. claimed that quasi-induced exposure method "had been tested in several occasions and its statistical validity had been verified," citing a study by Hodge et al. (1985). Unfortunately, the referred paper described extensively the use of group exposure measures in the assessment of risks incurred by people within particular population groupings and reviewed briefly the use of induced exposure measurements. There was no statistical validation whatsoever on the underlying assumptions of the quasi-induced exposure technique.

In summary, although underlying assumptions of quasi-induced exposure have been known for decades, very few traffic researchers actually devote efforts to validate them. This is important because the use of quasi-induced exposure technique hinges on the validity of underlying assumptions. Although several authors have attempted to validate them with different methodologies, some are not theoretically precise or incomplete (e.g., Lighthizer 1989). This provides motivation for this research effort to explore/develop different techniques to test the validity of quasi-induced exposure in a more comprehensive and convincing manner.

2.4.2.2 Responsibility assignment

The first assumption of quasi-induced exposure requires a specific accident data format—two-vehicle accidents with one responsible driver and one innocent driver. Therefore, it requires a responsibility-assigning scheme to determine which driver in a two-vehicle accident is responsible for the accident.

Suspecting the validity of Thorpe's fourth assumption, traffic researchers begin to assign responsibility for accident causation based on police investigators' judgment, since assigning and recording responsibility for accident causation is a common practice for

police officers in investigating accidents. A series of quasi-induced exposure-related studies (Carr 1969, Hall 1970, Carlson 1970, Joksch 1973, Cerrelli 1973) have also demonstrated the use of police officer's interpretation as the standard to allocate responsibility in an accident. For example, Cerrelli (1973) initiated a study of determining the insurance premium rates for various driver classes based on the "hazard index" which is similar to the concept of the relative accident involvement ratio in the quasi-induced exposure approach. The hazard index is defined as the percentage liability of certain driver divided by the percentage exposure of the corresponding driver. While determining the liability of each driver involved in accidents, the author simply used the responsibility code in the accident record and split drivers into two groups: responsible and non-responsible.

Although reviewing recent publications pertaining to quasi-induced exposure (Lighthizer 1989, DeYoung 1997, Stamatiadis 1998) did not specifically reveal what technique had been used to assign responsibility, communication with these three authors confirmed that the police's citation was the main standard. Nevertheless, using a police officer's judgment in assigning responsibility in two-vehicle accidents might cause some problems. The validity of assignment of responsibility by police was questioned by Haight (1970) who stated that "it would, except in very well-defined circumstances, be assuming too much if we supposed that the proportions of guilty and innocent parties were decided by the reporting authorities." Quasi-induced exposure emphasizes only whether a driver's behaviors should be responsible for the accident. When investigating police officers issue a citation to the "responsible" driver, their judgments are not necessarily solely based on the driver's driving behavior but a combination of factors. For

example, when police officers investigate traffic accidents, there may be a sort of "negative halo effect" (DeYoung 1997), where the investigating officers are more likely to assign the responsibility for the accident causation to a driver once they determine that the license status of the driver is suspended or revoked, alcohol or drug usage is involved, or an open container is found in the vehicle, regardless of whether the driver is involved any hazardous driving actions. If this occurs, it would inflate the involvement ratios for the groups exhibiting these behaviors. Furthermore, with fatal and serious injury accidents, the investigating police officers most probably will not issue a citation to the deceased or seriously injured driver even if that driver performed hazardous actions before the accident (DeYoung 1997). If this "under-assignment" occurs, it would underestimate the involvement ratios for the corresponding groups.

When utilizing quasi-induced exposure technique to investigate young drivers' (16-20) behaviors with Kentucky accident data, Aldridge et al. (1999) and Kirk et al. (2001a) happened to address the issue of how to determine the accident fault in a crash with the same methodology. They recommended using variables describing human factors in each accident entry, which indicated what each driver did to contribute to the accident. If the record shows no contributory factors present for a particular driver, then the driver is defined as non-responsible; if any contribution to accident causation is evident, the driver is considered to be responsible. In Kentucky accident data, the human factors include unsafe speed, failure to yield right-of-way, improper passing, and drug or alcohol involvement. The authors also argued that there was an inherent bias in relying on the judgment and inclination of the police officer to assign contributing factors.

Therefore, the validity of solely using police officers' citations in assigning responsibility is suspect. It could result in over- or under- estimated involvement of certain driver-vehicle combination in accidents. It is more appropriate to assign responsibility by collectively considering several variables in the accident report such as, hazardous actions, most harmful event, and violations. It is emphasized that, as far as the quality is concerned, these variables are fundamentally identical, since these variables are observed by the same investigating police agent. What makes the difference is that driver's citation status is determined according to a combination of violation phenomena (which could be hazardous actions, drinking, or revoked/suspended license), while driver's hazardous action is directly collected or estimated from the accident scene.

2.4.2.3 Applications

Since Carr (1969) developed quasi-induced exposure, it has been used more frequently than any other induced exposure formulation. Interestingly, for more than a decade after the concept of quasi-induced exposure was first introduced, there was no real practical application of this approach. Both Kuroda, et al. (1985) and Maleck and Hummer (1987) presented a similar study on investigating the relationship among driver characteristics, vehicle size, and IR by using quasi-induced exposure method. The results showed that relationship between driver age and vehicle size were not apparent, while driver age appeared to affect the IR for all vehicle weight classes, especially in urban driving conditions. However, there are some fundamental problems associated with the study. In the event of utilizing Michigan DOT accident data, the authors simply assumed that "accident reports were properly completed" and treated the second vehicles in two-vehicle accidents as the not-at-fault drivers (D2s). First, it has been shown that the raw

accident data contain substantial data errors and the quality of accidents is questionable. Second, in two-vehicle crashes, there were always some cases where both vehicles involved in the same accident were responsible or non-responsible for the accidents.

Considering the problems with the use of quasi-induced exposure, Lyles et al. (1994) summarized their experience and previous researcher's work and recommended some general guidelines for when/how/where to use (and not use) the quasi-induced exposure method in the real world applications. The following table is adapted from the original work.

Table 2.4. General guidelines to use or not use quasi-induced exposure

to use	not to use
when relative rates are adequate	when D2 distribution is biased
when accident data are available and other data are not	when the sample size is small
when data can be "cleaned"	when data are not "cleaned"
when the D1-D2 matrix is "stable"	when the D1-D2 matrix is "unstable"

In table 2.4, "relative rates are adequate" refers to the fact that quasi-induced exposure can only produce *relative* exposure, and not explicit rates such as accidents per million vehicle miles. For example, if there is a necessity to calculate an accident rate for a driving cohort, an exposure method other than quasi-induced exposure must be used. If it is sufficient to compare the relative accident rates between two driving-vehicle cohorts, quasi-induced exposure is certainly a good technique. "Data can be cleaned" refers to the fact that the data set must be capable of being "cleaned" to eliminate unreliable data, e.g., data from certain kinds of accidents (e.g., hit-and-run accidents) and where bias is evident. "D1-D2 matrix is stable" means that the row distributions in the matrix are similar or insignificantly different. In reference to table 2.3 where Lighthizer's (1989)

"complementary sets analysis" technique is discussed, the percentage of D2 auto vehicles (92.0%) in auto-auto accidents is close to those (90.5%) in pickup-auto accidents, suggesting that auto vehicles are randomly impacted in two-vehicle accidents.

DeYoung et al. (1997) applied the quasi-induced exposure method to fatal crash data to generate exposure and crash rate estimates for Suspended/Revoked (S/R) drivers in California. Based on California fatal two-vehicle crashes (1987-1992) on the highway, the D1-D2 matrix showed exposure rates of 8.8% and 3.3% for S/R and unlicensed drivers, respectively, and that, compared to valid licensed drivers, the former were overinvolved in fatal crashes by a factor of 3.7:1, and the latter 4.9:1. The study raised an issue concerning the degree of accuracy of the exposure estimates of S/R drivers. Using quasi-induced exposure, the D2 percentage for S/R drivers is 8.8%, while a random sample of all California drivers shows 5.5%. The authors noted that the former were adjusted for S/R accidents on the highway and thus accounted for driving, while the latter was simply a population proportion. In general, the authors argued that quasi-induced exposure yielded a reasonable approximation of the risks posed by S/R and unlicensed drivers and the reflected facts provided a compelling rationale for seeking more effective methods of enforcing laws prohibiting driving without a valid license. However, the authors fail to address the basic issue in using quasi-induced exposure—the validity of this exposure measurement. There is no evidence to ensure that the underlying assumptions of quasi-induced exposure are fully supported by the given accident dataset.

In a paper by Stamatiadis et al. (1998), the quasi-induced exposure technique was used to identify potential socioeconomic factors that could contribute to the relatively high fatality crash rates in the Southeast US and to develop preliminary relationships

between socioeconomic characteristics and crash trends. The authors calculated relative accident involvement ratios for single-vehicle and multi-vehicle crash rates and determined the real effect of one particular group and its tendency to be involved in crashes, assuming that 1) drivers involved in single-vehicle accidents are all "responsible" and 2) the exposure of a certain driver-vehicle combination is the total number of corresponding non-responsible driver in two-vehicle crashes. Although the first assertion can be reasonably assumed, the second assumption is questionable on the grounds that there is no discussion of the validity of using quasi-induced exposure.

In two similar studies, Aldridge et al. (1999) and Kirk et al. (2001a) both used the quasi-induced exposure technique to evaluate young drivers' accident propensity with three different passenger groupings (no passengers, peer, and adult or child) and four prominent crashes (left-turn, rear-end, single-vehicle, and passing crashes), respectively. Both studies have made improvement in addressing the issue of assigning accident fault by examining the specific human factors instead of police officers' judgment and thus eliminated the assumption that all drivers in single-vehicle accidents are responsible for the occurrence of accidents. The human factors found in the accident database include unsafe speed, failure to yield the right-of-way, improper passing, drug or alcohol involvement. However, no attention has been given to the validity of using quasi-induced exposure, which makes the results and conclusions less convincing and theoretically flawed.

2.5 Summary

In summary, if it works, the quasi-induced exposure method has a great advantage in being used to estimate the exposure of stratified vehicle/driver combination with

accident data, which can't be achieved using the conventional exposure methods, like VMT. The extensive review of the literature has demonstrated that there are two key elements of quasi-induced exposure that remain unresolved or have been resolved unsatisfactorily—validation of its fundamental assumptions and responsibility assignment.

Most of the authors simply apply quasi-induced exposure in their perspective studies without taking into account the validity of the underlying assumptions. Although some work has been done by Lighthizer (1989), Lyles et al. (1994), and Kirk et al. (2001b), it is either too qualitative or theoretically imprecise.

In terms of assigning accident responsibility, except for two most recent studies (Aldridge et al. 1999 and Kirk et al. 2001a) suggesting the use of human factors, the majority of literature heavily relies on the judgments (or the citation in the accident report) of investigating police. The police officers' "verdict" for accident fault is necessary in determining the accident responsibility, but not sufficient. Factors like a driver who has been drinking and/or using drugs, or has an invalid driver license might potentially lead police officers to issue citations. Certainly, this is of no help in determining the accident propensity for a specific driving cohort of interest.

The purposes of this dissertation are to develop a systematic procedure to assign responsibility for accident causation; to determine if the underlying assumptions of quasi-induced exposure are valid; and, ultimately, to determine if quasi-induced exposure can be used with confidence. If so, guidelines will be developed to indicate at what level of data aggregation or circumstances quasi-induced exposure is applicable.

Chapter 3

PROBLEM STATEMENT

The literature review revealed some of the problems with the assumptions inherent in using traditional exposure measures in traffic safety analysis. By way of summary, these include:

1. All driving involves the same exposure to accident hazards (assumption one). Not all driving (vehicle miles) is subject to similar driving hazards under certain circumstances. For example, vehicles traveling in a platoon at an identical speed where the leading vehicles might experience more driving hazards or differently than the following vehicles.

2. Exposure to accident hazards is always proportional to miles driven (assumption two). This assumption is really a matter of which level of accident data are considered. When VMT is used as an estimate of exposure at the system level, the assumption is generally valid. When exposure needs to be estimated in a disaggregated manner by other variables of interest (e.g., roadway type, driver age), the assumption becomes less appropriate.

3. The degree to which exposure is associated with miles driven is the same for all drivers (assumption three). Substantial differences exist between different drivers in terms of driving knowledge and experience and thus drivers might respond differently to the same type of driving hazard. Given this setting, for the same miles driven, experienced drivers might "feel" being exposed to fewer hazards than inexperienced drivers do, even if the hazards themselves are the same. The point is that the objectively same situations are not equally hazardous for different drivers.

4. The traveling speed for groups under scrutiny is inherently assumed to be equal (assumption four). In order to illustrate the point, imagine a potential at-fault driver waiting somewhere to encounter or interact with two innocent drivers, one traveling for 10 miles at 30 MPH and the other for 20 miles at 60 MPH. The at-fault driver will have two equal opportunities (time duration) to interact with an innocent driver. The probability of each of these innocent vehicles being impacted by an at-fault driver would be the same. Using the traditional VMT method would result in the vehicle that is traveling twice as far having twice the probability of being in an accident.

It is not argued that these assumptions should not be made (certainly VMT is a good measure of exposure "truth" in general), but to illustrate the point that use of VMT as the measure of exposure might introduce bias into the analytical results under certain circumstances. In addition to these assumptions, the use of VMT has one fundamental drawback in the calculation of accident exposure in practical applications: general availability of data. General availability of data refers to the fact that additional information (e.g., traffic volume, length of road segment) beyond the accident data themselves are necessary in order to compute the exposure and accident rates for certain driver-vehicle groups. Normally, this type of information is not available. Furthermore, limited availability of data leads to another issue—finer disaggregation of exposure. There are problems with the use of VMT as an estimate of exposure at other than the system level. VMT is typically unavailable for specific driver-vehicle groups under certain conditions, for example, young drivers' exposure in the downtown area.

In light of the theoretical and operational problems involved with using VMT, quasi-induced exposure seemingly provides acceptable and practical solutions to at least some of the problems. It craftily avoids the assumptions that have theoretical loopholes through a relatively simple postulation instead—non-responsible drivers involved in accidents are a random sample of the whole driving population on the road. Furthermore, from the operational perspective, its attractiveness lies in the utilization of already

available data to calculate exposure without requiring additional information and the capability of measuring exposure disaggregated by specific variables of interest. These two features, if both can validated, will overcome/avoid some the fundamental problems confronted in the most traditional exposure method (VMT), including assumptions one, two, and three.

However, quasi-induced exposure theory is far from being mature and has its own set of problems. It was seen in the literature review that although safety researchers have made substantial theoretical progress with quasi-induced exposure and implemented it in a variety of safety/crash studies, the fundamental concerns/problems with the method still remain unresolved. There are two major problems with quasi-induced exposure:

1. The first is in the preparation of accident data to make them readily available for use in analysis by quasi-induced exposure. This includes developing a systematic procedure to make accident data relatively error-free and accurately assigning responsibility for each accident.
2. The second is in validation of the underlying assumption that the collection of non-responsible driver-vehicle combinations involved in two-vehicle accidents is a random sample of the driving/vehicle population on the road at the time and place of those accidents.

The first problem arises from the requirement that accident data should be presented in a way such that:

1. Two-vehicle accident data are selected with other data being discarded and responsibility for accident causation clearly assigned (one is at-fault and the other is innocent). Note that word "discarded" does not imply throwing away data for

good, but not using it in this case. Three-or-more-vehicle accidents contain a lot of driver-vehicle information and their use will be detailed in a later chapter.

2. The information on each driver-vehicle combination in the accidents is sufficiently recorded such that variables of interest are not missing.
3. The information on each driver-vehicle combination is reasonable and contains minimum miscoded data information. For example, an accident labeled as "two-vehicle accident" should not include any one-vehicle or more than two-vehicle information.

The second problem is the basic issue in using quasi-induced exposure. The underlying assumption can be rephrased more specifically: non-responsible drivers (D2) involved in accidents are a random sample of the driving population on the road when and where accidents occur. Only when this assumption is satisfied, can quasi-induced exposure be used with confidence.

The resolution of these two issues is the goal of this research and, eventually, guidelines will be provided on how to use (or not use) the quasi-induced exposure approach. Each will be discussed in more detail in the following sections.

3.1 Development of systematic rules for preparing accident data

While accident data are readily available, they are often not useful in analysis without at least some preliminary manipulation or screening. Thus, "preliminary screening" of accident data is the first step in using quasi-induced exposure. It is of great importance because data errors introduced in the first stage will propagate through the rest of the process. For quasi-induced exposure, the accident data should only include two-vehicle accidents with clearly assigned responsibility for the accident (one is

responsible and the other is non-responsible). The literature review revealed that traffic researchers often do not pay sufficient attention to these concerns when using quasi-induced exposure. Therefore, it is worthwhile to explore the potential problems that occur when these issues are not addressed properly or sufficiently.

Examining the whole process of how accident data are obtained will help to reveal different types of errors hidden in the data. In terms of a sequence of events, an accident happens, police officers investigate the accident and record driver-vehicle information on an Accident Report (AR) according to a specific format, and, finally, the accident report is converted into a digitized format. The following table shows the possible errors during this process.

Table 3.1. Potential cause and description of errors in accident data

cause of errors	description of errors
limitation of AR format	missing descriptive information on accident, driver(s), or vehicle(s)
hit and run driver(s)	missing information on the hit-and-run driver-vehicle(s)
police officer records incomplete information in AR	missing descriptive information on accident, driver(s), or vehicle(s)
police officer records incorrect information in AR	conflicting information between variables in AR
AR is incompletely digitized	missing descriptive information on accident, driver(s), or vehicle(s)
AR is incorrectly digitized	conflicting information between variables in AR or in the digital form

In table 3.1, errors are combined together, which are shown in a variety of forms in the accident data. Descriptive statistics for raw accident data confirm that the errors such as those above occur in Michigan (using 2000 data as an example):

- Missing information in an AR, e.g., driver gender and/or driver age information;
- Abnormal values of the variables; and

- Conflicting information between variables, e.g., some variables in the record indicate that a particular accident involves only "one-vehicle" but data are shown for both driver 1 and 2 characteristics.

In the context of quasi-induced exposure, missing information in an AR can lead to incomplete description of characteristics of certain driver-vehicle combinations and/or the assignment of fault. Eventually, these errors will affect the relative accident involvement ratio (IR) of a specific driver-vehicle cohort in a profound manner, since they could cause the reduction concurrently in the magnitudes of the numerator and denominator in IR. Thus, whether the IR is underestimated or overestimated hinges on the relative reduction proportions of the numerator and denominator.

Using quasi-induced exposure without considering the quality of the raw accident data can result in biased, or erroneous analytical results. Since most of the raw accident data can not be ensured error-free, a logical procedure must be developed to "clean" accident data as much as possible before it is used, which includes specific operations aiming at eliminating more apparent data errors.

In addition to selecting and using only accident data that are relatively error-free, assigning responsibility is another important concern. Responsibility for accident causation is a critical issue in using the method. After the accident data are initially manipulated to be largely error-free, two-vehicle accident data are ready for responsibility assignment. It is noted that responsibility might be logically assigned to either driver in the AR, both, or neither in a two-vehicle accident. However, quasi-induced exposure explicitly requires that for each two-vehicle accident used in the analysis, only one driver (or driver-vehicle combination) involved be responsible (or "at-

fault") and the other be non-responsible. Thus, accidents with two "responsible" or "non-responsible" drivers cannot be used in application of quasi-induced exposure. Routinely, responsibility can be determined in individual cases on the basis of police reports or other supplementary investigations. The most commonly used indicator is whether the driver is issued a ticket by the investigating police officer. Several researchers (Lighthizer 1989, Davis et al. 1993, Stamatiadis et al. 1995 and 1997, David et al. 1997, Aldridge 1999, Kirk et al. 2001) using the quasi-induced exposure have simply assumed that the determination of responsibility in a two-vehicle accident is completely dependent on whether the driver is given a citation. When investigating police officers issue a citation to the responsible driver, their judgments may sometimes be based on a combination of factors. They may be more likely to assign the responsibility to a driver once they determine an indication of a traffic violation, regardless of the hazardous driving actions. If this occurs, it would inflate the involvement ratio for the groups containing this type of driver.

The point here is that a "violation" alone is not necessarily sufficient to allocate responsibility in a two-vehicle accident, and thus other variables need to be taken into consideration. At least a partial solution to this concern is to develop a systematic procedure for assigning responsibility to individual driver-vehicle combinations depending on the hazardous action(s) that a driver takes right before the accident and the violation indicator itself. Therefore, it is important that two variables be taken into consideration in a combinatory manner when relative guilt or innocence is assigned.

In summary, preparation of accident data is a necessary and important step in order to use quasi-induced exposure. It helps to screen out data errors and to assign responsibility for accident causation in a systematic manner.

3.2 Validation of assumptions

Fundamental to quasi-induced exposure is that distributions of the characteristics of the non-responsible driver-vehicle combinations in traffic accidents have an identical distribution to the same driver-vehicle combination in the driving population on the road at the time and place of the crash.

The problem still lies in how to develop practical approaches to validate this assumption. One approach is to compare the driver-vehicle characteristics from the non-responsible driver-vehicle category in accidents, with those from the driving population on the road. If similar distributions can be routinely confirmed, quasi-induced exposure is validated. The following two approaches are developed to address this issue:

- Compare various D2 distributions with more conventional estimates of exposure "truth" from other sources; and
- Compare D2 distributions derived from two-vehicle accidents with those from three-or-more-vehicle accidents.

As for the first approach, the idea is to compare the distributions of specific driver-vehicle combinations (D2) from the crash data to "truth" from other data sources. The following figure shows several exposure data sources that can provide comparisons (albeit with differing utility) with the D2 distributions derived from accident data.

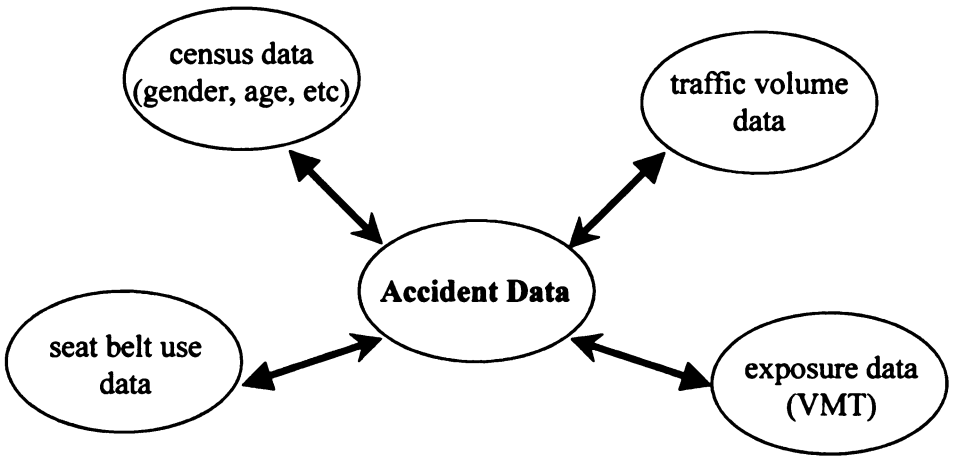

Figure 3.1. Exposure data sources comparable with accident data

It must be noted that some of the distributions calculated from other data sources can be directly compared with the accident data and some cannot (such as census data). For instance, the distributions of non-responsible driver-vehicle combinations in the accident can be directly compared with those of traffic volume data on condition that those two sets of data are collected under similar or identical road and traffic conditions (such as, the same route, time of day, and day of week). If not, some conversions and assumptions are needed to make them comparable. Depending on whether consistent distribution patterns can be found between accident data and other "true" data sources, the underlying assumption of quasi-induced exposure can be validated.

As for the second approach, the basic idea is to compare the non-responsible driver-vehicle distributions calculated from three-or-more-vehicle accidents with those from two-vehicle accidents. Traditionally, use of quasi-induced exposure only considers two-vehicle accident data with responsibility clearly assigned with other accident data being discarded. However, if the non-responsible driver-vehicle combination in two-vehicle accidents is truly a random sample of the driving population, one should expect

that the non-responsible driver-vehicle combinations in three-or-more-vehicle accidents are also random samples from the driving population.

3.3 Summary of problems

As discussed, the following list summarizes the problems that this research effort attempts to solve:

- Data errors inherent in the accident database;
- Responsibility assignment for accident causation; and
- Systematic validation of the underlying assumptions of quasi-induced exposure.

The work here will concentrate on the validation of the fundamental assertions that non-responsible driver-vehicle combinations in two-vehicle accidents are indeed a random sample of the driving population on the road. To the extent that the approach is validated, the research also includes development of guidelines/rules for when and where it is appropriate to use quasi-induced exposure.

Chapter 4

SYSTEMATIC RULES FOR PREPARING ACCIDENT DATA

As discussed in chapter 3, accident data are not readily usable for quasi-induced exposure application without some preliminary manipulation or screening. Examination of accident data reveals several associated issues:

- Missing accident records;
- Missing variable information in police accident reports;
- Unreasonable values for the variables; and
- Conflicting information between variables.

In addition, when responsibility assignment is explored in the context of quasi-induced exposure, it is necessary to determine if the accident causation assignment results from the "negative halo effect." If so, they will be eliminated.

It is the purpose of this chapter to develop a systematic procedure to screen the accident data obtained from state departments of transportation (DOT) to minimize data errors and make accident data dependable to be readily usable in quasi-induced exposure applications. In general, ready-to-be-used accident data records are fully coded for the variables of interest (e.g., driver age, gender), reasonable with respect to values taken for the variables, without contradiction between different variables, and, equally important,

have clearly and logically assigned responsibility for accident causation. In order to achieve the stated goal, a series of operations are developed in a stepwise manner:

1. The first step is to identify and eliminate the accident data with potential problems and errors. This can be achieved by choosing different screening variables corresponding to different accident data problems.
2. The second step is to logically assign responsibility for the accident to each individual driver-vehicle combination. Accident fault is determined on the basis of two variables, i.e., hazardous action and violation indication.

The first step is only executable on existing accident cases and does not address the "missing record" problem. The results of the literature review have shown that the missing record problem is attributed to factors such as weather conditions, severity of accidents, minimum threshold limitations, and availability of police in particular locations (e.g., Detroit area, MI). Unfortunately, there is no reasonable solution to the missing record problem. For example, it is known that the missing record problem in the Detroit area is due to the lack of police and a policy of simply not reporting minor accidents. When an analysis is developed at a statewide level, accidents occurring in the Detroit area should be excluded. The tradeoff is that some bias could be introduced if accidents in this area have some unique characteristics (e.g., a large percentage of young drivers).

4.1 Identification and elimination of bad data

Taylor and Maleck (1987) identified some problems in the Michigan accident database, indicating unreasonable variable values, conflicting information between variables, and vehicle type misclassification. Review of recent Michigan accident data

(2000) has also indicated that similar problems exist in the accident database. This includes cases with unreasonable values of variables, or conflicting information between values of variables or combinations of them. Identification and elimination of these errors are executed in a combinatory manner, since some cases have more than one type of error involved. Michigan accident data from 2000 (abbreviated as Michigan 2000) are chosen as an example to illustrate how to identify and/or eliminate problematic data.

Accident data will be classified into three categories by means of the number of vehicles involved in the accident (*numveh*): one-, two- and three-or-more vehicles. Note that the discussion here is focused on two- and three-or-more-vehicle accident data. Two-vehicle accidents are chosen because of the specified data requirement by quasi-induced exposure. Three-or-more-vehicle accidents are selected because they might provide an alternative to validate the underlying assumption of quasi-induced exposure. Comparison of the distributions for the non-responsible driver-vehicle combinations in three-or-more-vehicle accidents with those from two-vehicle accidents may be another way to validate the assumptions of quasi-induced exposure.

Several factors will be discussed for the purpose of minimizing errors in the accident data and in assigning the responsibility for accident causation. These include drinking/drug use, hit-and-run crashes, driver age, accidents with conflicting information, and accident type.

Drinking/drug use

"Drinking/drug use" is a responsibility assignment issue. With respect to accident causation, and as noted in the literature review, drivers with alcohol involvement or drug use **may** have a higher tendency to be issued tickets by the investigating police officers

regardless of what hazardous actions they took and their level of responsibility in causing the accident. The consequence is that the involvement ratio for groups containing drinking or drug use drivers will be inflated. However, there is a potential problem in eliminating the accidents with drinking/drug use drivers. Commonly, young drivers are believed to have a higher tendency to be associated with drink-and-drive accidents, and be more prone to take dangerous actions under the influence of alcohol/drugs. If this is so, the drinking young drivers are most likely to be the *real* guilty party in two-vehicle accidents. As a result, elimination of accidents with drinking/drug use by young drivers will reduce the number of D1s and the relative accident involvement ratio for young drivers will be underestimated correspondingly. Therefore, there is a tradeoff in eliminating accidents with drinking/drug use drivers.

In Michigan 2000, there is a variable *drinking* indicating whether any driver in the accident has been drinking. Statistics show that approximately 4.1% of all accidents involved at least one drinking driver, among which about 1.7% are two-vehicle accidents. Table 4.1 presents the frequencies and percentages of each driver in alcohol-related two-vehicle accidents. The drinking status of each individual driver is identified by another separate variable (*v1drink, v2drink*). From the table 4.1, 5,545 first drivers in two-vehicle accidents are involved with alcohol use and 1,824 second drivers.

Table 4.1. Frequencies and percentages of drinking drivers in two-vehicle accidents

drinking condition	first driver*		second driver**	
	N	%	N	%
drinking	5,545	79.2	1,824	31.0
non-drinking	1,460	20.8	4,058	69.0
total	7,005	100.0	5,882	100.0

* First driver refers to the first driver-vehicle in a typical accident record, not D1;
** Second driver refers to the second driver-vehicle in a typical accident record, not D2.

For these particular accidents with drinking driver(s), one issue is explored—how the incidence of drinking affects the decision-making of investigating officers issuing a citation.

Based solely on the accidents with drinking drivers, the cross tabulation between the variables *hazardous action* and *violation indication* is shown in table 4.2. Although drivers shown in table 4.2 are all intoxicated, drivers with no hazardous actions consistently do not receive citations (for both drivers).

Table 4.2. Cross-tabulation between hazardous action and violation indication

hazardous action	violation indication (first driver)		violation indication (second driver)	
	No	Yes	No	Yes
none	226	0	445	0
speed too fast	0	295	0	75
speed too slow	0	13	0	3
failed to yield	0	718	0	220
disobeyed TCD	0	427	0	92
drove wrong way	0	35	0	12
drove Lt of center	0	304	0	61
improper pass	0	61	0	16
improper lane use	0	238	0	50
improper turn	0	111	0	28
improper signal	0	8	0	4
improper backing	0	154	0	28
fail to stop ACD	0	1,273	0	311
other	0	957	0	270
unknown	103	0	39	0
reckless driving	176	0	54	0
careless/negligent	246	0	0	0
uncoded & errors	200	0	116	0
total	951	4,594	654	1,170

It suggests that drinking factors do not come in play in assigning crash responsibility. In other words, the investigating officers seem to determine the status of traffic violation without considering the drinking factor. When drivers commit specific

hazardous actions, they are issued tickets. The point is that based on Michigan data, drinking does not appear to contribute to a negative halo effect in determining "fault" for the accident. It is possible that drinking drivers receive two citations: one for drinking and one for apparent hazardous action. Unfortunately, these two citations can't be distinguished in the Michigan 2000. In table 4.2, it needs to be pointed out that the values in the shaded cells do not appear to be logical since the drivers' hazardous actions are evident and they receive no citation. This issue will be further explored in a later section.

In summary, analysis of Michigan 2000 has demonstrated that driver's drinking seemingly has no or little effect on the decision to issue a citation. This suggests that when quasi-induced exposure is used, the accidents with drinking driver(s) should not be excluded from analysis. Otherwise, young drivers as the guilty party will be underestimated.

It is noted that accidents involved with drug usage will cause a similar problem as those with drinking for corresponding driver-vehicle combinations. Unfortunately, in the Michigan accident database, there is no variable specifically showing whether drivers involved in accidents use drugs.

Hit-and-run crashes

"Hit-and-run crashes" also present problems with respect to the responsibility assignment issue. Once an accident occurs and one of the drivers flees, it is obvious that the investigating police officers will generally be unable to record the accident information correctly or completely. Moreover, the police officers may not issue a citation to the remaining driver regardless of his/her hazardous action before the accident. Table 4.3 shows the statistics for Michigan 2000—about 10.3 percent of all accidents

involve hit-and-run drivers. Because the records are incomplete, these accidents should be removed.

Table 4.3. Percentage and frequency of hit and run crash

hit and run	frequency	percentage
no	359,538	89.7
yes	41,313	10.3
total	400,851	100.0

Driver age

Examining the frequency distribution of driver age for Michigan 2000 shows that some drivers' age is below 15 and above 100 (table 4.4).

Table 4.4. Frequencies and percentage of unusual driver ages

age range	first driver		second driver		third driver	
	N	%	N	%	N	%
0-14	1,509	0.40	1,613	0.42	70	0.02
15-100	37,3507	99.59	379,966	99.57	423,317	99.98
>101	42	0.01	39	0.01	6	0.00
total	375,058	100.00	381,618	100.00	423,393	100.00

It can be seen from table 4.4 that 99.6% of all drivers are in the range of 15-100. For those under 15, based on Michigan law, drivers are not issued any type of valid driving license until the age of 14 years 9 months. It is possible that some young drivers illegally get behind the wheel, but, more likely, the phenomenon is due to data errors. For those above 100, some drivers' ages are as high as 400. This is obviously attributed to the errors during the data processing. However, since there is less than 1% driver age under 15 or above 100 in Michigan 2000, removal of accidents as such is not expected to make a significant difference in the analytical results. All drivers outside the range of 15-100 are removed.

Accidents with conflicting information

"Accidents with conflicting information" is another quality of data issue. A typical check for conflicting information that is (at least implicitly) built in the accident data is the instance when the number of vehicles involved in an accident (*numveh*) is not equal to the number of drivers identified in the record. For instance, some one-vehicle accidents have information for the second driver (e.g., 4,596 cases or 3.2% in Michigan 2000). It is more appropriate to treat these accidents as "multiple-vehicle" than "single-vehicle" accidents. A feasible solution is to physically count the number of vehicles and drivers information after a problematic accident is located and to redefine the value of the variable *numveh*. Since quasi-induced exposure only utilizes two-vehicle accidents, special attention will be given to mislabeled one-vehicle accidents (approximately 1.2% in Michigan 2000) or three-or-more-vehicle accidents (approximately 0.2% in Michigan 2000). Once these accidents are identified, they will be treated as two-vehicle accidents.

Accident type

"Accident type" is concerned with quality of data as well. Under a given road and traffic setting, there might be a limited number of accident types that could occur. The purpose of checking accident type is to see if there are any unreasonable accident types under a given circumstance. For example, on a freeway section the accidents should not consist of any types such as bicycle, dual left turn, or dual right turn, considering the limited accessibility and nature of freeway environments. In this context, I-94 (in Michigan) is chosen as an example to show accident types, frequencies, and percentages of two-vehicle accidents (table 4.5).

Table 4.5. Accident types, frequencies, and percentages on I-94

crash type	N	%
hit train	0	0.00
pedestrian	3	0.08
bicycle	0	0.00
hit parked vehicle	6	0.16
miscellaneous multiple vehicle	416	10.98
angle straight	275	7.26
angle turn	4	0.11
head on left turn	1	0.03
rear end straight	1,856	48.98
rear end left turn	9	0.24
rear end right turn	17	0.45
dual left turn	5	0.13
dual right turn	6	0.16
head on	63	1.66
side-swipe same	998	26.34
side-swipe opposite	113	2.98
angle drive	0	0.00
rear end drive	3	0.08
other drive	0	0.00
backing	10	0.26
parking	4	0.11
total	3,789	100.00

The first column in table 4.5 is the list of possible accident types in Michigan data. Since the accidents occurring on I-94 are examined, it is logical that there are no "hit train," "angle drive," and "bicycle" accidents. There are several accident types, which are most relevant to freeway off-ramps, rest areas, or illegal use of medium openings, including "angle turn," "head-on left-turn," "rear-end left-turn," "rear-end right-turn," "angle straight," "dual left-turn," "dual right-turn," "side-swipe opposite," "angle drive," "other drive," "backing," and "parking." The point is that if a study is developed to focus on the accidents on a freeway segment, accidents occurring at the freeway off-ramps or rest areas should be eliminated. The identification and elimination can be achieved based on the crash type (*crshtype*). In table 4.5, three shaded accidents

are believed to be the most commonly occurring types for two-vehicle accidents in the freeway sections and consequently are considered for further analysis. The rest of the accidents should be excluded for the mainline analysis.

The analytical process described above is graphically illustrated in figure 4.1:

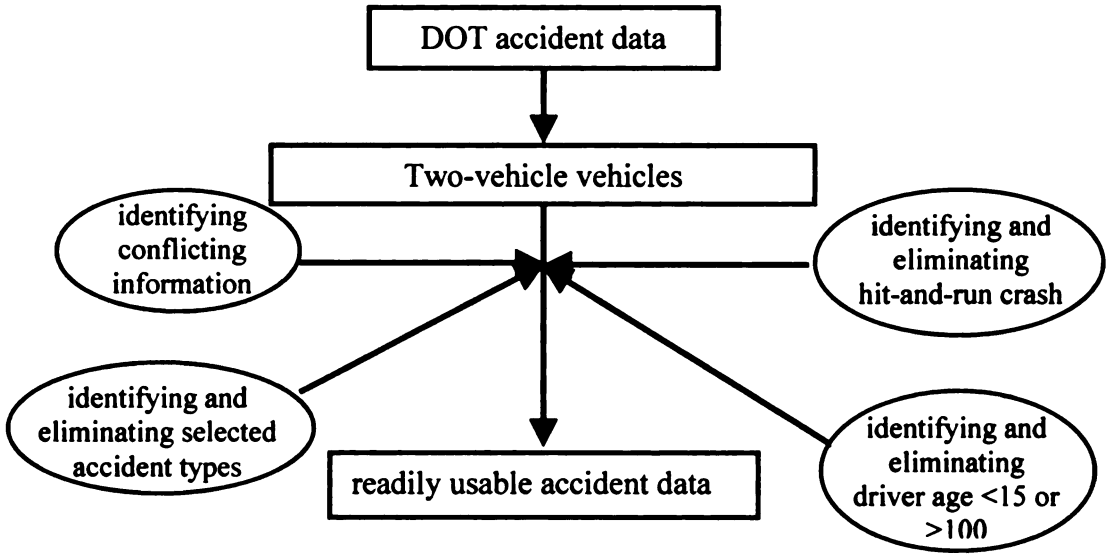

Figure 4.1. Flow chart of identifying and/or eliminating bad data

In summary, after these operations are executed, accident data are expected to be more reasonable and "cleaner" in the sense that there is less conflicting information in an accident case, no accident with driver age falling below 15 or above 100, no hit-and-run type of accidents, and no unreasonable accident types for a given circumstance.

A note on the issue of identifying and eliminating some accident data is the percentage of accidents that drop out. After the stated analytical process, approximately 56.0% of the accidents are filtered out from Michigan accident database (2000). Arguably, it is assumed that different driver-vehicle combinations are subject to the data errors in the same manner and elimination of those data will not introduce bias to the results. Table 4.6 shows the number of accidents and the percentages left after each step (Michigan 2000).

Table 4.6. The number of accidents and the percentages left after each step

Steps	number left (N)	percentage left (%)
1. total accidents	427,353	100.0
2. two-vehicle accidents	263,093	61.6
3. conflicting info	262,974	61.5
4. hit-and-run	226,347	53.0
5. driver age	187,845	44.0

Note that in each step, the operation is executed directly on the number of accidents left from the preceding step. The most significant step is where accidents are categorized based on the number of vehicles involved in an accident: about 38.4% of all accidents are filtered out. The percentages dropped out due to hit-and-run crashes accounting for 17.5%. The accidents shown in the last step are still not readily usable by the quasi-induced exposure technique. For each individual driver, the status of responsibility for accident fault is not clear at this point. How to assign responsibility will be discussed next.

4.2 Responsibility assignment

Two topics will be covered in this section: 1) how to assign responsibility for two-vehicle and three-or-more-vehicle accidents; and 2) whether accident data recorded by different police agencies make any difference in terms of responsibility assignment.

4.2.1 Assigning responsibility

Responsibility assignment is a critical issue because the construction of the D1-D2 matrix is based upon it. The literature review has indicated that several authors (Lighthizer 1989, Davis et al. 1993, Stamatiadis et al. 1995 and 1997, David et al. 1997, Aldridge 1999, Kirk et al. 2001) using the quasi-induced exposure have simply assumed that the determination of responsibility in a two-vehicle accident is completely dependent on whether the driver is given a citation.

Therefore, driver's hazardous action is investigated to track what was happening before the occurrence of an accident. It is argued that responsibility for accident causation can be more reasonably and consistently assigned, if they are used in a combinatory manner. Variables chosen for this process are:

- The hazardous action (*v1hazact*, *v2hazact*, *v3hazact*), which is used to describe the hazardous actions taken by various drivers; and
- The violation indication (*v1violtr*, *v2violtr*, *v3violtr*), which is used to identify whether the driver receives a citation from the investigating police officer.

The following section will cover the responsibility assignment schemes for two- and three-or-more-vehicle accidents.

4.2.1.1 Two-vehicle accidents

Before the responsibility-assigning scheme is developed for two-vehicle accidents, the variables violation indication and hazardous action are examined. Based on the coding menu, the variable violation indication takes only two values: 0 and 1. When a driver's violation indication is equal to 1, it means that the driver is issued a ticket by the police officer; when 0, there is no citation. The coding for hazardous actions is more complicated.

Table 4.7 shows the coding menu and the frequency of different hazardous actions (first driver, 187,845 cases in total) after the accident data have been systematically "cleaned." Of the hazardous actions in table 4.7, many are clearly interpreted as driving hazards while others either indicate that the first drivers have no hazardous actions (35,277 cases) or actions are unclear or unknown. There are 8,149 accidents with the hazardous action "other" which implies some hazardous actions, but

the actions are not apparent and thus it is impossible to determine what they are; 3919 accidents with "unknown" hazardous action, which suggest that there might or might not be hazardous action; and 1,904 accidents with hazardous action "uncoded and errors," which explicitly advises that certain (but unknown) data errors are affiliated with the coding of hazardous action. Accidents with the hazardous action "uncoded and errors" should be excluded from the analysis.

Table 4.7. Coding menu and the frequency of different hazardous actions (first driver)

description	N	description	N	description	N
none	35277	drove Lt of center	1844	fail to stop ACD	53743
speed too fast	7392	improper pass	2401	other	8149
speed too slow	479	improper lane use	7360	unknown	3919
failed to yield	43807	improper turn	4955	reckless driving	265
disobeyed TCD	9484	improper signal	458	careless/negligent	1038
drove wrong way	223	improper backing	5147	uncoded & errors	1904

For two-vehicle accidents, the relationship between the hazardous action and violation indication for each driver is studied. This helps to provide a general picture of the responsibility-assigning scheme. Table 4.8 shows the cross-tabulation between hazardous action and violation indication. Based on the information displayed in table 4.9, several important observations are made:

- When hazardous action is "none," the driver is not issued a ticket. There are 144,730 accidents with hazardous action "none" for second driver, while only 34,508 accidents for first driver. Comparatively, the first driver is likely to be "at-fault" driver although it is not always the case.
- While hazardous action is self-explanatory and specific (e.g., improper pass, improper signal), both the first and second drivers are consistently issued citations.

- When a driver's hazardous action is "other," the driver is consistently issued a citation by the police officers. This implies that certain specific hazardous actions occur even though the action is not clear.
- When a driver's hazardous action is "unknown," the driver is consistently NOT issued a citation by the investigating police officers.

Table 4.8. Cross-tabulation between hazardous action and violation indication

hazardous action	violation indication			
	first driver		second driver	
label	no	yes	no	yes
none	34508	0	144730	0
speed too fast	0	7349	0	2424
speed too slow	0	477	0	116
failed to yield	0	43617	0	9789
disobeyed TCD	0	9442	0	1956
drove wrong way	0	223	0	49
drove Lt of center	0	1839	0	404
improper pass	0	2384	0	720
improper lane use	0	7321	0	2171
improper turn	0	4927	0	1182
improper signal	0	455	0	266
improper backing	0	5127	0	1238
fail to stop ACD	0	53534	0	12914
other	0	8091	0	2960
unknown	3898	0	3443	0
reckless driving	261	0	86	0
careless/negligent	995	0	0	0
total	39662	144786	148259	36189

As highlighted in the shaded cells, when a driver's hazardous action is "reckless driving" or "careless/negligent," the investigating police agent does not issue a citation to the driver. Based on 2000-2001 Michigan Vehicle Code and Law Related to Ownership and Use of Vehicle (Candice S. Miller, 2001), the definitions of reckless driving and careless/negligent driving are:

> Reckless driving—any person who drives any vehicle upon a highway or other place to the general public, including any area designed for the parking of motor vehicles, within this state, in willful or wanton disregard for the safety of persons or property is guilty of reckless driving (pp. 214).
>
> Careless driving—a person who operates a vehicle upon a highway or other place open to the general public, including an area designated for the parking of vehicles in a careless or negligent manner likely to endanger any person or property, but without wantonness or recklessness, is responsible for a civil infraction (pp. 215).

According to the definitions, these two actions are indeed two driving hazards, which likely triggered the occurrence of the accidents. They could be any driving behaviors endangering the safety of persons or property (no matter intentionally or unintentionally). Therefore, hazardous actions as such are indicative of traffic violations and, thus, drivers with "reckless driving" and "careless/negligent driving" should be assigned accident fault.

According to the above analyses and observations, the following scheme is developed to assign responsibility in a two-vehicle accident:

- When there is no hazardous action, the involved driver is not responsible for the accident occurrence.
- When the hazardous action is apparent, including reckless driving and careless/negligent driving, the driver is responsible for the accident.
- When the hazardous action is "other" (not apparent but nonetheless occurs), the driver is responsible for the accident.
- When the hazardous action is "unknown," the driver is not responsible for the accident.

Given this responsibility-assigning algorithm, each two-vehicle accident will take one of the following three forms: one responsible driver and one non-responsible driver,

two responsible drivers, or two non-responsible drivers. Based on the theory of quasi-induced exposure, the number of non-guilty parties in two-vehicle accidents with one responsible and one non-responsible driver is the relative exposure. For the example of Michigan 2000, 13,707 accidents are eliminated in this step, accounting for 3.2% of all the accidents recorded, and 170,741 two-vehicle accidents remain, approximately 40% of all the accidents in the database. Note that other accident data can be discarded: 1) for accidents with no innocent driver, the involvement ratio will be inflated because D1s are counted but D2s are not; and 2) for accidents with no guilty driver, the involvement ratio will be underestimated because D2s are counted but D1s are not. Now accidents are usable for quasi-induced exposure.

4.2.1.2 Three-or-more-vehicle accidents

Three-or-more-vehicle accidents provide an alternative to validate the underlying assumptions of quasi-induced exposure through a comparison of the distributions for non-responsible driver-vehicle combinations with those in two-vehicle accidents. Three-or-more-vehicle accidents are contained in the same accident database as two-vehicle accidents. Therefore, three-or-more-vehicle accident data are coded in the same manner as those in two-vehicle accidents in the accident database. The responsibility-assigning logic developed for two-vehicle accidents still applies.

An issue surfaces after the responsibility is assigned to the individual driver based on the hazardous action: are all the three-or-more-vehicle accidents eligible for the comparison? The answer is no. In a typical three-or-more-vehicle accident there must be at least one responsible and one non-responsible driver, so accidents with no responsible or non-responsible drivers will be eliminated. It is consistent with the theory of quasi-

induced exposure: there must be an accident initiator and an innocent victim in an accident. However, it is possible that in a three-or-more-vehicle accident there are two or more non-responsible or responsible drivers. The assumption that innocent driver-vehicle combinations are randomly impacted by the guilty party in both two- and three-or-more-vehicle accidents, lays the foundation for the stated comparison. Given this, the distributions of all the non-responsible drivers in three-or-more-vehicle accidents are comparable to D2s in two-vehicle accidents.

4.2.2 Accident data recorded by different police agencies

With responsibility assigned to each individual driver, one might be interested to know whether reporting practice for different enforcement agencies makes any difference in investigating and recording accidents in terms of assigning responsibility for accident causation. The literature review has shown that there is a consistency issue in combining accident data from different states in terms of different definitions of defective vehicles, injury severity, and accident location. However, no past research has been found to explore the issue of combining accident data from different police agencies within the same state.

It is possible that even a well-trained investigating police agent could misjudge the accident scene and record inconsistent accident information. The degree of agreement between a driver's hazardous action and the violation indication will be indicative of the performance of the particular police agency in recording the accident data.

The examination is done in a stepwise manner:

1. Determine responsibility for accident causation based on hazardous actions;

2. Identify accidents where drivers are responsible but not issued citations or drivers are not responsible but issued citations;
3. Determine the frequency of inconsistency (step 2) for different law enforcement agencies;
4. Determine the total number of accidents investigated by different law enforcement agencies (based on total records); and
5. Calculate the percentage of inconsistency by different law enforcement agencies, defined as the ratio of step 3 to step 4.

Conveniently, step 1 can be achieved with the procedure developed in the previous section. In steps 2 and 3, if a driver is assigned the responsibility but not issued a citation or not assigned the responsibility but issued a citation, his/her information has been inconsistently recorded by the police officer. For ease of explanation, the accidents with such information are labeled as "inconsistent" accidents. Based on the percentage of inconsistently reported accidents, the performance of each police agency can be evaluated by comparing it with others.

Using Michigan 2000 data as an example (without any data manipulation), three tables (4.9-4.11) show the cross-tabulations between driver responsibility and violation indication for the first, second, and third drivers, respectively. The highlighted cells in tables 4.9, 4.10, and 4.11 are the number of drivers where responsibility and violation indication are inconsistent. Interestingly, for those inconsistent accidents, it is always the case that drivers have hazardous actions but receive no citations. The percentages for the first, second, and third drivers are 2.0%, 0.1%, and 0.1%, respectively. Since some accident records contain more than one inconsistent driver, the number of accidents adds

up 8477 in total. This subtotal will be further allocated among different Michigan police agencies.

Table 4.9. Crosstabulation of responsibility and violation indication (first driver)

		violation indication		total
		no	yes	
hazardous action	no	155511	0	155511
	yes	8208	249649	257857
total		163719	249649	413368

Table 4.10. Crosstabulation of responsibility and violation indication (second driver)

		violation indication		total
		no	yes	
hazardous action	no	224453	0	224453
	yes	274	51685	51959
total		224727	51685	276412

Table 4.11. Crosstabulation of responsibility and violation indication (third driver)

		violation indication		total
		no	yes	
hazardous action	no	18321	0	18321
	yes	23	3231	3254
total		18344	3231	21575

Four types of law enforcement agencies collect Michigan accident data: Michigan State Police (MSP), county sheriffs, township police, and city and other police. Table 4.13 presents the number of inconsistent accidents (step 4) and total accidents (step 5) reported by different agencies. In table 4.13, the percentage is formulated as the ratio of step 4 to step 5, that is, the proportion of inconsistent accidents recorded by different police agencies. The table shows that township police officers are the most consistent, and then sheriffs, city police, and the MSP. On the other hand, the percentages are relatively close (within one percentage point). It is argued that, operationally there is no

significant difference in the quality of accident data among the four different types of police agencies.

Table 4.12. Inconsistent and total accidents by different police agencies

police type	inconsistent (step 4)	total (step 5)	percentage (%)
MSP	616	34228	1.8
sheriff	1070	71322	1.5
township	363	30215	1.2
city and others	6428	277603	2.3
total	8477	413368	2.1

In summary, Michigan accident data do not show a conspicuous discrepancy of accident data among the data collected by different levels of police agencies, at least as hazardous actions and violation indication are concerned.

4.3 Conclusion

Accident data from state DOTs are a convenient data source for the use of quasi-induced exposure, although the raw data are typically not ready for immediate analysis. In this chapter, the attempt was made to develop a systematic procedure to make the raw accident data relatively error-free, with the example of Michigan accident data. The procedure can be generalized to other states, since the accidents obtained from other states also encompass all the necessary screening variables and information. The "rules" to "clean" accident data are summarized as follows:

- For drinking, the incidence of drinking does not appear to adversely affect the decision-making of investigating officers' issuing citations. Therefore, accidents with drinking driver(s) should not be removed.
- For hit-and-run crashes, the accidents are incompletely recorded. Hit-and-run accidents are eliminated.

- For driver age, some accidents are identified as having drivers with unusual ages (<15 or >100). Although the percentage is negligible, such accidents will be removed.
- The most typical accidents with conflicting information are those two-vehicle accidents miscoded as one- or three-or-more-vehicle accidents. A feasible solution is to count the number of drivers/vehicles after a problematic accident is identified and to redefine the value—the number of vehicles involved in an accident.
- For accident type, special attention should be given to the fact that under a specific circumstance, only certain types of accidents are reasonable. For example, if a study is developed to focus on the accident on a freeway segment, accidents occurring at the freeway on- and off-ramps should be eliminated.

Reasonable responsibility assignment is developed for two- and three-or-more-vehicle accidents, the logic of which is summarized:

- When there is no hazardous action or it is "unknown," the involved driver is not responsible for the accident.
- When a hazardous action is apparent or hazardous action is "other," including reckless driving and careless/negligent driving, the driver is responsible for the accident.
- Accidents with hazardous actions "uncoded and errors" are eliminated.

In addition, it is found that there is no operational difference among the accidents investigated by different police agencies in Michigan.

After this procedure is completed, the data are ready to be used in quasi-induced exposure-related studies.

Chapter 5

VALIDATION OF THE ASSUMPTIONS OF QUASI-INDUCED EXPOSURE

Fundamental to quasi-induced exposure is the assertion that D2s in two-vehicle crashes constitute a random sample of the driving population on the road at the time and place of the crash. The goal of this chapter is to employ three externally available data sources to validate the underlying assumptions. The data used in this chapter include vehicle miles of travel (VMT), safety-belt use data from the University of Michigan Transportation Research Institute (UMTRI), and truck volume data (W-2 data) from the Federal Highway Administration.

The basic idea is to compare the D2 distributions for driver-vehicle characteristics calculated from accident data with those derived from externally available data sources, to see if they are the same. Note that the driver-vehicle characteristics studied include: driver gender, age, and vehicle type. Criteria are established to determine if the difference between two distributions is significant from operational and/or statistical (chi-square test) perspectives. Now, the question remains how to justify the significance thresholds for the practical and statistical differences.

As for statistical difference, a significance level of $\alpha = 0.05$ is a commonly used and accepted threshold in the scientific and engineering-related studies. Based on the chi-square contingency table, if the p-value given by the test is greater than 0.05 (significance

level), the null hypothesis that there is no significant difference of D2 distributions is accepted; otherwise it will be rejected. For practical difference, the significance threshold (μ_0) is determined by satisfying the following constraint:

$$P\left(x \in \left[\bar{\mu} - \mu_0, \bar{\mu} + \mu_0\right]\right) \geq 0.95$$

where,

P is the probability function;
x is an observed percentage (distribution) for certain driver-vehicle characteristic (i.e., driver age, gender, and vehicle type);
$\bar{\mu}$ is the average percentage (distribution) for certain driver-vehicle characteristic under various circumstances (e.g., different functional roadways, weather conditions, time of day, day of week, year, month, etc.);
μ_0 is the significance threshold for certain driver-vehicle characteristic; and
$\left[\bar{\mu} - \mu_0, \bar{\mu} + \mu_0\right]$ is the confidence interval.

The statistical meaning of the constraint is that there is at least 95% probability that a randomly observed percentage (x) falls within a specified confidence interval ($\left[\bar{\mu} - \mu_0, \bar{\mu} + \mu_0\right]$). The minimum value of μ_0, which satisfies the inequality, is the significance threshold. Based on the empirical practices, the threshold ($\mu_0 = 0.04$) can satisfy the conditions for all three driver-vehicle characteristics. Therefore, if the difference for each characteristic is more than 4 percentage points, it is operationally significant; otherwise it is not.

It needs to be pointed out that the "operational" results do not necessarily match the results of the chi-square test. It would be desirable that both methods generate the same results—that is, the difference is operationally and statistically significant or insignificant. However, there are always some cases that the two results differ. In addition, in order to argue that two distributions are similar, all the driver-vehicle characteristics examined must be consistent concurrently. For example, it is assumed that

male young drivers use the same vehicles as male old drivers. Since the former drives more aggressively than the latter, it is reasonably argued that young male drivers belong to a different driving cohort from old male drivers. Therefore, similarity of two distributions requires the consistency of all characteristics of interest.

In next sections how to use three externally available data to validate the underlying assumptions of quasi-induced exposure will be discussed.

5.1 Validation using vehicle miles of travel data

5.1.1 Introduction

Estimates of VMT can be used to validate the underlying assumptions of quasi-induced exposure. Generally, VMT data are directly collected from household travel surveys conducted by state DOTs or others, or indirectly estimated from traffic counts observed at counting stations. VMT estimates can also be derived from statewide gasoline consumption surveys. In theory, these data will give a true indication of the annual miles of travel for a typical cohort. Comparing the calculated exposure (D2s) from accident data with "true" exposure (VMT) would be one test for the validity of quasi-induced exposure.

In chapter two, several theoretical and operational difficulties in using VMT as an exposure measurement were identified. In terms of validating the assumptions of quasi-induced exposure, the most critical issue is unavailability or limited availability of VMT data disaggregated by desirable variables (driver age, gender, and vehicle type). For instance, VMT data obtainable from the Michigan DOT are disaggregated by trunkline types. So, the fundamentally useful information is the statewide VMT. In some cases state VMT data are calculated based on traffic counts and can be disaggregated by

functionally classified roadway types. Other than that, VMT data disaggregated by driver characteristics are difficult to obtain. Fortunately, the dilemma can be partially resolved with the aid of other data sources such as the National Household Travel Survey (NHTS). A typical NHTS report contains detailed information on a respondent's household, personal characteristics, annual vehicle miles traveled, and so on. Based on census statistics of respondents with the same characteristics (e.g., gender, age group) at the state level, it is possible to calculate the average annual VMT (AVMT) by gender and age group. However, the NHTS data are not usable to compute VMT for different vehicle types. Thus, driver-vehicle characteristics examined in this section only consist of driver age and gender.

Once average annual VMT by gender and age, serving as "true" exposure, are indirectly calculated from NHTS, they can be compared with D2s derived from accident data. Note that accident data used in this section are obtained from the Highway Safety Information System (HSIS). HSIS is "a multistate database that contains crash, roadway inventory, and traffic volume data for a select group of states" (HSIS website). The most obvious advantage of using HSIS data is that they are available to the general public. Currently, there are nine participating states in HSIS, including California, Minnesota, Illinois, North Carolina, Maine, Ohio, Michigan, Utah, and Washington. The accident data for most of the states contain sufficient and necessary information to satisfy the data requirements for using the quasi-induced exposure method.

The following sections include the discussion of the methodology to compare VMT to D2s, the description of the HSIS data, and then the manipulation of the HSIS data, followed by introduction of NHTS, the development of annual VMT estimated for

different driver genders and age groups, the comparisons of VMT and relative exposure estimated by quasi-induced exposure, and conclusions.

5.1.2 Methodology

Using the NHTS data, it is theoretically possible to generate VMT data disaggregated by variables of interest—driver age and driver gender for individual states. Examination of the extent of agreement between relative quasi-induced exposure and "true" exposure helps to reveal if the assumption of quasi-induced exposure will be satisfied by accident data.

VMT data can be directly compared to the relative exposure derived from accident data for D2 characteristics. The steps in this comparison include:

1. Calculate the percentage of "true" exposure for each state based on annual VMT data from the NHTS, disaggregated by age and gender;
2. Calculate the relative quasi-induced exposure stratified by the same two D2 characteristics (age and gender), for each state;
3. Compute the percentage distributions of age and gender based on step 2; and
4. Compare the results from step 1 with step 3 to validate D2s as a measure of exposure.

If D2 distributions disaggregated by age and gender are consistent with those based on VMT data, it indicates that the underlying assumptions of quasi-induced exposure are validated for the given data stratification (e.g., the state level).

5.1.3 Highway safety information system (HSIS)

In order to improve the consistency of data across multiple states, two generic variable tables are developed for the nine states in HSIS. The first table lists the crash-

related variables for each state, and the second table contains the roadway-related variables. The basic crash-related variables are divided into three separate subfiles for ease of computer handling: accidents, vehicles, and occupants. Each accident in the accident file is identified by a unique case number and has only one record (observation) per crash. The vehicle file can have multiple records per crash, depending on the number of vehicles involved, and all vehicles in a given crash are also identified by the (same) unique case number. Similarly, the occupant file has one or more occupants per vehicle with all occupant records being identified with the unique case number. The accident, vehicle, and occupant files can be linked together using the unique case number. The roadway inventory variables describe a section of roadway and are available in a variety of different files for different states. According to the nature of the HSIS data structure, only the accident and vehicle files in the crash-related table are chosen. They are sufficient for satisfying the data need in validating the underlying assumptions of quasi-induced exposure.

Note that not all HSIS states are chosen for study. Whether a state is chosen depends on the availability of a detailed data guidebook (that is, the accident data coding menu) for each state. As listed in the HSIS website, guidebooks are only available for California, Maine, Minnesota, and Utah. Although Minnesota is on the list, it is not chosen due to its mismatched coding menu: variables in the accident data do not match with those in the guidebook. While Michigan does not have a guidebook, the coding menu is available from the Michigan DOT. As a result, accident data from California, Maine, Michigan and Utah will be utilized to develop comparisons.

The HSIS database contains only those crashes that occur on state-maintained roads. The reason is that the locations of these crashes are assumed to be more accurate than the crashes occurring on the local streets and roads for a given state. In order that VMT data are comparable with D2 data calculated from HSIS, only those accidents occurring on corresponding parts of the system will be selected.

5.1.4 Manipulation of HSIS data

HSIS vehicle and accident data are contained in two separate files. The accident file contains basic information on the nature of the accident, such as accident type, location, time, environment, and number of vehicles involved. The vehicle file has information on the characteristics of the driver(s) and vehicle(s) in the crash, such as driver age, driver gender, and vehicle type. As mentioned earlier, there is one unique case number in the accident file, while there are several identical case numbers in the vehicle file depending on the number of vehicles involved in a crash. Linkage of these two files is necessary to develop quasi-induced exposure estimations.

A problem arises in that different drivers in the same accident are coded as different data records in the vehicle file. All driver-vehicle information in the same accident, listed as individual records, must be reformatted as a single data record. The resultant data are then handled by an available statistical software package (i.e., SPSS).

The combined accident records (containing accident and involved driver-vehicle information) are not necessarily coded in an error-free manner. Simple descriptive statistics reveal that some accidents are missing information for the variables of interest (e.g., driver age), or have unreasonable values (e.g., other than male and female for gender). Therefore, the systematic rules developed in an earlier chapter to prepare

Michigan accident data must be applied here in order to make HSIS data usable. In the event of assigning fault, the developed responsibility assignment scheme can also be applied here, considering the similar coding for hazardous actions and violations between Michigan data and other states.

The objectives of the preliminary manipulation of HSIS data are to link the vehicle and accident information, reformat all the driver-vehicles in the same accident as a single record, "clean" the erroneous or miscoded accident data, and, finally, assign responsibility for the accident to each driver-vehicle combination.

With Utah data chosen as an example, the following section will cover how the HSIS accident data are manipulated in a systematic manner. The values shown in the tables below are adapted from the original Utah accident coding as noted by HSIS. In table 5.1.1, one accident case is coded as one row (record); while in table 5.1.2, one driver-vehicle combination is coded as one row (record). Both of the tables contain an identical variable which can be used to combine them into one record—accident case number (*caseno*).

Table 5.1.1. Sample data from Utah accident file (2000)

caseno	accident variable list								
	rodwycls	*agency*	*weekday*	*month*	...	*hour*	*acctype*	*county*	*light*
200000002	08	00207	7	1	...	10	4	03	2
200000003	06	0000F	1	1	...	21	2	03	5
200000004	00	01000	2	1	...	12	2	19	2
200000005	08	00100	2	1	...	15	7	01	2
200000006	00	01600	2	1	...	11	10	31	2

Table 5.1.2. Sample data from Utah vehicle file (2000)

caseno	driver-vehicle variable list						
	Vehno	*vehtype*	*contrib1*	...	*contrib2*	*drv_sex*	*drv_age*
200000002	1	2	1	...	0	2	50
200000003	1	32	7	...	14	1	42
200000003	2	5	3	...	15	1	20
200000003	3	2	7	...	14	2	29
200000004	1	20	51	...	0	1	51
200000005	1	33	30	...	50	1	40
200000006	1	40	1	...	0	1	50
200000006	2	6	42	...	0	1	67

The diagram in figure 5.1.1 shows the whole process for manipulating the Utah accident data, which consists of several "operation" modules.

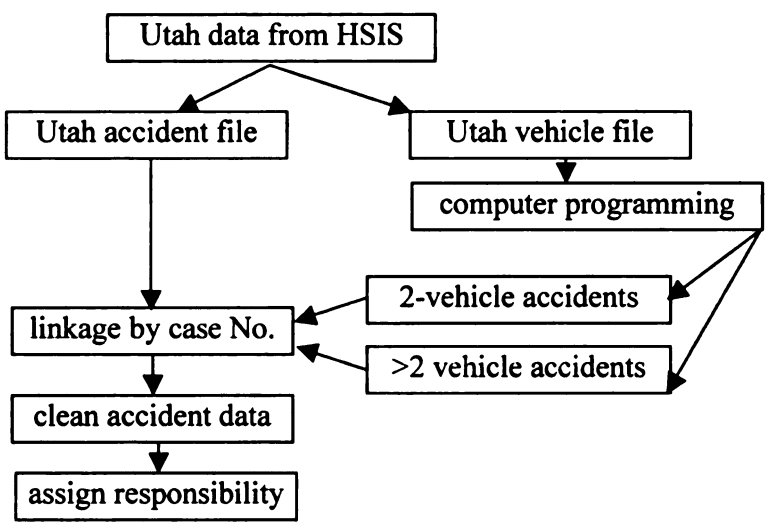

Figure 5.1.1. Flow chart of manipulating Utah accident data

In the flow chart, the "computer programming" module is specially designed to eliminate one-vehicle accidents in the vehicle file and then split the vehicle file into two subfiles: two-vehicle accidents and three-or-more-vehicle accidents. Each subfile is manipulated in a way that all driver-vehicle information for the same accident is rearranged in the same row (record). This was done using Microsoft C++ (programming code can be found in Appendix A). The function "linkage by case No." is to link the

accident file to the vehicle file. The systematic rules developed to select useful and dependable accident data and assign responsibility for accident causation in chapter four are contained in the "clean accident data" and "assign responsibility" modules, respectively. After this is done, a typical HSIS accident data record will take the form shown in table 5.1.3, with responsibility clearly assigned for each vehicle.

Table 5.1.3. Desirable data format for a typical HSIS data record

caseno	accident			vehicle 1			...	vehicle n		
	hour	...	county	age	...	sex	...	age	...	sex
20004	15	...	19	29	...	2	...	51	...	1

At this point, the work with three-or-more-vehicle accidents will not be discussed. The details of how to use three-or-more-vehicle accidents to validate the underlying assumptions of quasi-induced exposure will be covered in a later chapter. The following section is about how to use NHTS data to estimate VMT disaggregated by driver age and gender.

5.1.5 National household travel survey (NHTS)

The National Household Travel Survey (NHTS) is the nation's inventory of daily and long-distance travel by telephone interviews. The NHTS is a survey of the civilian, non-institutionalized population of the United States. The survey includes demographic characteristics of households, people, vehicles, and detailed information on daily and longer-distance travel for all purposes by all modes. NHTS survey data are collected from a sample of U.S. households and expanded to provide national estimates of trips and miles by travel mode, trip purpose, and a host of household attributes. These data provide planners and decision makers with up-to-date information to assist them in effectively improving the mobility, safety, and security of the nation's transportation systems. It

should be pointed out that a relatively low response rate (41%) was achieved in 2001 NHTS survey, which might result in the non-response bias, and coverage bias could have been introduced by excluding the households without land-line telephone. The margin of error in the survey results are assessed to be up to 10%.

In the research context here, 2001 NHTS data are utilized to derive average AVMT information for a specified driver group on a state basis. The 2001 NHTS data are composed of four major files: household, person, vehicle, and day trip. The person file is of the most interest, which contains key variables that can be used as the basis for calculating disaggregated AVMT.

Ideally, gender/age group VMT can be estimated for several different states for which accident data are readily available (California, Maine, Michigan and Utah). However, in the 2001 NHTS data, Maine is one of the states where household location is not categorized individually but combined with other states having population less then one million. In this context, respondents from Maine cannot be separated from other states. Consequently, comparisons of VMT with D2 estimates are limited to three states: California, Michigan, and Utah.

5.1.6 Average annual vehicle miles traveled

Four key variables from the person file have been identified to be the most valuable and relevant: state-household location, respondent age, respondent gender, and miles respondent drove during the last 12 months. It has been found that some cases are not useful because respondents refused to answer what they were asked. Appropriate actions are necessary to screen out the unclear data. And then, based on the state-household location, separate datasets are created for each state. In each individual state,

data are further disaggregated by D2 characteristics (driver age and gender in this case), with which annual VMT is averaged over all the respondents. For an example of driver gender, average annual VMT for male drivers is the total annual VMT by all the male respondents divided by the number of male respondents; the same concept is adopted for female drivers and other characteristics. The following flow chart shows the process graphically (figure 5.1.2).

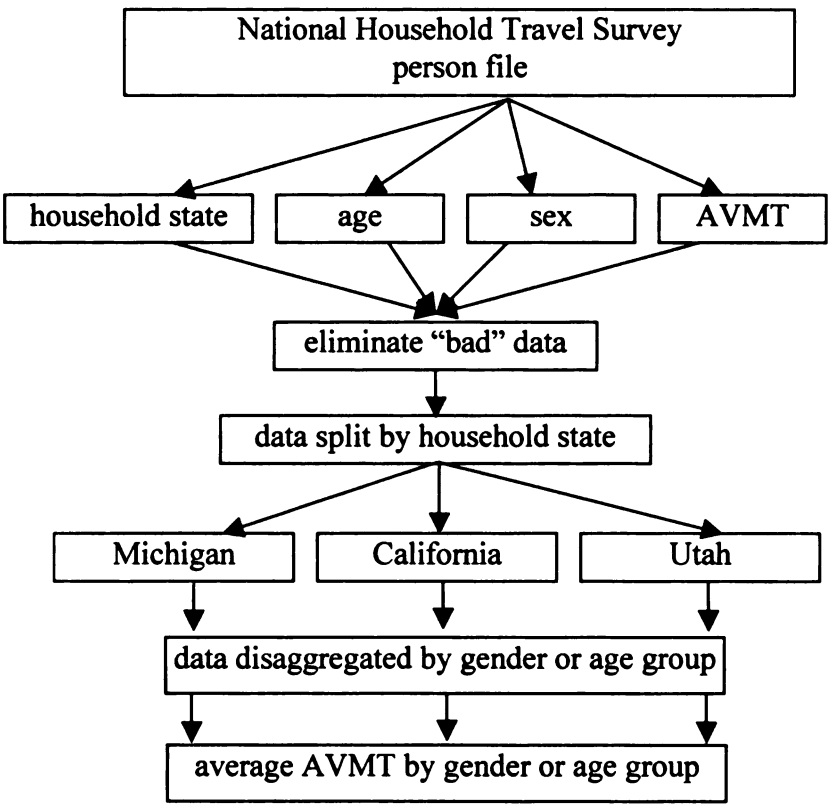

Figure 5.1.2. The process to derive average AVMT

In the above process, three assumptions are implicitly made to estimate average annual VMT:

1. the survey respondents primarily travel within his/her household state;
2. the survey respondents are a random sample of the driving population on the road; and

3. the respondent can accurately estimates AVMT and each sex and each age group makes the same average errors.

Practically speaking, the first assumption is likely to be true. For the second assumption, since NHTS is a national survey, data collected from a nationally representative sample of households should be reliable to derive statistically reliably travel estimates at the national and state levels.

VMT percentages

Mathematically, the average annual VMT disaggregated by driver characteristics (e.g., female AVMT) multiplied by the corresponding population yields total annual VMT (e.g., total female AVMT). The population information comes from state census data which are available to the public.

The following series of tables is the presentation of total annual VMT estimates and percentages (the distributions) disaggregated by driver gender and age group for Michigan, Utah, and California in 2001.

Table 5.1.4. Total AVMT and percentages disaggregated by age group (MI)

age group	average AVMT (mi)	census	total AVMT (million)	percentage (%)
15-19	6,526	659,423	4,303	3.9
20-24	14,646	642,409	9,409	8.5
25-34	16,118	1,335,227	21,521	19.4
35-44	17,858	1,590,477	28,403	25.6
45-54	17,207	1,429,001	24,589	22.2
55-64	14,736	895,921	13,202	11.9
65+	8,083	1,156,882	9,351	8.4
total	13,956*	7,709,340	110,778	100.0

*This number is averaged AVMT for driver age or gender; same as below.

Table 5.1.5. Total AVMT and percentages disaggregated by gender (MI)

gender (age>14)	average AVMT (mi)	census	total AVMT (million)	percentage (%)
male	17,607	3,713,910	65,391	59.7
female	11,058	3,995,430	44,182	40.3
total	14,333*	7,709,340	109,573	100.0

As highlighted in tables 5.1.4 and 5.1.5, the total AVMT by age group and gender are fairly close. The estimation error, defined as the difference between two AVMT totals divided by the arithmetic mean, is approximately 1.1%, which is small enough to be negligible. And, a relatively small estimation error (2.6%) is also found between the average AVMT among different age groups (13,956 miles/year) and genders (14,333 miles/year).

Table 5.1.6. Total AVMT and percentages disaggregated by age group (UT)

age group	average AVMT (mi)	census	total AVMT (million)	percentage (%)
15-19	5,344	190,821	1,020	4.6
20-24	12,488	229,278	2,863	12.9
25-34	15,134	339,667	5,141	23.2
35-44	19,536	287,678	5,620	25.1
45-54	14,897	246,105	3,666	16.5
55-64	17,962	147,017	2,641	11.9
65+	6,639	186,648	1,239	5.6
total	13,143*	1,627,214	22,190	100.0

Table 5.1.7. Total AVMT and percentages disaggregated by gender (UT)

gender (age>14)	average AVMT (mi)	census	total AVMT (million)	percentage (%)
male	16,866	804,152	13,563	61.2
female	10,428	823,062	8,583	38.8
total	13,649*	1,627,214	22,146	100.0

Compared to Michigan data, the estimation error between two AVMT totals in UT is much smaller, less than 0.3%; but, the estimation error for average AVMT among

all age groups (13,143 miles/year) or genders (13,649 miles/year) is approximately 3.8%. These errors can be ignored as well. Note that the calculated estimation error is intended to show the internal consistency between the average vehicle miles traveled data by driver gender and driver age. For UT and MI data, the AVMT data obtained from different approaches are operationally close.

For CA, the AVMT totals from different age groups and genders are fairly close—the estimation error is 0.6% (table 5.1.8). And, the average AVMT among different age groups (12,438 miles/year) and genders (13,039 miles/year) are also consistent, approximately 1% (table 5.1.9).

Table 5.1.8. Total AVMT and percentages disaggregated by age group (CA)

age group	average AVMT (mi)	census	total AVMT (million)	percentage (%)
15-19	7,049	2,368,932	16,699	5.1
20-24	15,288	2,279,533	34,850	10.6
25-34	14,187	5,028,008	71,332	21.7
35-44	15,505	5,335,260	82,723	25.1
45-54	14,049	4,354,118	61,171	18.6
55-64	13,088	2,665,045	34,880	10.6
65+	7,900	3,481,862	27,507	8.3
total	12,438*	25,512,758	329,161	100.0

Table 5.1.9. Total AVMT and percentages disaggregated by gender (CA)

gender (age>14)	average AVMT (mi)	census	total AVMT (million)	percentage (%)
male	15,603	12,488,228	194,854	58.8
female	10,475	13,024,530	136,432	41.2
total	13,039*	25,512,758	331,286	100.0

In general, for each individual state, the total or average AVMT for different driver genders agree with different driver age groups, which suggests the reliability and integrity of VMT and census data.

It needs to be pointed out that the VMT distributions are calculated based on the census data. Using census population to replace driving population (i.e., traffic volume data disaggregated by three key characteristics) will likely cause some age groups to be overrepresented while others are underrepresented. For example, young drivers are assumed to drive more frequently than their proportion in census distribution while old drivers tend to drive less. Consequently, young drivers will be underrepresented in the driving population, while older drivers are overrepresented. So, in order to be technically accurate and comparable with D2 distributions, VMT distributions should be computed based on driving population data. At this point, due to the unavailability of traffic volume data, it is assumed that driving population is characteristic of the census population at each state.

5.1.7 D2 distributions

Calculating the D2 distributions is straightforward and based on accident data. For Michigan, accident data are readily available from the Michigan DOT. California and Utah accident data are obtained from the Highway Safety Information System (HSIS). However, California and Utah data are not complete as only the accidents occurring on state-maintained roads are filed with HSIS. Examination of the accident guidebook for each state reveals that accidents recorded in the HSIS system occur on mainline roadway segments and related areas (e.g., on-ramp, off-ramp, rest areas). It has been known that under a given road and traffic setting, there might be a limited number of accident types. Therefore, selection of accidents on state-maintained routes could be a case-specific process.

For Michigan data (tables 5.1.10 and 5.1.11), D2 distributions are developed for three categories: all accidents (total), trunkline accidents, and non-trunkline accidents. Trunkline accidents are distinguished from non-trunkline accidents according to the variable "control section." The control section is a five-digit code that identifies the portion of the trunkline system: the first two digits are the MDOT county number and the last three digits are the unique trunkline segment number. If the five-digit code is composed of five zeros, the route is non-trunkline. So, the trunkline can be distinguished from the non-trunkline based on the control section values.

Trunkline and non-trunkline accidents are compared to see if the D2 percentage distributions are different. Based on the Michigan experience, one can generally estimate/predict whether D2 percentage distributions in trunkline and non-trunkline accidents follow the same pattern, the results of which can be reasonably presumed for two other states: California and Utah. It has been implicitly assumed that the definitions of the truckline are similar among the states of interest.

Table 5.1.10. D2 gender distributions for Michigan data (2001)

gender	Michigan total		Michigan trunkline		Michigan non-trunkline	
	N	%	N	%	N	%
male	75512	53.7	24301	55.0	51211	53.1
female	65137	46.3	19899	45.0	45238	46.9
total	140649	100.0	44200	100.0	96449	100.0

According to table 5.1.10, there is a 1.9 percentage point difference of D2 male drivers between trunkline and non-trunkline accidents. Since this difference is operationally small, it is argued that the D2 gender distribution on trunklines is not significantly different from that on non-trunklines. Examination of the differences for age groups in table 5.1.11 also reveals the same phenomenon.

Table 5.1.11. D2 age distributions for Michigan accident data (2001)

age	Michigan total		Michigan trunkline		Michigan non-trunkline	
	N	%	N	%	N	%
15-19	14166	10.1	4109	9.3	10057	10.4
20-24	16929	12.0	5461	12.4	11468	11.9
25-34	30299	21.5	9572	21.7	20727	21.5
35-44	31357	22.3	9972	22.4	21385	22.2
45-54	24974	17.8	7940	18.0	17034	17.6
55-64	12581	8.9	3991	9.0	8590	8.9
65+	10343	7.4	3155	7.0	7188	7.5
total	140649	100.0	44200	100.0	96449	100.0

Based on the Michigan experience, it is reasonable to predict that for CA and UT accident data, the D2 driver-vehicle combinations involved in the trunkline-related accidents are representative of those in all accidents. Tables 5.1.12 and 5.1.13 show the D2 distributions of gender and age on trunkline roads for CA and UT.

Table 5.1.12. D2 gender distributions for CA and UT data (2001)

gender	Utah trunkline		California trunkline	
	N	%	N	%
male	13819	57.6	12823	64.1
female	10179	42 4	7178	35.9
total	23998	100.0	20001	100.0

Table 5.1.13. D2 age distributions for CA and UT data (2001)

age	Utah trunkline		California trunkline	
	N	%	N	%
15-19	3415	14.2	1197	6.0
20-24	4290	17.9	2174	10.9
25-34	5450	22.7	4851	24.3
35-44	4589	19.1	5110	25.5
45-54	3252	13.6	3720	18.6
55-64	1704	7.1	1818	9.0
65+	1298	5.4	1131	5.7
total	23998	100.0	20001	100.0

As shown in tables 5.1.12 and 5.1.13, the total number of accidents for CA and UT is fairly small, compared to MI (140,649 cases in total). Based on the CA guidebook,

there are about 500,000 total accidents that occur each year, and approximately 160,000 occur on the state-maintained highway system. However, for CA data, significant percentage of accidents is filtered out and approximately 17.5% is usable by quasi-induced exposure.

In table 5.1.12, the percentage of D2 male drivers in CA is much larger (6.5 percentage points) than UT. Operationally, the difference is significant. In table 5.1.13, there is considerable difference (5-8 percentage points) for the several age groups when D2 age distributions are compared across states.

Although the differences in D2 distributions among states are recognized, it is more important to know whether the D2 distributions for age and gender are significantly different from those based on VMT for each state.

5.1.8 Comparison of D2 and VMT estimations

Comparisons between D2 and VMT exposure estimates are made for Michigan, California, and Utah and the results are evaluated with operational and statistical methods.

Tables 5.1.14 and 5.1.15 are created from information shown in tables 5.1.4 through 5.1.13. The percentage difference is defined as the VMT percentage minus the corresponding D2 percentage. A positive sign indicates that the VMT percentage is larger than the D2 percentage; a negative sign shows the opposite.

Table 5.1.14. Gender percentage difference (VMT-D2) among different states

gender	Michigan trunkline	Utah trunkline	California trunkline
male	4.7	3.6	-5.1
female	-4.7	-3.6	5.1

Table 5.1.15. Age percentage difference (VMT-D2) among different states

age	Michigan trunkline	Utah trunkline	California trunkline
15-19	-6.5	-9.6	-0.9
20-24	-3.4	-5.0	-0.3
25-34	-2.1	0.5	-2.6
35-44	3.4	6.2	-0.4
45-54	4.6	2.9	0.0
55-64	3.0	4.8	1.6
65+	0.9	0.2	2.6

The shaded cells in tables 5.1.14 and 5.1.15 are indicative of where the percentage difference is operationally significant. D2 gender distributions for both Michigan and California data show a disagreement with those based on VMT data; two age groups (15-19 and 45-54) in Michigan data and four age groups in Utah data show that percentage differences are considerable. Although there is consistency in the gender distribution for Utah data and in the age distribution for California data, D2 distributions are claimed to be different from VMT disaggregated by age and gender for the three states discussed.

The percentage differences by gender and age for each state are also tested using chi-square test to see if they are statistically significant. The chi-square statistic and *p*-value for each comparison are presented in table 5.1.16. Note that for driver gender and vehicle type, the critical value $\chi_{0.05}(1)$ is equal to 3.84; for driver age, the critical value $\chi_{0.05}(2)$ is equal to 6.00.

Table 5.1.16. Summary of chi-square statistics and *p*-values

	Michigan trunkline	Utah trunkline	California trunkline
gender	286 (0.00*)	64 (0.00)	219 (0.00)
age	2969 (0.00)	1851 (0.00)	300 (0.00)

**p*-values are shown in the parenthesis.

As shown in table 5.1.16, the *p*-values for gender and age given by chi-square test are constantly equal to zero. Since *p*-values are less than 0.05, the null hypothesis that

there is no significant difference of distributions between the VMT and D2s is rejected. Statistically speaking, the underlying assumptions of quasi-induced exposure are not supported for Michigan, California, and Utah data.

5.1.9 Conclusion

It has been shown that quasi-induced exposure estimates are operationally and statistically different from VMT estimates for Michigan, Utah, and California. The VMT estimates are derived from NHTS in combination with census data. The NHTS is a national survey and collects data from a nationally representative sample of households at the national and state levels. Therefore, the travel estimates should be statistically reliable and representative of individual states. However, the VMT distributions are calculated based on the census data instead of the driving population data (i.e., traffic volume data). The data are actually indicative of the distribution of census-population for each state. The assumption that driving population is characteristic of the census population at each state must not hold.

As indicated earlier, the survey results are subject to up to 10% margin of error. Since the VMT disaggregated by driver age and gender for each individual state, supposedly as an exposure "truth," can not be justified, the validity of D2 data can not be determined. Therefore, it is inconclusive whether quasi-induced exposure is a legitimate approach to measure the relative exposure on the state-maintained routes as the accident data represent. Further research efforts are necessary to derive more dependable and accurate VMT estimates.

The following section will cover how to use safety-belt use data to develop a similar analysis.

5.2 Validation using safety-belt use data

5.2.1 Introduction

Another source of data for validation of the quasi-induced exposure technique is a series of direct observation surveys conducted by the University of Michigan Transportation Research Institute (UMTRI) for the purpose of studying safety-belt use among motor vehicle occupants in Michigan. The data were collected as part of an evaluation of the effectiveness of Michigan's mandatory safety-belt use law/policy and enforcement. The safety-belt data can be viewed as scientifically sampled traffic counts disaggregated by driver and passenger gender, age, day of week, time of day, and observation site type (intersection or freeway off-ramp). The safety-belt data include the information on front-outboard vehicle occupants (driver and front-seat passenger) in eligible commercial and noncommercial vehicles (passenger car, vans/minivans, sport-utility vehicles, and pickup trucks). In terms of the research here, driver and vehicle information in these safety-belt data are the potentially useful elements.

The safety-belt data are systematically observed and collected following a strict sampling plan. The sampling design plan used by UMTRI follows federal guidelines for state observational surveys for safety-belt use as developed by the National Highway Traffic Safety Administration (NHTSA, 1992 and 1998). A systematic sampling procedure is employed which has been designed in a manner such that selection of observations sites is random. Since the NHTSA guidelines are strictly followed in the data collection, the precision of the observed data is expected to be less than 5% relative error. The resultant data represent a random sample at the state level. Therefore, at this

level, the safety-belt data serve as a measure of "true" exposure of the statewide driving population.

The underlying assumption of quasi-induced exposure is that D2s in two-vehicle accidents constitute a random sample of the driving population on the road at the time of accident occurrence. Comparison of the distributions for driver-vehicle characteristics between safety-belt data and non-responsible driver-vehicle combinations (D2s) will help to determine if D2s are really a random sample of the driving population on the road.

5.2.2 Methodology

In the attempt to validate the assumption of quasi-induced exposure, D2 data must be comparable to the observed safety-belt use data, i.e., in terms of roadway type, weather, day of time, and vehicle classification. In the comparison, both the safety-belt and D2 data contain three basic characteristics of the driving population that can be compared: driver age, driver gender, and vehicle type.

D2s are examined at the state level, where the sampling plan for the safety-belt data is developed. Further comparison will be conducted at the stratum level (multi-county) as defined in the UMTRI study and at the county level to see if D2 assumption can be validated at more disaggregated levels. The following is the methodology used here:

1. Calculate the distribution for of characteristics of the driving population at the state level, using the UMTRI data;
2. Determine the distribution for D2 characteristics at the state level;
3. Compare the results from steps 1 and 2;
4. Disaggregate the safety-belt and D2 data into several strata;

5. Repeat steps 2 and 3;
6. Further disaggregate the safety-belt and D2 data into counties; and
7. Repeat steps 2 and 3.

If similarity exists between the characteristic distributions of D2 and safety-belt data, the attempt to validate the underlying assumption of quasi-induced exposure is successful at the level represented by the data.

5.2.3 Preparation of safety-belt data

Since the safety-belt data are initially targeted for another purpose, the presentation of data is not immediately usable for purposes here. The safety-belt data are organized in three files: a list of survey sites, individual site descriptions, and observation data from each site.

The list of survey sites (see table 5.2.1 for example) contains a complete listing of sites chosen for field data investigation. It is a descriptive table including survey site number, the county where the site is located, the specific description of site location (direction observed, name of intersection or freeway off-ramp), the site type (intersection or freeway off-ramp), and the stratum number. It is noted that each observation site belongs to a certain stratum (table 5.2.2) defined in the UMTRI study (Eby et al. 2001).

Table 5.2.1. Sample of survey site list

site number	county	site location	type	stratum number
002	Kalamazoo	NB 34th St. & V. Ave	I	1
077	Ottawa	NBR I-196 & Byron Rd	ER	2

Note that the strata were constructed based on historical safety-belt use rates and VMT for each county. Total VMT by strata are roughly equal and observation sites are randomly assigned (Eby et al. 2001) within each.

Table 5.2.2. Stratification scheme in safety-belt study

stratum	county
1	Ingham, Kalamazoo, Oakland, Washtenaw
2	Allegan, Bay, Eaton, Gr. Traverse, Jackson, Kent, Livingston, Macomb, Ottawa
3	Berrien, Calhoun, Genesee, Lapeer, Lenawee, Monroe, Muskegon, Saginaw, Shiawassee, St. Clair, St. Joseph, Van Buren
4	Wayne

The description file contains the characteristics of the observation site (table 5.2.3), i.e., site number, site type, traffic control device, observation day, weather, and observation start and end times. Most of the variables listed in table 5.2.3 are self-explanatory, e.g., day of week, weather. The variable "interruption" refers to the total number of minutes during which safety-belt data collection is interrupted due to some unexpected situations, such as traffic congestions or extremely adverse weather.

Table 5.2.3. Coding of site description form—1998, 2000

variable name	codes
site number	001 – 168 (observation site number)
site type	1=intersection, 2=freeway
site choice	1=primary, 2=alternate
traffic control device	1=traffic light, 2=stop sign, 3=none, 4=other
Date	MMDD
observer number	1 – 7=observers alone, 8 – 0=observer pairs
day of week	1=Monday, 2=Tuesday, 3=Wednesday, 4=Thursday, 5=Friday, 6=Saturday,7=Sunday
Weather	1=mostly sunny, 2=mostly cloudy, 3=rain, 4=snow
start time	HHMM – military time
end time	HHMM – military time
Interruption	MM – number of minutes
Median	1=yes (median present), 2=no (no median present)

The observation file includes the characteristics of driver and/or passenger (if available) observed at the site: site number, estimated driver age, driver gender, vehicle type, and passenger gender and age (if available). The file information is shown in table 5.2.4.

For the variables listed in table 5.2.4, it is relatively easy to identify the driver and/or the passenger gender and the vehicle type, but is obviously not as accurate for driver and/or passenger age. Based on the defined age ranges, it is especially hard to tell the ages in the upper bound of range 16-29 from those in the lower bound of range 30-59.

Table 5.2.4. Coding of observation data–1998, 2001

variable name	codes
site number	001 – 168 (observation site number)
driver gender	1=male, 2=female, 9=missing
driver age	2=4-15, 3=16-29, 4=30-59, 5=60+, 9=missing
vehicle type	1=passenger car, 2=van, 3=utility, 4=pickup, 9=missing
passenger gender	1=male, 2=female, 8=no passenger, 9=missing
passenger age	1=0-3, 2=4-15, 3=16-29, 4=30-59, 5=60+, 8=No passenger, 9=missing
vehicle number	### - each vehicle within each site is assigned a unique number
commercial vehicle	1=no (vehicle is not commercial) 2=yes (vehicle is commercial)

These data are aggregated to the counties, strata, and state level. Thus, the characteristic distributions of driving populations at the state, stratum, and county levels can be compared with those based on accident data. The comparisons under different aggregations can be further developed for intersections and freeway.

The data in the three existing UMTRI files can be combined into a larger one with several variables including site number, county, site location, site type, observation day,

day of week, observation time, and characteristics of the drivers and passengers (if available). The format of this file is shown in table 5.2.5. In the context of validating the underlying assumptions of quasi-induced exposure, some of variables are irrelevant to the discussion at hand and are eliminated. These include the characteristics of the passenger, median type, and traffic control devices. The final data record contains the following variables: site number, site type, county, driver age, driver gender, and vehicle type.

Table 5.2.5. Coding of resultant data–1998, 2001

variable name	Codes
site number	001 – 168 (observation site number)
site type	1=intersection, 2=freeway
county	list of county name
driver gender	1=male, 2=female, 9=missing
driver age	2=4-15, 3=16-29, 4=30-59, 5=60+, 9=missing
vehicle type	1=passenger car, 2=van, 3=utility, 4=pickup, 9=missing

The following is a sample case in the resultant data file:

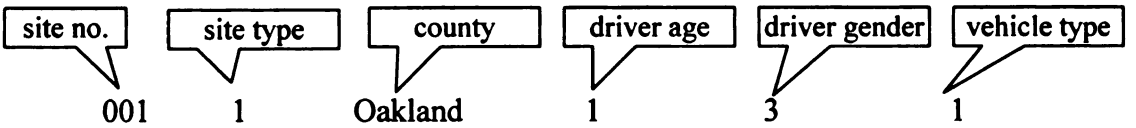

Figure 5.2.1. Sample data case in the resultant data file

5.2.4 Discussion of variables used

The following is the discussion of how accident data were selected to match a variety of circumstances where safety-belt data were observed. In this research context, timeframe, driver age, driver gender, vehicle type, day of week, observation period, site type, and weather are discussed.

Timeframe

Michigan 2001 accident data are selected to match the 2001 safety-belt data obtained from UMTRI. Furthermore, Michigan accidents are also selected to reflect the seasonal observations of safety-belt data: spring (May or June) and fall (August, September, or October), with the aid of variable *month*.

Driver age

In the safety-belt data, driver age is coded as a general range (e.g., 16-29, 30-59), since it was impossible to know the exact age of drivers during on-site observations. As stated, they are prone to error. Errors notwithstanding, the age distributions of D2s are also aggregated into the same ranges.

Driver gender

Driver gender should not be an issue in this exercise, since it is relatively easy to distinguish driver gender in the field survey.

Vehicle type

Vehicle type in the safety-belt data is classified into four categories: passenger car, van, utility, and pickup. More specifically, the "passenger car" category includes passenger vehicles and station wagons; the "van" consists of full-size and mini vans; the "utility" encompasses sports utility vehicles; and the "pickup" is composed of large and small pickup trucks. Note that other vehicles were not observed and thus not recorded. In the Michigan accident data (2001), vehicle type has 30 categories (table 5.2.6).

Table 5.2.6. Vehicle type in Michigan accident data (2001)

00 uncoded & error	11 miscellaneous commercial	22 single, hazardous
01 passenger car & station wagon	12 combo unit	23 single, tank
02 van, motorhome	13 combo, hazardous	24 single, passenger
03 pickup	14 combo, tank	25 single, tank & haz
04 truck under 10,000 lb	15 combo, passenger	26 under 26k, hazardous
05 cycle	16 combo, double or triple	27 under 26k, passenger
06 moped	17 combo, tank & hazardous	28 under 26k, tank & haz
07 go-cart	18 combo, double or triple tank	29 others
08 snowmobile	19 combo, double or triple hazardous	
09 ORV, ATV	20 combo, double or triple tank hazardous	
10 other non-commercial	21 single over 26k	

Based on the values and descriptions listed in table 5.2.6, values do not match the categories in safety-belt data precisely. Category 01 approximately matches the category "passenger car" in the safety-belt data; and category 03 reasonably matches "pickup" in the safety-belt data. The category "utility" in the safety-belt data (13.4% in total) is captured in categories 01 and 02 in the accident data; the category "van" (14.3% in total) can't match the category 02 in the accident data because of the motorhome vehicles. That means for the rest of values, there is no category (or combination) in accident data that can match the categories "utility" and "van" in the safety-belt data. With this classification scheme, vehicle type is divided into two classifications: passenger car and pickup truck. The rest of the data will not be used.

Other variables

There are also some other variables to be considered in the safety-belt data, which might affect the selection of accident data. Table 5.2.7 shows the descriptive statistics for 168 observation sites for the 2001 safety-belt use study.

Table 5.2.7. Descriptive statistics for the 168 observation sites (2001)

day of week	observation period	site type
Monday 15.5%	7-9 AM 11.9%	intersection 76%
Tuesday 14.3%	9-11 AM 20.2%	freeway (off-ramp) 24%
Wednesday 11.9%	11-1 PM 15.5%	
Thursday 17.8%	1-3 PM 20.8%	
Friday 14.3%	3-5 PM 19.7%	
Saturday 16.1%	5-7 PM 11.9%	
Sunday 13.1%		

It can be seen that the observations were fairly evenly distributed over the day of week, while the observation periods are limited to the daytime hours. In the corresponding accident data, only accidents occurring during the daytime hours (7AM to 7PM) will be selected with the variable *hour*.

Furthermore, in table 5.2.7 the safety-belt data were collected at intersections and freeway off-ramps. However, the safety-belt data were observed in a way to represent the statewide population; correspondingly, the accidents must be statewide numbers instead of those occurring at intersections and freeway off-ramps.

In brief, after the systematic rules to clean accident data are applied on Michigan data, appropriate selections are undertaken to make them comparable with the safety-belt data in terms of timeframe (2001, fall and spring), observation period (7AM-7PM), and site type (all locations). After the responsibility is assigned, D2 distributions for driver gender, age, and vehicle type are calculated.

5.2.5 Comparison of safety-belt and D2 data

This section includes the comparison of safety-belt with D2 data disaggregated by the three main driver-vehicle characteristics. As indicated in the methodology, the comparison will be conducted at three levels: state, stratum, and county. Furthermore, an add-on comparison is developed at intersections (state and stratum levels only), to see if

D2 distributions are able to fit the safety-belt data. Note that for all the tables listed in this section, the column "difference" stands for the percentage difference between safety-belt and D2 data in terms of different driver-vehicle characteristics. Positive values indicate that the percentages for the safety-belt data are larger than those for D2 data.

State level

Table 5.2.8 shows the comparisons of different driver-vehicle characteristics statewide. In table 5.2.8, the percentage of male drivers observed in the safety-belt data is 2.9 points higher than D2s; the age group (30-59) has the largest difference (5.0 percentage points); and the percentage difference of passenger car is fairly large, 7.3 percentage points. Operationally, these differences are insignificant for driver gender but significant for driver age and vehicle type (larger than 4 percentage points). In addition to the operational analysis, the differences between D2s and safety-belt use data are also examined with chi-square test. Table 5.2.9 shows the statistics and *p*-values for three key driver-vehicle characteristics.

Table 5.2.8. Safety-belt versus D2 distributions for statewide

characteristics		safety-belt		accident data (D2)		difference (%)
		N	%	N	%	
gender	male	6975	58.8	24733	55.9	2.9
	female	4887	41.2	19494	44.1	-2.9
age	16-29	3331	28.1	13642	30.8	-2.8
	30-59	7371	62.2	25296	57.2	5.0
	60+	1156	9.7	5289	12.0	-2.2
vehicle type	passenger car	6394	74.5	31557	81.9	-7.3
	pickup truck	2184	25.5	6996	18.1	7.3

Table 5.2.9. Summary of chi-square statistics and *p*-values (safety-belt versus D2 data)

gender	age	vehicle type
32 (0.00*)	102 (0.00)	239 (0.00)

**p*-values are shown in the parenthesis.

In table 5.2.9, the *p*-values yielded by the test are consistently equal to zero ($p<0.05$). These results suggest that the differences are statistically significant for these three characteristics.

A similar analysis is conducted for the accidents occurring at intersections. These accidents are a subset of statewide accidents, which can be largely pinpointed by examining the area type of each accident record (variable *areacode*). Accidents of interest mainly consist of those with *areacode* "within intersections," "driveway within 150 ft of intersection," and "other intersection related." In the safety-belt data, the observations at intersections can be conveniently identified with the built-in variable (site type, table 5.2.5). The comparison of the safety-belt and D2 data is presented in table 5.2.10.

Table 5.2.10. Safety-belt versus D2 distributions statewide—intersection

characteristics		safety-belt		D2		difference (%)
		N	%	N	%	
gender	male	4849	57.2	12218	53.2	4.0
	female	3635	42.8	10792	46.7	-4.0
age	16-29	2299	27.1	7152	31.1	-4.0
	30-59	5298	62.5	13057	56.7	5.8
	60+	885	10.4	2802	12.2	-1.8
vehicle type	passenger car	4570	74.4	16447	81.9	-7.5
	pickup truck	1572	25.6	3637	18.1	7.5

Shown in table 5.2.10, the percentage of male drivers in the safety-belt data is approximately 4.0 percentage points greater than that in D2 data, right around the significance threshold; among the three age groups, the middle-aged group has the biggest difference: 5.8 points; compared to the differences in driver gender and age, the difference in vehicle type is larger, 7.5 percentage points. This result is expected, since it

has been found that there is a disagreement of distributions for driver age and vehicle type between the safety-belt and D2 data at the statewide level.

Given the fact that the sampling plan for the safety-belt data collection was implemented at the statewide level, the characteristics of the driver-vehicle combinations in the safety-belt data are supposed to represent the real driving population on the road. However, data errors might be introduced to the driver age, since it is relatively difficult to identify accurate driver age during the field observations, especially for the 30-59 age group. For the vehicle type distribution, there might be some errors due to the mismatched classification. As discussed, the passenger cars in accident data might include some sport utility vehicles, which consequently overestimate the percentage (81.9%).

So, driver gender seems to be the most reliable characteristic. The comparison results demonstrate a fairly good agreement of driver gender distributions between the safety-belt and D2 data both at the statewide and intersections levels, from the operational perspective. In this confined context, it is safe to argue that the fundamental assumption of quasi-induced exposure, that D2s constitute a random sample of driving population (at least, in terms of driver gender) on the road, is validated at the state level. That is to say, quasi-induced exposure is a generally good approach at the state level, although more validation is needed for driver age and vehicle type.

In the following section, the research continues to explore to see if the validation can be achieved at the stratum level. Since the data are more disaggregated at the stratum level, the differences of distributions for three characteristics are expected to be greater.

Stratum level

Safety-belt data can be also aggregated to the stratum level according to the definition in table 5.2.2. Analogous to the analysis at the state level, the percentages of driver-vehicle characteristics for D2s will be compared with those from safety-belt data for each stratum. Discussion is developed separately for the stratum and intersections.

The data are organized on a stratum basis with percentages disaggregated by different driver-vehicle characteristics and displayed in the same table. The following four tables (5.2.11, 5.2.12, 5.2.13, and 5.2.14) show the comparisons for each stratum.

Table 5.2.11. Safety-belt versus D2 distributions for stratum 1

characteristics		safety-belt		D2		difference (%)
		N	%	N	%	
gender	male	1970	56.5	6685	50.5	6.0
	female	1518	43.5	6547	49.5	-6.0
age	16-29	1073	30.8	4069	30.8	0.0
	30-59	2126	61.0	7797	58.9	2.0
	60+	289	8.3	1366	10.3	-2.0
vehicle type	passenger car	1855	74.1	9690	83.8	-9.6
	pickup truck	647	25.9	1878	16.2	9.6

Table 5.2.12. Safety-belt versus D2 distributions for stratum 2

characteristics		safety-belt		D2		difference (%)
		N	%	N	%	
gender	male	1328	61.1	6652	52.7	8.4
	female	844	38.9	5966	47.3	-8.4
age	16-29	576	26.5	3970	31.5	-4.9
	30-59	1367	62.9	7123	56.5	6.5
	60+	229	10.5	1525	12.1	-1.5
vehicle type	passenger car	1092	71.3	8747	79.8	-8.5
	pickup truck	440	28.7	2221	20.2	8.5

Table 5.2.13. Safety-belt versus D2 distributions for stratum 3

characteristics		safety-belt		D2		difference (%)
		N	%	N	%	
gender	male	991	57.9	4518	50.8	7.0
	female	721	42.1	4367	49.2	-7.0
age	16-29	426	24.9	2755	31.0	-6.1
	30-59	1067	62.4	4953	55.7	6.7
	60+	216	12.6	1177	13.2	-0.6
vehicle type	passenger car	821	64.3	6101	78.1	-13.9
	pickup truck	456	35.7	1706	21.9	13.9

Table 5.2.14. Safety-belt versus D2 distributions for stratum 4

characteristics		safety-belt		D2		difference (%)
		N	%	N	%	
gender	male	2686	59.8	4920	51.8	8.0
	female	1804	40.2	4572	48.2	-8.0
age	16-29	1256	28.0	2848	30.0	-2.0
	30-59	2811	62.6	5423	57.1	5.5
	60+	422	9.4	1221	12.9	-3.5
vehicle type	passenger car	2626	80.4	7019	85.5	-5.1
	pickup truck	641	19.6	1191	14.5	5.1

The comparisons of the safety-belt and D2 distributions are fundamentally the same for each stratum. The operational and statistical significances for each comparison are summarized in table 5.2.15:

Table 5.2.15. Summary of operational and statistical significances for each stratum

stratum		gender	age	vehicle type
stratum 1	operational	yes	no	yes
	chi-square	$p<0.05$	$p<0.05$	$p<0.05$
stratum 2	operational	yes	yes	yes
	chi-square	$p<0.05$	$p<0.05$	$p<0.05$
stratum 3	operational	yes	yes	yes
	chi-square	$p<0.05$	$p<0.05$	$p<0.05$
stratum 4	operational	yes	yes	yes
	chi-square	$p<0.05$	$p<0.05$	$p<0.05$

As shown in table 5.2.15, the majority of the differences are both operationally and statistically significant, since the percentage differences are more than 4 percentage points and the *p*-values given by chi-square are less than 0.05. Although for stratum 1 the difference for age distributions is small between the safety-belt and D2 data, 2 percentage points, it is not sufficient to suggest that D2 data are representative of safety-belt data for stratum 1. The differences of distributions for gender and vehicle type at the same stratum are found to be operationally significant. Based on the information shown in table 5.2.15, it can be concluded that for all four strata examined, there is no agreement of D2 versus seat belt data from operational or and statistical perspectives.

Similar to the comparison at the state level, some other observations are revealed according to the data shown in tables 5.2.11, 5.2.12, 5.2.13, and 5.2.14:

1. The percentages of male drivers in safety-belt data are consistently higher than those in D2s, at least 6 percentage points.
2. The percentages of the middle-age group (30-59) in safety-belt data are always larger. It is consistent with the discussion at the state level, which is most likely caused by the observation errors of driver age during the safety-belt data collection.
3. The passenger car percentages for safety-belt data are consistently smaller. Similar to the state level, it accords with the fact that the "passenger car" category in D2 data contains some sports utility vehicles, while the safety-belt data do not.

Comparisons are also developed for the individual strata at intersections, the results of which are represented in Appendix B. Identical observations as above can be

offered. At the stratum level (intersections), the differences are more conspicuous between D2s and safety-belt data from both operational and statistical point-of-views.

In summary, the discussions have illustrated that there is a disagreement in driver-vehicle characteristic estimates between the safety-belt and D2 data for each stratum and intersection. That means that the validity of the underlying assumptions of quasi-induced exposure could not be confirmed at the stratum level. At the state level, it has been shown that D2 distributions for driver age and vehicle type are significantly different from the safety-belt data. Therefore, it is no surprise that at the stratum level where data are more disaggregated, the results could not show an agreement between the D2s and safety-belt data. However, the results should not lead to the conclusion that quasi-induced exposure is not a good exposure measurement that could be used at the stratum level, since the safety-belt data at the stratum level can't be proved to be "exposure truth."

Although it has been demonstrated that the underlying assumptions of quasi-induced exposure could not be supported using the safety-belt data at the stratum level, there are some insights gained. Table 5.2.16 presents the distributions for passenger cars at the levels of state, state at intersections, stratum, and stratum at intersections.

Table 5.2.16. The distributions for passenger cars at different levels

levels	safety-belt		D2		difference (%)
	N	%	N	%	
state	6394	74.5	31557	81.9	-7.3
state at intersections	4570	74.4	16447	81.9	-7.5
stratum (3)	821	64.3	6101	78.1	-13.9
stratum (3) at intersections	446	60.1	3864	78.2	-18.1

As shown in the "difference" column, the differences in percentages for passenger cars between the safety-belt and D2 data are consistently shown as being negative, indicating that D2 data contain higher percentages of passenger cars at all four levels. It is

probably due to the fact that the "passenger car" category in D2 data encompasses some sports utility vehicles while that in the safety-belt data does not. However, a discernible pattern can be identified from the "difference" column—the differences in absolute values become more conspicuous at increasing levels of disaggregation. The percentages for passenger cars in the D2 data are relatively stable between different levels, while the percentages vary considerably in the safety-belt data. Therefore, the distribution for passenger cars in the safety-belt data is more sensitive to data aggregation than the D2 data; the increasing differences shown in table 5.2.16 are mainly due to the frequency/percentage change in the safety-belt data (exposure data).

Similar observations can be found in the age distributions. An example is given for the middle-age group (30-59) and the results are shown in table 5.2.17.

Table 5.2.17. The distributions for the age group (30-59) at the different levels

levels	safety-belt		D2		difference (%)
	N	%	N	%	
state	7371	62.2	25296	57.2	5.0
state at intersections	5298	62.5	13057	56.7	5.8
stratum (2)	1367	62.9	7123	56.5	6.5
stratum (2) at intersections	948	63.8	4691	56.1	7.7

In table 5.2.17, the differences for the age group (30-59) become more prominent when the safety-belt and D2 data are disaggregated at the fine levels. Another important fact is that the percentages of age group (30-59) in the safety-belt data are consistently greater than those in D2 data. It seems that rough estimates of driver age in the safety-belt data contribute to errors and these errors persist while the data are disaggregated. It is possible that drivers around the upper-bound of age group (16-29) and the lower-bound of age group (60+) are mistakenly counted as age group (30-59), although some errors might cancel out, e.g., 30-31 year-old drivers are counted as being 28-29 year-old.

County level

With the recognition that the differences between the safety-belt and D2 data are more obvious with the data disaggregation, it can be predicted that the distributions of D2 data will be considerably different from the safety-belt data at the county level. The comparisons of distributions for three key driver-vehicle characteristics are presented in Appendix C. The results demonstrate a considerable variation (maximum 22 percentage points) of distributions between the seat-belt and D2 data.

5.2.6 Conclusion

Analysis has been undertaken to compare the driver-vehicle characteristics between the safety-belt and D2 data at the state, stratum, and county levels. Statistically, the results consistently suggest that driver characteristics observed in the safety-belt data are significantly different from D2s as derived from accident data. Practically, at the state level, quasi-induced exposure is a generally good technique in terms of driver gender; however, further justifications are needed for driver age and vehicle type, since there are subject to data observation errors and mismatched vehicle classification.

At the stratum and county levels where operationally significant distribution differences have been shown for three key characteristics, the validity of the quasi-induced exposure can not be verified, because at these levels the safety-belt data do not necessarily represent the "exposure truth." Notwithstanding that, it has been observed that with the safety-belt and accident data at the more finely-disaggregated levels, the variation of distributions by three D2 characteristics becomes more significant. The D2 estimates yielded by quasi-induced exposure are relatively stable, while the distributions given by the safety-belt data vary considerably. Given that, quasi-induced exposure

seems less sensitive to the data disaggregation, especially at the levels such as state, statewide intersections, stratum, and stratum-intersections.

The next section will cover how to validate the underlying assumptions of quasi-induced exposure with truck volume data.

5.3 Validation using truck volume data (W-2)

5.3.1 Introduction

Another source that can be employed to validate the underlying assumptions of quasi-induced exposure is vehicle classification data collected by the Federal Highway Administration's (FHWA) Office of Highway Policy Information (OHPI). The Vehicle Travel Information System (VTRIS) so-called "W-tables" maintained at the FHWA's website. Currently, there are 14 years (1990 - 2003) of data available to the general public.

The W-tables are designed to provide a standard format for presenting the summary of vehicle weighing and classification efforts at truck weigh sites around the country (FHWA website). There are six types of W-tables which contain different truck-related information and are designed for different purposes. For example, the W-4 tables contain "information on truck axle loadings and their effect on flexible and rigid pavement based on 18-KIP equivalent axle loads," which is most commonly used in pavement designs (FHWA website). However, in the research context here, special attention is given to the W-2 tables, which contain potentially useful information for validating the underlying assumptions of quasi-induced exposure.

The W-2 table is a summary of vehicle counts at Weigh-In-Motion (WIM) stations by vehicle classification. At each station, the vehicle classification data are

averaged for each hour and the 24-hour averages are added for the average daily count over a year. So, fundamentally, traffic counts in W-2 tables are Average Annual Daily Traffic (AADT). Note that WIM stations are typically located on freeway links demarcated by intersecting major roads.

Since W-2 data are observed directly from the freeway system, they reflect the actual vehicle type distributions at the specific locations where the WIM sites are located. Thus, the information that W-2 data represent can be regarded as a measurement of exposure "truth." In theory, comparing the exposure as measured using the quasi-induced exposure approach with the W-2 data is a potential validation of the underlying assumption of quasi-induced exposure. This comparison is made for selected Michigan freeway links. However, the nature of W-2 tables and availability of W-2 data limit the efforts. For example, in Michigan (1999) several highway routes have no W-2 tables (e.g., US131); and, for other routes, although W-2 tables are available, no traffic volume data were collected at some WIM sites (e.g., WIM station #6069 at I-69). Given the available data from the FHWA website, three interstate routes (I-94, I-75, and I-96) and two US routes (US23 and US12) are chosen for analysis.

The next section covers the methodology for identifying comparable D2 and W-2 data and utilizing W-2 data to validate the underlying assumptions of quasi-induced exposure.

5.3.2 Methodology

In order to use W-2 data, the accident data must be cleaned according to the procedure described in chapter 4 to eliminate one- and three-or-more-vehicle accidents and accidents with conflicting information. Further, accident data need to be selected in a

way to reflect the same conditions under which W-2 traffic counts were collected in terms of timeframe, vehicle type classification, and accident locations.

Vehicle type

Considering the nature of the W-2 data, vehicle type is the only driver-vehicle characteristic that will be used in the comparison. Due to the different vehicle-type classification schemes used in the W-2 and D2 data, it is necessary to adjust the vehicle-type categorizations to make them comparable. The data in the W-2 tables are broken down by 13 vehicle types shown in table 5.3.1. The vehicle type classifications for accident data are in table 5.3.2.

Table 5.3.1. FHWA vehicle classes in W-2 tables

1. motorcycles	2. passenger cars	3. single-unit vehicles: 2-axle, 4-tire
4. buses	5. single-unit trucks: 2-axle, 6-tire	6. single-unit trucks: 3-axle
7. single-unit trucks: 4-axle, or more	8. single-trailer trucks: 4-axle, or less	9. single-trailer trucks: 5-axle
10. single-trailer trucks: 6-axle, or more	11. multi-trailer trucks: 5-axle, or less	12. multi-trailer trucks: 6-axle
13. multi-trailer trucks: 7-axle, or more		

Table 5.3.2. Vehicle type in Michigan accident data (2001)

00 uncoded & error	01 passenger car & station wagon	02 van, motorhome
03 pickup	04 truck under 10,000 lb	05 cycle
06 moped	07 go-cart	08 snowmobile
09 ORV, ATV	10 other non-commercial	11 misc commercial
12 combo unit	13 combo, hazardous	14 combo, tank
15 combo, passenger	16 combo, double or	17 combo, tank &
18 combo, double or triple tank	19 combo, double or triple hazardous	20 combo, double or triple tank
21 single over 26k	22 single, hazardous	23 single, tank
24 single, passenger	25 single, tank	26 under 26k,
27 under 26k, passenger	28 under 26k, tank	29 others

Comparison of tables 5.3.1 and 5.3.2 shows that although vehicles are classified differently, five consistent vehicle types can be identified:

1. Motorcycles. Value 05 "cycle" in the accident data is equivalent to value 1 "motorcycle" in the W-2 data.
2. Passenger cars. In the W-2 data, value 2 "passenger cars" includes all sedans, coupes, and station wagons manufactured primarily for the purpose of carrying passengers. Thus, value 01 "passenger car & station wagon" in the accident data matches value 2 "passenger cars" in the W-2 data.
3. Pickups and vans. In the W-2 data, value 3 "single-unit vehicles, 2-axle, 4-tire" includes pickups, panels, vans, and other vehicles such as campers, motorhomes, ambulances, and minibuses. Thus, value 3 in W-2 data is equivalent to values 02 "van and motorhomes" and 03 "pickups" in the accident data. However, this fit may not be as good as the first two types.
4. Bus. Values 15 "combo, passenger", 24 "single, passenger," and 27 "under 26k, passenger" combined in the accident data go with value 4 "bus" in the W-2 data;
5. Other trucks. According to tables 5.3.1 and 5.3.2, the rest of the trucks are categorized with different schemes. However, summation of them will belong to the same category, which includes all kinds of trucks, exclusive of pickups. Values 04 and 12-28 (exclusive of 15, 24, and 27) in the accident data match values 5-13 in the W-2 data.

It is noticed that several values in table 5.3.2 do not fit in the above five categories, including 06-11 and 29. In W-2 data, there is no vehicle class that can match those values. Therefore, vehicle types 06-11 and 29 in the accident data will not be used.

Timeframe

For Michigan, there is great flexibility in selecting different years of Michigan accident data and W-2 data to develop the comparisons. Currently, both accident and W-2 data are available from 1990 to 2003.

Locations

Since all of the WIM stations are located on interstates or US routes, selected two-vehicle accidents must occur on freeway sections. This would eliminate all accidents on ramps, local surface roads, and intersections. Note that the accidents occurring on ramps are filtered out because the non-guilty drivers in these accidents, if actually on surface roads when struck (e.g., at the end of an exit ramp), will not be representative of the driving population on the freeway sections. Accidents on freeway sections can be located by studying the control section and area code in the accident data. In the meanwhile, the accident data must be selected to reflect the accident types which could reasonably occur on freeways. Similar to the discussion in chapter four, accidents occurring on freeway sections take forms of three major types: miscellaneous multiple vehicle, rear-end straight, and side-swipe same. Correspondingly, the D2 data only consist of these four accident types.

After the appropriate screening operations are undertaken, accident responsibility is assigned to two-vehicle accidents. Accidents with one responsible and one non-responsible driver-vehicle combination will be used for analysis. Thus, the accident data are chosen to be coincident with the circumstances where and when the W-2 data were observed. The following section describes how the W-2 data are used in the comparison.

An example is given for a freeway segment on I-94 in Michigan. It is ascertained that traffic composition along a freeway may change where two major intersecting traffic streams merge or diverge. Therefore, it is necessary to locate the major intersecting roads along the freeway section. In figure 5.3.1, the locations of the several WIM stations and major intersecting highways along I-94 in Michigan are depicted.

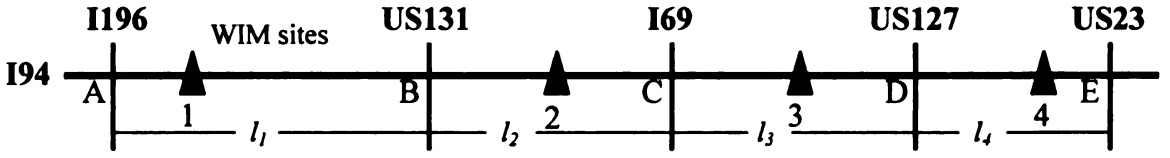

Figure 5.3.1. Illustration of stations and major intersecting roads on I-94 in Michigan

Two assumptions have been made in order to develop a logical comparison between D2 and W-2 data. First, it is assumed that the traffic count observed at each WIM station will be representative of the traffic composition along the corresponding link (l_i). For instance, the traffic count from WIM station 1 will represent the traffic composition along link AB. Second, it is assumed that the traffic characteristics along each link remain the same and traffic at minor intersecting roads does not affect the traffic composition along the link. Based on these two assumptions, D2s can be compared with W-2 data for I-94 (with reference to figure 5.3.1). The comparison is developed in a stepwise manner:

1. Calculate the vehicle type distribution of D2s on the entire roadway segment (AE) between I-196 and US-23;
2. Determine the average traffic counts at four weighting stations (1, 2, 3, and 4), disaggregated by vehicle type (that will be representative of the average traffic count on the roadway segment AE); and
3. Compare the results from steps 1 and 2 (overall comparison).

If similarity is found in the comparison, it can be argued that the underlying assumption of quasi-induced exposure is validated under the given circumstances.

5.3.3 Comparison of W-2 and D2 data

As stated, the comparisons of vehicle type distributions between the W-2 and D2 data are developed for three freeway routes and two US routes. Based on the methodology, the vehicle type distribution for a highway (e.g., I-94) in W-2 data is averaged over all of the available WIM stations for roadway segments in a year. The accident data used in the analysis are reflective of the same freeway, timeframe, and the reasonable accident types. The details of W-2 and D2 distributions by vehicle type are presented in table 5.3.3 (frequencies) and table 5.3.4 (percentages).

Table 5.3.3. Comparing vehicle type distributions between D2s and W-2 (frequencies)

vehicle type	I-94		I-96		I-75		US23		US12	
	W2	D2	W2	D2	W2	D2	W2	D2	W2	D2
motorcycle	11	2	57	0	37	4	54	0	1	0
passenger car	11879	941	19382	718	10006	998	15960	234	2112	62
pickups and vans	2808	353	4843	260	2570	406	4270	129	911	24
bus	72	2	60	0	98	1	55	0	8	0
other trucks	5406	100	4908	67	2155	120	4697	32	612	8

Table 5.3.4. Comparing vehicle type distributions between D2s and W-2 (percentages)

vehicle type	I-94		I-96		I-75		US23		US12	
	W2	D2	W2	D2	W2	D2	W2	D2	W2	D2
motorcycle	0.1	0.1	0.2	0.0	0.2	0.3	0.2	0.0	0.0	0.0
passenger car	58.9	67.3	66.3	68.7	67.3	65.3	63.7	59.2	58.0	66.0
pickups and vans	13.9	25.3	16.6	24.9	17.3	26.6	17.1	32.7	25.0	25.5
bus	0.4	0.1	0.2	0.0	0.7	0.1	0.2	0.0	0.2	0.0
other trucks	26.8	7.2	16.8	6.4	14.5	7.8	18.8	8.1	16.8	8.5

In addition, chi-square is also employed to compare the D2 and W-2 distributions for individual routes. The chi-square statistics and *p*-values are summarized in table 5.3.5.

Table 5.3.5. Chi-square statistics and *p*-values

	I-94	I-96	I-75	US23	US12
statistics	331	113	119	81	5
p-values	<0.05	<0.05	<0.05	<0.05	0.28

Based on these tables, there are several observations:

1. There is considerable variation in the vehicle type distributions. The percentages of "pickups and vans" in the D2 data are consistently larger than these in the W-2 data, while the percentages of "other trucks" in the D2 data are smaller.
2. Operationally, there is no agreement of vehicle type distributions for the W-2 and D2 data for any of the five routes. The differences in percentage for pickups and vans (9 percentage points), and other trucks (11 percentage points) are greater than the 4-point criterion adopted earlier.
3. Statistically, except for US12, the differences between the W-2 and D2 data are significant at 0.05 significance level.
4. The percentages of "motorcycle" and "bus" are very low in both D2 and W-2 data, which seems realistically true. However, for three highways (I-96, US23, and US12) these percentages are consistently equal to zero in D2 data. It is most likely that the sample sizes of D2 motorcycle data are insufficient to represent the (motorcycle) driving population.

Two approaches are used to increase the sample size: 1) regrouping the vehicle types into two categories (passenger car and others); and 2) combining several years of D2 data for a freeway route. For the first approach, based on tables 5.3.3 and 5.3.4, the

vehicle type distributions in percentage for the D2 and W-2 data are presented in table 5.3.6 and the chi-square statistics and *p*-values are shown in table 5.3.7.

Table 5.3.6. Vehicle type distributions between D2s and W-2 (regrouped)

vehicle type	I-94		I-96		I-75		US23		US12	
	W2	D2	W2	D2	W2	D2	W2	D2	W2	D2
passenger car	58.9	67.3	66.3	68.7	67.3	65.3	63.7	59.2	58.0	66.0
others	41.1	32.7	33.7	31.3	32.7	34.7	36.3	40.8	42.0	34.0

Table 5.3.7. Chi-square statistics and *p*-values (regrouped)

	I-94	I-96	I-75	US23	US12
statistics	38	2.7	2.6	3.4	2.4
p-values	<0.05	0.20	0.20	0.10	0.20

Based on the information displayed in tables 5.3.6 and 5.3.7, determined by the criteria established earlier, the statistical and operational significances of differences between the D2 and W-2 data for these five freeway routes are summarized as follows:

Table 5.3.8. Summary of statistical and operational significances

	I-94	I-96	I-75	US23	US12
statistical	yes	no	no	no	no
operational	yes	no	no	yes	yes

It can be seen from table 5.3.8 that by regrouping the vehicle types the differences become statistically insignificant for four freeway routes (except for I-94); and, the differences for two freeway routes (I-96 and I-75) turn out to be operationally insignificant.

For the second approach, comparisons using multiple years of data (2001-2003) are conducted for I-94 and I-75, considering the availability of W-2 data. The vehicle type distribution on a certain freeway in the W-2 data is calculated according to the traffic count averaged over all the WIM stations on that freeway for the three years under

scrutiny (2001, 2002, and 2003). Table 5.3.9 shows the vehicle type distributions for I-94 and I-75.

Table 5.3.9. Vehicle type distributions for I-94 and I-75 (2001-2003)

vehicle type	I-94				I-75			
	W2		D2		W2		D2	
	N	%	N	%	N	%	N	%
motorcycle	35	0.1	5	0.1	66	0.2	11	0.2
passenger car	34011	61.3	2723	66.2	25931	65.5	3041	67.5
pickups and vans	8110	14.6	1010	24.5	7104	17.9	1049	23.3
bus	216	0.4	4	0.1	189	0.5	3	0.1
other trucks	13070	23.6	374	9.1	6304	15.9	402	8.9

In contrast to the information displayed in table 5.3.4, the discrepancy for vehicle type distributions between the W-2 and D2 data in table 5.3.9 is much smaller for both I-94 and I-75. For example, the percentages of motorcycles in W-2 data become equal to those in D2 data; the difference in percentage for passenger cars is 4.9 percentage points on I-94 with data aggregation over time, as opposed to 8.4 percentage points without. Obviously, the change is due to the increased sample size in the D2 and W-2 data. However, in general, the vehicle type distributions for the two sets of data still differ significantly (e.g., for "other trucks," the difference is as much as 14.5 percentage points). Statistically, the *p*-values yielded by the chi-square test are less than 0.05 for both I-94 and I-75, which suggests that the differences are significant as well.

5.3.4 Conclusion

Analytical results have consistently demonstrated that there are significant differences between the W-2 and D2 data (with five vehicle type categories) for the five freeway routes examined, although the difference is statistically insignificant for US12. The sample size seems to play an important role in the results. With the sample size

increased by regrouping the vehicle types into two categories, the results become mixed regarding whether the D2 data fit the W-2 data. For I-96 and I-75 there is no significant difference between the D2 and W-2 data; for US23 and US12 the difference is statistically insignificant; and for I-94 the differences are both statistically and operationally significant. Using multiple years of data, notwithstanding that the difference for vehicle type distribution remains operationally and statistically significant for I-94 and I-75, the discrepancy is much smaller with large sample size. And, the results suggest that quasi-induced exposure is sensitive to the sample size (data aggregation) when it is used for freeway routes.

Chapter 6

VALIDATION USING THREE-OR-MORE-VEHICLE ACCIDENTS

6.1 Introduction

The previous chapter was directed to a validation of the underlying assumptions of quasi-induced exposure through using available exposure data from other sources (i.e., VMT, seat belt, and W-2 data). The effort here is to develop another technique to achieve the same goal, using internally available information from the accident data themselves—specifically, three-or-more-vehicle accidents.

Generally, applications of quasi-induced exposure use only two-vehicle accident data, discarding the accidents involving only one and three-or-more vehicles. In the context of validating the underlying assumptions of quasi-induced exposure, three-or-more-vehicle accident data may contain some very useful information. The basic idea is to compare the distributions for non-responsible driver-vehicle combinations calculated from three-or-more-vehicle accidents with those from two-vehicle accidents. The argument is that if non-responsible driver-vehicle combinations in two-vehicle accidents are truly a random sample of the driving population at the time of accident occurrence, one should expect that the non-responsible driver-vehicle combinations in three-or-more-vehicle accidents are also representative of the same driving population. Similarly, the

comparison can be conducted between different non-responsible drivers in three-or-more-vehicle accidents.

As discussed in chapter 4, three-or-more-vehicle accidents that are usable must consist of at least one responsible and one non-responsible driver. The involvement mechanism is analogous to that used in two-vehicle accidents—a typical accident must have one responsible driver and one non-responsible driver. Based on the responsibility assignment scheme developed in chapter 4, there are some accidents without responsible drivers. It is essential that those accidents should be excluded from the analysis here as well. A comparison will also be made between the first non-responsible drivers (denoted as D2's as opposed to D2s) and the rest of the non-responsible drivers (denoted as D3s) in three-or-more-vehicle accidents. Note that D2' and D3 denoted in this way is for ease of explanation. Comparison of the characteristic patterns of these two non-responsible driver-vehicle combinations could help validate the underlying assumptions of quasi-induced exposure.

There are several advantages in utilizing three-or-more-vehicle accidents in validating the underlying assumptions. First, three-or-more-vehicle accidents are readily available in the accident database. There is always a variable available to identify/separate accident cases with three-or-more vehicles involved. Second, three-or-more-vehicle accidents are coded in the same manner as two-vehicle accidents in the accident database. Therefore, when data are grouped or manipulated in the same way, there should not be any consistency issues between two- and three-or-more-vehicle accidents. In addition, the systematic rules developed for preparing two-vehicle accidents for analysis and the responsibility assignment scheme can be conveniently applied to

three-or-more-vehicle accidents as well. However, there are also some issues in using three-or-more-vehicle accidents. It is perceived that the conditions/circumstances for three-or-more-vehicle accidents might be inherently different from those for two-vehicle crashes. For example, three-or-more-vehicle accidents are most likely to occur during the peak hour or under otherwise congested traffic conditions. Therefore, it is necessary to investigate to see if the characteristics of two-vehicle accidents are indeed different where and when three-or-more-vehicle accidents take place.

The accident data used to develop the comparisons in this section are obtained from HSIS and Michigan DOT. The same three participating states are selected based on the availability of data coding guidebook from HSIS: California, Maine, and Utah. In addition, Michigan data are also utilized, which are available from Michigan DOT. For each state, the variables used to depict the characteristics of driver-vehicle combinations are fundamentally identical: driver age, driver gender, and vehicle type. Since there is some inconsistency in vehicle type definitions from state to state, generally, the vehicle type is simplified to cars, pickups, and other vehicles.

A stepwise procedure for validating the underlying assumptions will be discussed first. Then, the characteristics of three-or-more-vehicle accidents will be explored. Then, comparisons of D2 distributions between two- and three-or-more-vehicle accidents will be conducted at the state level. Finally, a summary will cover what has been learned from the comparisons.

6.2 Methodology

After the HSIS data have been manipulated as described in chapter 5, they can be used in the quasi-induced exposure application. A number of factors need to be taken into

account to select the appropriate accident data, such as accident location, time of day, and crash type. The following stepwise procedure is employed:

1. Clean and select two-vehicle accident data, identify accidents with one responsible and one non-responsible driver, and calculate D2 distributions for the three key characteristics;
2. Clean and select three-or-more-vehicle accident data, identify accidents with at least one responsible and at least one non-responsible driver and calculate the distributions for the three key characteristics for all non-responsible drivers in three-or-more-vehicle accidents;
3. Calculate the distributions for the three key characteristics for D2's and D3s in three-or-more-vehicle accidents; and
4. Examine the differences between steps 1 and 2, and D2's and D3s in step 3, to see if they are significantly different from zero.

If no significant difference is found, the underlying assumptions of quasi-induced exposure are validated for the basic situation that the accident data represent; otherwise, the assumptions can't be validated. Note that the criteria as the previous chapter are used to determine if a difference is significant.

6.3 Characteristics of three-or-more-vehicle accidents

In order to use three-or-more-vehicle accidents to validate the underlying assumptions of quasi-induced exposure, the validation must be firmly grounded on the understanding of the characteristics of three-or-more-vehicle accidents. Compared to two-vehicle accidents, the underlying causes for three-or-more-vehicle accidents may be different—they generally occur where several (more than 2) vehicles to drive closely.

Three-or-more-vehicle accidents (Michigan 2001) are utilized as an example to identify the circumstances where these accidents are prone to occur. The variables of interest include accident location, time of day, crash type, and speed limit. The characteristics of three-or-more-vehicle accidents are compared to those of two-vehicle accidents for different variables.

Accident location

For Michigan data (2001), simple descriptive statistics of two- and three-or-more-vehicle accident locations indicate that the majority of accidents occur at three mutually exclusive locations: intersection areas, freeways, and straight surface roads. The accident location is identified with variable "area code." Table 6.1 presents the percentages at each of the three major locations:

Table 6.1. Percentages of accidents by three major locations (%)

locations		2-vehicle accidents	>2 vehicle accidents
intersection	within intersection	25.2	17.4
	driveway within 150 ft of intersection	8.2	5.9
	other intersection related	13.7	17.7
freeways		5.2	10.7
straight surface road		29.9	36.8
total		82.2	88.5

From table 6.1, it seems that for 82.2% of two-vehicle and 88.5% of three-or-more-vehicle accidents occurring at these three locations. Approximately 47.5% of three-or-more-vehicle accidents occur on freeway areas and straight surface roads—12 percentage points more than two-vehicle accidents; however, three-or-more-vehicle accidents are less likely to take place at intersections (7 percentage points less). Based on the criterion adopted earlier, in term of accident location the difference between two- and three-or-more-vehicle accidents is claimed to be significant.

Time of day

The investigation of time of day shows that two- and three-or-more-vehicle accidents are liable to take place in hours where traffic is more congested, including morning peak hours (7-9am) and afternoon hours (12-7pm). The percentages for two-vehicle and three-or-more-vehicle accidents in these hours are presented in table 6.2

Table 6.2. Percentages of accident by 10 major hours (%)

hour	7-8am	8-9am	11-12am	12-1pm	1-2pm
2-vehicle accidents	4.9	5.0	5.6	6.9	6.3
>2 vehicle accidents	6.5	5.1	5.1	6.1	6.3
hour	2-3pm	3-4pm	4-5pm	5-6pm	6-7pm
2-vehicle accidents	7.6	9.8	9.4	9.3	6.4
>2 vehicle accidents	8.0	10.9	10.9	11.5	6.9

The percentages of accidents in these 10 hours are 71.3% for two-vehicle accidents and 77.2% for three-or-more-vehicle accidents. For 7-9am, 11am-3pm, and 6-7pm, the percentages of two-vehicle accidents are approximately equal to those of three-or-more-vehicle accidents; for hours (3-6pm) including evening peak hour, the percentages of three-or-more-vehicle accidents are consistently higher by 1-2 percentage points. Given the small percentage variation, it is argued that three-or-more-vehicle accidents are not different from two-vehicle accidents in terms of time of day.

Crash type

For two- and three-or-more-vehicle accidents there are three major accident types (table 6.3), including angle straight, rear-end straight, and side-swipe same.

Table 6.3. Percentages of accidents by three major accident types (%)

crash type	angle straight	rear-end straight	side-swipe same	total
2-vehicle accidents	16.8	30.3	12.3	59.5
>2 vehicle accidents	10.2	60.2	5.3	75.7

The results show that for each accident type the differences in percentage are quite significant between two- and three-or-more-vehicle accidents. For rear-end straight, difference is substantial, approximately 30 percentage points. It can be stated that the predominant crash type of three-or-more-vehicle accidents is different from that of two-vehicle accidents.

Speed limit

The speed limit on roadways where accidents occur is also examined. It is found that accidents mostly happen on roadways with speed limits from 25mph to 55mph. The results are shown in table 6.4.

Table 6.4. Percentages of accidents at different speed limits (%)

speed limit	2-vehicle accidents	>2 vehicle accidents
<25	1.0	0.4
25	24.4	11.9
30	8.2	8.2
35	19.2	19.7
40	9.9	12.6
45	17.4	22.5
50	3.9	5.5
55	11.3	10.9
60	0.1	0.1
65	1.5	3.2
70	3.3	5.1

In table 6.4, the percentages of accidents are generally close between two- and three-or-more-vehicle accidents on roads with speed limits of 30 and 35. On roadways with speed limit of 25mph or less, the percentage of two-vehicle accidents is about 13 percentage points higher than three-or-more-vehicle accidents, which is significant; on roads with speed limits of 40mph or more, the percentages of the three-or-more-vehicle accidents are consistently larger (except for speed limit 55mph). It suggests that,

comparatively, three-or-more-vehicle accidents are prone to occur on roadways with higher speed limits. Conclusively, three-or-more-vehicle accidents occur on roads with different speed limits (versus two-vehicle accidents).

In summary, based on the above analyses, it has been identified that three-or-more-vehicle accidents have several noticeable differences when compared to two-vehicle accidents:

- Three-or-more-vehicle accidents are prone to occur on freeway sections and straight roads;
- Three-or-more-vehicle accidents have three primary accident types—angle-straight, rear-end straight, and side-swipe same; and
- Three-or-more-vehicle accidents are likely to take place on roadways with speed limits (>40mph).

Given these differences, D2s in two-vehicle accidents must be compared to D2's in three-or-more-vehicle accident only under these conditions where the latter are more likely to occur. If there is no significant difference, D2s in two-vehicle accidents are argued to be comparable with the non-responsible in three-or-more-vehicle accidents; if there is difference, they are not comparable.

With the use of Michigan accident data (2001), the characteristics of D2s in two-vehicle accidents are compared under two circumstances: three major accident types (set one) and others (set two). Table 6.5 shows the distributions for three characteristics for D2s under these two distinctive circumstances.

Table 6.5. Characteristics of D2s under two defined circumstances

characteristics		set one		set two		difference (%)
		N	%	N	%	
gender	male	47959	54.0	29188	53.6	0.4
	female	40812	46.0	25253	46.4	-0.4
vehicle type	cars	70220	79.1	43642	80.2	-1.1
	pickups	13453	15.2	7818	14.4	0.8
	others	5098	5.7	2981	5.5	0.2
age	15-19	8609	9.7	6152	11.3	-1.6
	20-29	20099	22.6	12407	22.8	-0.2
	30-39	20228	22.8	11746	21.6	1.2
	40-49	18578	20.9	11002	20.2	0.7
	50-59	12064	13.6	6977	12.8	0.8
	60-69	5288	6.0	3369	6.2	-0.2
	70-79	3063	3.5	2120	3.9	-0.4
	80+	842	0.9	668	1.2	-0.3

In table 6.5, the column "difference" refers to the difference between the two sets of accidents. As shown, the differences for all three characteristics are consistently smaller than 2 percentage points. This suggests that the D2 distributions under "set one" and "set two" conditions are operationally close.

The chi-square test was also used to test the differences for individual characteristics. The statistics and *p*-values generated by the test are shown in table 6.6.

Table 6.6. Chi-square statistics and *p*-values for three key characteristics

	gender	vehicle type	age
statistics	2.3	23.4	176.4
p-values	0.20	<0.05	<0.05

In table 6.6, although the results show no statistical difference for driver gender, the *p*-values for vehicle type and age are less than 0.05, indicating that the differences are statistically significant between "set one" and "set two."

Thus, it has been shown that the prevalent accident types of three-or-more-vehicle accidents are angle straight, rear-end straight, and side-swipe same. Under these three

accident types, operationally D2s in two-vehicle accidents are not different. In this context, two-vehicle accidents are arguably comparable with three-or-more-vehicle accidents.

6.4 Comparing D2s between two- and three-or-more-vehicle accidents

The attempt to validate the underlying assumptions of quasi-induced exposure is undertaken for California, Maine, Michigan, and Utah (2000) at the state level. The procedures to clean accidents and assign the responsibility for accident causation developed in chapter 4 are applied to the accident data from these four states.

For each state, two types of comparisons will be conducted:

1. The D2 distributions in two-vehicle accidents for three key characteristics are compared to <u>all</u> the non-responsible drivers in three-or-more-vehicle accidents.
2. In three-or-more-vehicle accidents, the D2' distributions for the three characteristics will be compared to D3s.

<u>California</u>

Table 6.7 presents the frequencies and percentages of D2s in two- and three-or-more-vehicle accidents for California data.

As shown in the "difference" column, there is no difference in driver gender between two- and three-or-more-vehicle accidents. For vehicle type, there is more of a difference, but it is not operationally significant. The difference is a maximum of 3.4 percentage points for passenger cars. The maximum percentage difference for age groups is 3.5 for the 20-29 group, while the differences for other age groups all fall below 2 percentage points. Operationally, the difference is less than the adopted criterion and thus, D2 distributions between two- and three-or-more-vehicle accidents are claimed to

be similar. The D2 distributions are further compared using the chi-square test. The statistical results show that chi-square statistics (p-values in the parenthesis) are 0.01 (p=0.92), 160.6 (p<0.05), and 90.2 (p<0.05) for gender, vehicle type, and age, respectively. The differences for vehicle type and age are statistically significant at 0.05 significance level, while not significant for driver gender. Thus, for California data D2 distributions in two-vehicle accidents agree well with those in three-or-more-vehicle accidents operationally albeit not statistically.

Table 6.7. Comparison of D2 distributions for three key characteristics (CA, 2000)

characteristics		2-veh acc.		>2 veh acc.		difference
		N	%	N	%	
gender	male	12823	64.1	12302	64.1	0.0
	female	7178	35.9	6902	35.9	0.0
vehicle type	cars	13714	68.6	13831	72.0	-3.4
	pickups	4717	23.6	4451	23.2	0.4
	others	1570	7.8	922	4.8	3.0
age	15-19	1197	6.0	1161	6.1	-0.1
	20-29	4492	22.5	4976	25.9	-3.5
	30-39	5119	25.6	4974	25.9	-0.3
	40-49	4577	22.9	4149	21.6	1.3
	50-59	2760	13.8	2412	12.6	1.2
	60-69	1189	5.9	1020	5.3	0.6
	70-79	524	2.6	416	2.2	0.5
	80+	143	0.7	96	0.5	0.2

Further comparisons between D2's and D3s in three-or-more-vehicle accidents are shown in table 6.8. The column "D2's" shows the characteristic distributions of the first non-responsible driver-vehicle combinations in three-or-more-vehicle accidents; the column "D3s" shows the characteristic distributions of the rest of the non-responsible driver-vehicle combinations. The columns "D1" and "D2" show the responsible and non-responsible driver-vehicle combinations in two-vehicle accidents, respectively.

Table 6.8. Comparing distributions for three key characteristics (CA, 2000)

characteristics		>2 vehicle accidents (%)			2-vehicle accidents (%)		
		D2's	D3s	diff.	D1	D2	diff.
gender	male	64.1	64.0	0.1	66.7	64.1	2.6
	female	35.9	36.0	-0.1	33.3	35.9	-2.6
vehicle type	cars	72.5	71.3	1.2	70.2	68.6	1.6
	pickups	22.5	24.1	-1.6	22.7	23.6	-0.9
	others	5.0	4.5	0.5	7.1	7.8	-0.7
age	15-19	6.6	5.3	1.3	12.8	6.0	6.8
	20-29	26.9	24.5	2.4	29.7	22.5	7.2
	30-39	25.5	26.5	-1.0	21.9	25.6	-3.7
	40-49	21.0	22.6	-1.6	17.1	22.9	-5.8
	50-59	12.3	13.0	-0.7	9.3	13.8	-4.5
	60-69	5.0	5.8	-0.8	4.7	5.9	-1.2
	70-79	2.3	2.0	0.3	3.2	2.6	0.6
	80+	0.6	0.4	0.2	1.3	0.7	0.6

For all three characteristics, the differences between D2' and D3 distributions are consistently smaller than 3 percentage points, which suggests that the differences are practically insignificant. In other words, the distributions of first non-responsible drivers in three-or-more-vehicle accidents are similar to the rest of non-responsible drivers. This implies that the non-responsible drivers in three-or-more-vehicle accidents are "randomly selected" by the responsible drivers, regardless of the order in which they are impacted. However, the comparison among D2' and D3 and D2 with chi-square test indicates that the distributions are significantly different.

It has been demonstrated that in table 6.7, the D2s in two-vehicle accidents follow the same distribution as the non-responsible drivers in three-or-more-vehicle accidents; in table 6.8, the first non-responsible drivers in three-or-more-vehicle accidents are representative of the rest of the non-responsible drivers. Deductively, it is safe to argue that D2s in two-vehicle accidents are randomly selected by the responsible drivers. Thus,

in this case the basic assumptions of quasi-induced exposure can be validated for California data.

In table 6.8, D1 and D2 distributions in two-vehicle accidents are also compared. For driver gender and vehicle type, D1 and D2 distributions are operationally close, both below 3 percentage points. However, for driver age there is a conspicuous (operationally significant) variation. For two young driver groups 15-19 and 20-29, the D1 percentages are approximately 6.8 and 7.2 percentage points larger than D2 percentages, respectively; for the old driver group (70+), D1 is greater than D2, although the difference is not significant. The discrepancy between D1s and D2s has been expected; otherwise, the conjecture of D1s and D2s being equal will lead to the conclusion that different drivers have the same propensity for accident causation, which is not factually true.

The results suggest that young drivers (15-29) and old drivers (70+) are more likely to be responsible for the occurrence of accidents. For age group (30-69), the D1 percentages are consistently smaller than D2 percentages, indicating that drivers in this range cause proportionately fewer accidents than their presence on the roads.

Maine

Maine data have also been manipulated with the same procedure as described above. Table 6.9 is the presentation of D2 distributions for three main characteristics for two- and three-or-more-vehicle accidents. The difference for male drivers is about 0.3 percentage point; among the three vehicle types, the maximum difference lies in pickups, about 3.3 percentage points (which is operationally insignificant); for all the age groups, the differences consistently fall below 1 percentage point. Thus, D2 distributions in two-vehicle accidents show a positive agreement with those in three-or-more-vehicle

accidents for all the three characteristics of interest. Conclusively, overall consistency is found between two- and three-or-more-vehicle accidents. However, the results of the chi-square test show that there is no statistical difference for driver gender only—the *p*-values is equal to 0.78 for driver gender. The *p*-values for vehicle type (chi-square statistics: 44.8) and age (chi-square statistics: 136.8) are both less than 0.05, indicating that the difference is statistically significant.

Table 6.9. Comparison of D2 distributions for three key characteristics (ME, 2000)

characteristics		2-veh		>2 veh		diff.
		N	%	N	%	
gender	male	9652	55.4	1455	55.1	0.3
	female	7768	44.6	1185	44.9	-0.3
vehicle type	cars	13783	79.1	2133	80.8	-1.7
	pickups	3282	18.8	408	15.5	3.3
	others	355	2.0	99	3.8	-1.8
age	15-19	3470	9.1	255	9.6	-0.5
	20-29	5049	20.0	539	20.3	-0.3
	30-39	3848	22.1	606	22.6	-0.5
	40-49	3801	21.8	554	21.0	0.8
	50-59	2532	14.5	366	14.3	0.2
	60-69	1226	7.0	187	7.1	0.0
	70-79	629	4.2	97	3.7	-0.1
	80+	335	1.3	36	1.4	0.6

Analogous to what was done with California data, the D2' and D3 distributions for the three characteristics are also compared. The results are shown in table 6.10. The percentage differences listed in the "diff." column are constantly below 4 percentage points, which leads to the conclusion that operationally in three-or-more-vehicle accidents the first non-responsible drivers are representative of all the non-responsible drivers as a whole. Consistent with California data, chi-square test also finds that D2' and D3 and D2 distributions are statistically different.

It is also found that considerable variations also exist between D1s and D2s with regard to three driver-vehicle characteristics for Maine data. Comparatively, the variation is more conspicuous for age distributions than for driver gender and vehicle type. For example, for age group (15-19) the difference is 7.8 percentage points; for age group (40-49) the difference is 6.0 points. These differences are operationally significant. Another observation is that D1s in young (15-29) and old (70+) driver groups are consistently larger than D2s, while it is opposite for age group (30-69). This finding is consistent with that from the California data (table 6.8).

Table 6.10. Comparing distributions for three key characteristics (ME, 2000)

characteristics		>2 vehicle accidents (%)			2-vehicle accidents (%)		
		D2's	D3s	diff.	D1	D2	diff.
gender	male	54.2	56.3	-2.1	58.5	55.4	3.1
	female	45.8	43.7	2.1	41.5	44.6	-3.1
vehicle type	cars	82.0	78.7	3.3	74.9	79.1	-4.2
	pickups	14.8	17.4	-2.6	18.7	18.8	-0.1
	others	3.2	3.9	-0.7	6.4	2.0	4.4
age	15-19	10.3	6.5	3.8	16.9	9.1	7.8
	20-29	21.4	20.2	1.2	23.0	20.0	3.0
	30-39	22.6	22.7	-0.1	17.8	22.1	-4.3
	40-49	20.3	22.7	-2.4	15.8	21.8	-6.0
	50-59	13.2	15.4	-2.2	10.9	14.5	-3.6
	60-69	6.9	8.2	-1.3	6.7	7.0	-0.3
	70-79	3.8	3.4	0.4	6.2	4.2	2.0
	80+	1.5	0.8	0.7	3.5	1.3	2.2

Michigan

Similarly, the comparisons of D2 distributions between two- and three-or-more-vehicle accidents are conducted for Michigan data (table 6.11). Similar to the results in tables 6.6 (CA) and 6.8 (ME), it is shown in table 6.11 that the differences for driver age, gender, and vehicle type are all below 2 percentage points. This indicates that an overwhelming consistency of D2 distributions between two- and three-or-more-vehicle

accidents. However, the statistical results given by the chi-square test suggest that the differences between D2 distributions are significant at the 0.05 level for all the three D2 characteristics, because *p*-values are all less then 0.05.

Table 6.11. Comparison of D2 distributions for three key characteristics (MI, 2000)

characteristics		2-veh		>2 veh		diff.
		N	%	N	%	
gender	male	77105	54.6	6495	56.3	-1.7
	female	64157	45.4	5049	43.7	1.7
vehicle type	cars	101850	72.1	8404	72.8	-0.7
	pickups	37293	26.4	3025	26.2	0.2
	others	2119	1.5	127	1.1	0.5
age	15-19	14267	10.1	1177	10.2	-0.1
	20-29	28252	20.0	2355	20.4	-0.4
	30-39	36304	25.7	2771	24.0	1.7
	40-49	30230	21.4	2436	21.1	0.3
	50-59	18082	12.8	1478	12.8	0.0
	60-69	8476	6.0	762	6.6	-0.5
	70-79	4520	3.2	462	4.0	-0.7
	80+	989	0.7	115	1.0	-0.3

Similar to the exercise with CA and ME data, the D2's versus D3s distributions are compared. The data are presented in table 6.12. It can be seen from table 6.12 that, overall, the D2' and D3 distributions for the three basic characteristics appear to show a generally good agreement—the differences are relatively small (all below 3 percentage points).

The D1 and D2 distributions are also compared for Michigan data (table 6.12). Analogous to the findings in CA and ME data, for driver gender and vehicle type, the differences between D1 and D2 distributions are smaller than 4 percentage points. Other similar observations are offered:

1. For the young driver group (15-29), D1 percentages are larger than D2s and the differences are practically significant.

2. For the older driver group (70+), D1 percentages are also greater than D2s, but the differences are not significant.

3. However, for the middle-aged group (30-69), D1 percentages are smaller.

Table 6.12. Comparing distributions for three key characteristics (MI, 2000)

characteristics		>2 vehicle accidents (%)			2-vehicle accidents (%)		
		D2's	D3s	diff.	D1	D2	diff.
gender	male	57.6	55.1	2.5	57.5	54.6	2.9
	female	42.4	44.9	-2.5	42.5	45.4	-2.9
vehicle type	cars	74.5	71.6	2.9	73.6	72.1	1.5
	pickups	24.8	27.1	-2.3	25.3	26.4	-1.1
	others	0.7	1.3	-0.6	2.1	1.5	0.6
age	15-19	10.9	9.2	1.7	14.3	10.1	4.2
	20-29	20.8	19.8	1.0	24.4	20.0	4.4
	30-39	24.6	22.8	1.8	24.0	25.7	-1.7
	40-49	20.7	22.9	-2.2	19.4	21.4	-2.0
	50-59	12.5	14.1	-1.6	9.2	12.8	-3.6
	60-69	5.5	6.7	-1.2	4.3	6.0	-1.7
	70-79	4.1	3.5	0.6	3.5	3.2	0.3
	80+	0.9	1.0	-0.1	0.9	0.7	0.2

Utah

Table 6.13. Comparison of D2 distributions for three key characteristics (UT, 2000)

characteristics		2-veh		>2 veh		diff.
		N	%	N	%	
gender	male	13819	57.6	4280	57.3	0.3
	female	10179	42.4	3195	42.7	-0.3
vehicle type	cars	17879	74.5	6481	76.6	-2.1
	pickups	4968	20.7	1375	18.4	2.3
	others	1152	4.8	374	5.0	-0.2
age	15-19	3415	14.2	992	13.3	1.0
	20-29	7373	30.7	2415	32.3	-1.6
	30-39	4657	19.4	1508	20.2	-0.8
	40-49	4177	17.4	1302	17.4	0.0
	50-59	2380	9.9	703	9.4	0.5
	60-69	1181	4.9	342	4.6	0.3
	70-79	629	2.6	167	2.2	0.4
	80+	186	0.8	46	0.6	0.2

Table 6.14. Comparing distributions for three key characteristics (UT, 2000)

characteristics		>2 vehicle accidents (%)			2-vehicle accidents (%)		
		D2's	D3s	diff.	D1	D2	diff.
gender	male	56.8	57.7	-0.9	57.7	57.6	0.1
	female	43.2	42.3	0.9	42.3	42.4	-0.1
vehicle type	cars	76.7	76.4	0.3	78.1	74.5	3.6
	trucks	15.5	18.7	-3.2	18.2	20.7	-2.5
	others	7.8	4.9	2.9	3.7	4.8	-1.1
age	15-19	15.0	11.6	3.4	25.6	14.2	11.4
	20-29	33.5	31.2	2.3	30.1	30.7	-0.6
	30-39	18.4	21.9	-3.5	15.8	19.4	-3.6
	40-49	16.8	18.0	-1.2	11.7	17.4	-5.7
	50-59	8.8	10.0	-1.2	7.2	9.9	-2.7
	60-69	4.2	4.9	-0.7	4.0	4.9	-0.9
	70-79	2.4	2.1	0.3	3.6	2.6	1.0
	80+	0.9	0.4	0.5	1.9	0.8	1.1

The identical comparisons (D2 distributions between two- and three-or-more-vehicle accidents, D2's and D3s, D1s and D2s in two-vehicle accidents) are also conducted for Utah data (2000). Due to the redundancy, the data are presented in tables 6.13 and 6.14 (shown as above). Note that similar conclusions are drawn.

6.5 Conclusion

In summary, accident data from California, Maine, Michigan, and Utah have demonstrated generally good agreements of D2 distributions between two- and three-or-more-vehicle accidents. A similar comparison is developed in three-or-more-vehicle accidents between D2's (the first non-responsible drivers) and D3s (the rest of the non-responsible drivers). The results consistently show that there is an overwhelming agreement between D2's and D3s, which suggests that the responsible drivers in three-or-more-vehicle accidents "randomly select" the non-responsible drivers. Deductively, it can be concluded that D2s in two-vehicle accidents are indeed randomly impacted by D1s on the road. Overall, these observations imply that at the state level, the underlying

assumptions of quasi-induced exposure are supported from the operational perspective. That is, quasi-induced exposure is a good technique to measure the exposure of specific driver-vehicle groups of interest at the state level. However, the comparison among D2' and D3 and D2 with chi-square test indicates that the distributions are significantly different.

During the process of validation, the comparisons have confirmed some facts. The distributions for D2s are significantly different from D1s in two-vehicle accidents. The difference is especially obvious for age distributions. The results show that young drivers (15-29) and old drivers (70+) are more likely to be responsible for the occurrence of accidents; however, driver groups (30-69) are likely to be the non-responsible party, because they cause proportionately fewer accidents than their presence on the roads.

Chapter 7

MEASURING EXPOSURE CHANGE—MICHIGAN GRADUATED DRIVER LICENSING

Graduated Driver Licensing (GDL) is a system designed to "teach" teens to drive responsibly by gradually increasing their driving privileges as they advance through the system. The goal of the GDL is to reduce crashes, serious injuries and traffic-related deaths involving teen drivers and their passengers. The GDL imposes specific rules on certain young drivers with regard to nighttime driving—no driving is allowed from midnight to 5 AM. Due to the nighttime restriction, it is expected that the exposure of the affected young drivers during the restricted driving period would be drastically reduced. The research purpose here is to examine whether quasi-induced exposure is able to capture and then represent the exposure change due to the implementation of the GDL program. In order to quantify the effect of the GDL program with respect to exposure, the exposure of young drivers using quasi-induced exposure within and around the restricted nighttime period will be compared before and after the program was put into practice.

7.1 Michigan graduated driver licensing (GDL)

The GDL program implemented in Michigan on April 1, 1997 was a three-tiered (levels) licensing system for young drivers. The eligibility and restrictions of each tier is detailed below. Note that the GDL ends for all teens when they are 18.

Level 1

Level 1 license permits the holder to operate a motor vehicle only when accompanied by a licensed parent or legal guardian, or with the permission of the parent or legal guardian, any licensed driver 21 years of age or older. Note that there is no nighttime driving restriction for young drivers at level 1, if they are accompanied by eligible guardians. To obtain a Level 1 license a driver must:

- be at least 14 years, 9 months of age;
- complete a "Segment 1" driver education course approved by the Michigan Department of Education;
- pass a vision test and meet the physical and mental standards set by the Secretary of State;
- obtain written approval from a parent or legal guardian.

Level 2

A Level 2 license permits the holder to drive unrestricted between 5:00 a.m. and midnight. To obtain a Level 2 license, a driver must:

- be at least age 16;
- complete a "Segment 2" driver education course approved by the Michigan Department of Education;
- have no convictions/civil infractions, license suspensions or crashes during the 90-day period immediately prior to applying for a Level 2 license;

- hold a Level 1 license for six months and complete a minimum of 50 hours of behind-the-wheel driving, including 10 hours of nighttime driving that is certified by a parent or legal guardian; and
- pass a road test conducted by a Secretary of State approved examiner.

Level 3

A Level 3 license allows the license holder unrestricted privileges. To obtain a Level 3 license, a driver must:

- be at least age 17;
- hold a Level 2 license for six months;
- complete 12 consecutive months of driving without a moving violation, an at-fault crash that resulted in a moving violation, a license suspension, or a violation of the graduated license restrictions.

According to the three-tiered driving licensing system, at different levels young drivers have different driving privileges and restrictions.

7.2 Affected young drivers and program in-effect years

Based on the three-tiered licensing program, fewer and different driving restrictions are phased in as the young drivers advance through the program. In a specific nighttime driving period (midnight-5am) drivers with level 1 or 2 license are not allowed to drive alone, in other words, they are restricted; otherwise, drivers are allowed. The diagram below (figure 7.1) graphically illustrates how different ages will be affected by the restricted nighttime driving (midnight to 5am) due to the three-tiered GDL program over the first years of its existence. In this diagram, it is assumed that all young drivers are able to advance to a higher level without any failure.

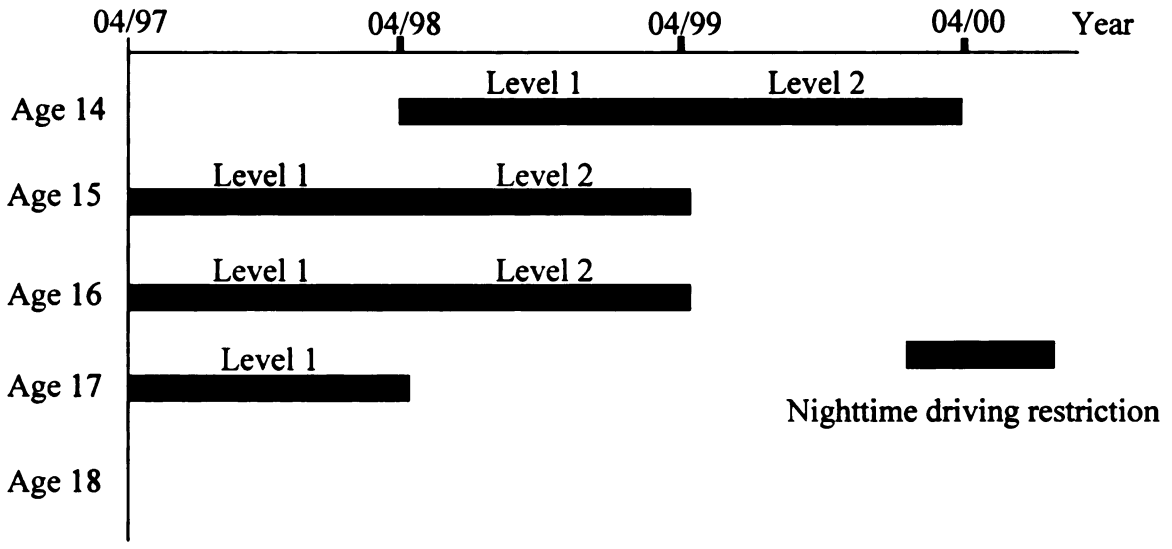

Figure 7.1. Demonstration of different ages affected by the restricted nighttime driving

As shown in figure 7.1, when the GDL program was implemented in April 1997, drivers with different ages were examined individually. For the 14-year-old or younger drivers, they had to wait at least one year to join the GDL program and they were expected to have a minimum of two years of restricted nighttime driving. If the drivers were 15- or 16-year-old when the GDL program started, they would have at least two years of restricted nighttime driving, which depended on how they performed at the levels 1 and 2. Considering that some drivers might fail to advance to level 2 in April 1998, the restricted nighttime driving for some drivers could be extended to period from 04/98 to 04/99. If the drivers were 17, they stayed in the GDL program for one year (level 1) and then were automatically out of the program. If the drivers were 18 or older, they were not in the program. Based on the descriptions above, the affected young drivers were those 15, 16, and 17 years old; the GDL program was "partially" effective in the period from 04/97 to 04/99 and year 2000 was the first year that the GDL program was fully in effect.

7.3 Exposure data analysis

Six years of Michigan accident data are examined, including three years before GDL implementation (1994, 1995, and 1996) and three years after (2000, 2001, and 2002). The main purpose is to determine the frequency shifts for young (15-17) D2 drivers during the restricted nighttime period. The Michigan accident data are screened to make them usable for quasi-induced exposure with specific attention given to driver age, in terms of eliminating data cases with missing and/or miscoded driver age information. The reason is that in this exercise driver age is the only driver-vehicle characteristic that will be examined. Table 7.1 shows the frequencies of non-responsible drivers, disaggregated by driver age, in the restricted driving time period before and after the enforcement of GDL.

Table 7.1. The frequencies of non-responsible drivers by age (midnight-5am)

year	driver age							total
	15	16	17	18	19	20	>20	
1994	2	70	111	164	185	168	3087	3787
1995	2	68	112	163	206	176	3095	3822
1996*	2	71	113	168	174	165	2932	3625
2000**	1	30	78	147	161	156	2836	3409
2001	2	24	72	129	127	145	2200	2699
2002	3	21	61	115	115	116	2265	2696

Note: * Michigan GDL law was effective on April 1997, so data from year 1996 and earlier did not have GDL restrictions; ** data from year 2000 and later have full GDL restrictions.

In table 7.1, the affected driver ages are highlighted with shaded cells. The total number of D2s for 15 years old drivers is very small (1-3 accidents/year) for all six years, and it is hard to quantify the effectiveness of the GDL program. Therefore, 15-year old drivers are not included in the discussion. However, a noticeable reduction can be seen in the absolute number of D2s for all ages examined along the boundary where the GDL

program was implemented. Before the GDL was implemented, the number of D2s for each age is generally consistent within the three years examined (1994, 1995, and 1996).

However, the number of D2s shows a general trend of decreasing from 2000 to 2002. For ease of explanation, the average D2s before and after the GDL for each age and D2 changes in frequencies and percentages are calculated in table 7.2.

Table 7.2. Average D2s before and after the GDL and changes (midnight-5am)

GDL	driver age						total
	16	17	18	19	20	>20	
before	70	112	165	188	170	3038	3745
after	25	70	130	134	139	2434	2935
change	45	42	35	54	31	604	810
change (%)	64.1	37.2	21.0	28.7	18.1	19.9	21.6

It can be seen from table 7.2 that, overall, there is a considerable reduction of D2s for all ages and the total after the GDL was implemented. The total number of D2s decreases by 21.6% for some reason (e.g., traffic safety policy or programs), which presumably contributes, at least partially, to D2 percentages decreasing for all drivers. Given that, the D2 change in percentage for 16-year-old drivers is still significant: 64.1% is considerably greater than the 21.6% decrease overall. The 37.2% of D2 reduction for 17-year-old drivers is also larger than the overall decrease.

Since the total number of D2s (last column in table 7.1) varies considerably from year to year, it is more meaningful to compare the D2s in percentages over the years (table 7.3).

As illustrated, there is a conspicuous gap in the D2 percentages for the affected drivers at the point where the GDL program was implemented, which is not so obvious for other ages. D2 percentages for 16-year-old drivers before the GDL program (averaging 1.8%) are consistently higher than those after the GDL program was initiated

(averaging 0.8%) by 1.0%. Although this number itself does not suggest a significant amount, the relative proportion change, defined as the ratio of average percentage change (1.0%) to average D2 percentage without the GDL program (1.8%), is quite substantial, as much as 54.4%. The results also show D2 decrease in percentage for 17-year-old drivers (0.59%), however, the relative proportion change is less significant, 19.6%. This analysis demonstrates that with the quasi-induced exposure technique, the exposure reduction for 16- and 17-year-old drivers is reflected in the exposure data. Meanwhile, it is no surprise to see some 16- and 17-year-old D2 drivers in the 2000, 2001, and 2002 data, since drivers at these ages can still drive legally between midnight to 5am with guardians when they are at level 1 or 2. An alternative explanation is that many young drivers will continue to drive during the curfew in spite of the law.

Table 7.3. The percentages of non-responsible drivers (midnight-5am)

year	driver age						total
	16	17	18	19	20	>20	
1994	1.85	2.93	4.33	4.89	4.44	81.52	100.00
1995	1.78	2.93	4.26	5.39	4.60	80.98	100.00
1996	1.96	3.12	4.63	4.80	4.55	80.88	100.00
2000	0.88	2.29	4.31	4.72	4.58	83.19	100.00
2001	0.89	2.67	4.78	4.71	5.37	81.51	100.00
2002	0.78	2.26	4.27	4.27	4.30	84.01	100.00

Moreover, D1 percentages for all age groups are also computed before and after the GDL implementation. Table 7.4 shows the frequencies of responsible drivers in the restricted driving period before and after the GDL program was implemented. Comparing the D2s (table 7.1) to D1s (table 7.4), it can be seen that D1s are generally larger than D2s in the corresponding year for the affected drivers (16 and 17) and several other young drivers (18-20), no matter whether the GDL is implemented. In other words, the

involvement ratios for young drivers are consistently greater than 1. These results generally confirm the fact that young drivers are indeed the more dangerous driving cohorts in the traffic stream.

Table 7.4. The frequencies of responsible drivers (midnight-5am)

year	driver age						total
	16	17	18	19	20	>20	
1994	125	147	211	208	151	2938	3787
1995	119	142	225	215	187	2932	3822
1996	126	138	212	162	172	2808	3625
2000	60	109	213	180	173	2672	3409
2001	44	79	151	166	138	2116	2699
2002	48	113	173	159	148	2051	2696

In addition, D2 percentages for the 16- and 17-year-old drivers in three hours just before and after the restricted period are also examined. The purpose is to check if the D2 distributions of the affected drivers in the restricted driving period shift to the adjacent non-restricted time period for the years after the GDL program was executed. Six years including three before the GDL implementation (1994, 1995 and 1996) and three after (2000, 2001 and 2002) are chosen to illustrate how D2 percentages evolve with the time of day. The results are shown in figure 7.2 for the 16-year-old drivers. It is seen from figure 7.2 that the time-of-day D2 distribution for each year examined appears to be U-shaped. Three D2 trendlines for years before the GDL implementation (abbreviated as "D2 trendlines before") and three for years after (abbreviated as "D2 trendlines after") are similar in nature. Although at some hours the relative position of D2 trendlines before and after fluctuate up and down, some general trends are observed based on these two sets of D2 trendlines:

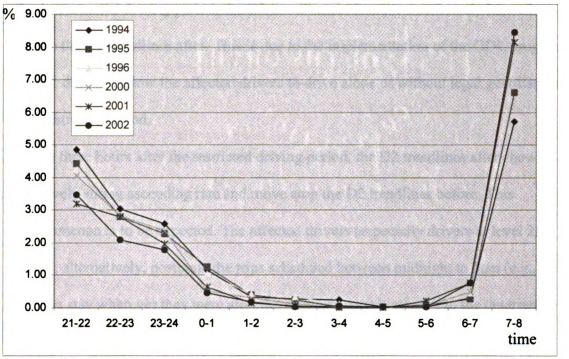

Figure 7.2. The D2 percentages of the 16-year-old drivers (time-of-day distribution)

1. In the three hours before the restricted driving period, generally, the D2 trendlines after fall below the D2 trendlines before. This is not expected, since the restriction of nighttime driving should cause the affected drivers (especially those at level 2) to drive relatively more in this period to compensate the travel scheduled between midnight to 5am (e.g., young drivers end the midnight activities earlier). Thus, the nighttime driving restriction pushes the affected drivers out of time period (midnight-5am) and D2 percentages after the GDL was implemented should be larger in the period (9-12pm). However, the further investigation of D2 distribution in the earlier hours indicates that D2 trendline after is generally located atop D2 trendline before. It suggests that affected young drivers will drive relatively more outside the restricted nighttime driving period to compensate the driving activity suppressed by the GDL law in a long-time stretch.

2. In the restricted driving period, the D2 trendlines before are consistently located above the D2 trendlines after. This is due to the implementation of the GDL program, which does not allow the affected drivers to drive alone or without legal guardians in the restricted period.
3. In the three hours after the restricted driving period, the D2 trendlines after show a relatively strong ascending rate and move atop the D2 trendlines before. This phenomenon is to be expected. The affected drivers (especially drivers at level 2) could, alternatively, postpone the trips scheduled between midnight to 5am (e.g., drivers stay wherever they were after the midnight activities), due to the deterrent effect of the GDL program.

These trends have shown that with the quasi-induced exposure approach, the D2 time-of-day distribution for 16-year-old drivers can only partially represent the effectiveness of the GDL program. There is a shift of D2 distributions in three hours after the nighttime driving restrictions but not before.

7.4 Summary

In summary, using several years of accident data, quasi-induced exposure could partially pick up the exposure change resulting from the GDL implementation. It is reflected from two perspectives: 1) the exposure is noticeably reduced for the affected 16- and 17-year-old drivers; 2) the time-of-day distribution indicates that D2 percentages are shifted to hours after the period of midnight-5am, but not before (at least three hours as the data present). From this point-of-view, it can be argued that quasi-induced exposure is a legitimate approach in this instance. As a closing note, it needs to be

pointed out that if the results do not show the exposure change of the affected group, it does not necessarily constitute the evidence that quasi-induced exposure is invalid.

Chapter 8

DIFFICULTIES WITH QUASI-INDUCED EXPOSURE

In chapter 5, it has been observed that even with large sample size, for I-94 the D2 data differ significantly from the W2 data even with the data aggregation. While the sample size is an issue, a question is also raised regarding potential theoretical difficulties associated with the quasi-induced exposure technique. It is known that freeway I-94 is especially favored by heavy truck traffic, 33% at some WIM locations. Could heavy truck traffic be an essential element affecting the D2 measurement on I-94?

On a typical freeway, it is argued that substantial speed variation exists between passenger cars and trucks. Based on Taylor's study of freeway speeds (2000), the mean speed of automobiles and light trucks is, on average, 7 MPH higher than that of heavy trucks. The problem arises from the fact that using quasi-induced exposure implicitly assumes that the travel speeds are identical for different types of vehicles.

An example with passenger cars and trucks moving along an identical freeway segment is given to illustrate how the relative accident involvement ratio for different vehicle types can be incorrect when speed variation is evident. The relative accident involvement ratio is expressed as the number of times that a certain vehicle type is responsible (hits other vehicles) to the number of times that the same type of vehicle is non-responsible (is hit by other vehicles). Mathematically

$$IR_i = \frac{D1_i}{D2_i} \tag{8.1}$$

where:

IR_i – the relative accident involvement ratio for vehicle type i;
i – a vehicle type;
$D1_i$ – the number of times that vehicle type i is responsible; and
$D2_i$ – the number of times that vehicle type i is non-responsible.

The formulation above is used to calculate the relative accident involvement ratio for different types of vehicles. However, when the traveling velocity is taken into consideration, there are problems involved with the computation of the numerator and the denominator of the equation. The problem becomes more typical on freeways where the speed variation between different vehicles is conspicuous and it can be safely assumed that passenger cars generally move faster than trucks. This is explained by comparing the magnitude of the number of non-responsible and responsible drivers between two scenarios ("no speed variation" and "speed variation"). It is achieved in a stepwise manner, including assumptions, model development, constraints, and comparison.

Assumptions

In order to develop the proposed comparisons, three assumptions are made:

1. The symbols used in two scenarios are distinguished with a prime ('). For the "no speed variation" scenario, the number of non-responsible and responsible drivers are labeled as D1 and D2, respectively; for the "speed variation" scenario, they are labeled as D1' and D2'.
2. Two-vehicle accidents are involved with only two types of vehicles: passenger cars or trucks. Once one vehicle initiates an accident, this vehicle is responsible; otherwise, it is non-responsible.

3. Speed variation between different vehicle types has no effect on the propensity of accident occurrence between the same vehicle types. However, it will affect the accident propensity between different vehicle types. As reflected in the traffic characteristics, overtaking is one of the most common driving maneuvers to cause speed variation. Based on this deduction, overtaking between different vehicle types is likely to cause accidents.

Model development

Under these two scenarios, the number of non-responsible and responsible drivers is calculated based on the information shown in figure 8.1. In figure 8.1, N_{ij} is the number of times that vehicle *i* hit vehicle *j*. The subscript P stands for passenger cars and T stands for trucks.

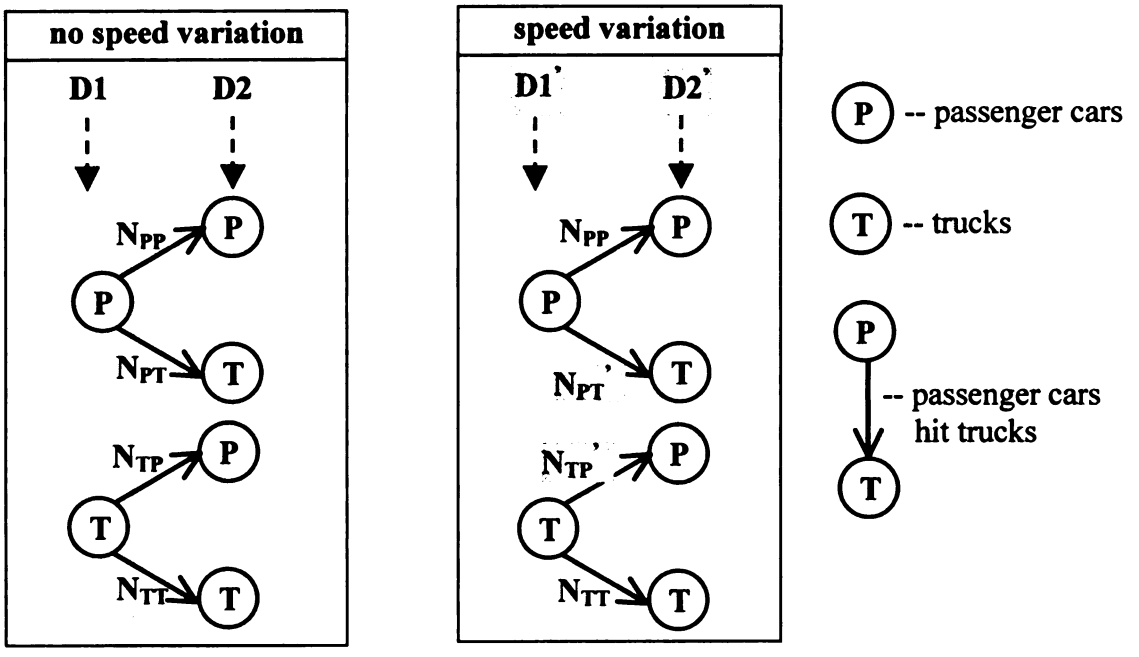

Figure 8.1. D1 and D2 for passenger cars and trucks under two scenarios

Based on the denotation in figure 8.1, the number of non-responsible and responsible drivers for passenger cars and trucks with and without considering speed variation can be expressed as:

"no speed variation:"

$$D1_P = N_{PP} + N_{PT},\ D1_T = N_{TT} + N_{TP} \tag{8.2}$$

$$D2_P = N_{PP} + N_{TP},\ D2_T = N_{TT} + N_{PT} \tag{8.3}$$

"speed variation:"

$$D1'_P = N_{PP} + N'_{PT},\ D1'_T = N_{TT} + N'_{TP} \tag{8.4}$$

$$D2'_P = N_{PP} + N'_{TP},\ D2'_T = N_{TT} + N'_{PT} \tag{8.5}$$

Constraints

When speed variation between different vehicle types is taken into account on freeways, it suggests the consideration of a specific driving maneuver—overtaking between passenger cars and trucks. However, given the "speed variation" scenario, passenger cars move faster and will overtake trucks more than they do with no speed variation. Thus, in the event of accident occurrences, passenger cars are more likely to be the guilty party and trucks are the "victims." Given this illustration, a relationship between N_{PT} and N'_{PT} is established:

$$N_{PT} < N'_{PT} \tag{8.6}$$

The same theory applies to the event of trucks overtaking passenger cars—the probability of overtaking is certainly less with "speed variation." So, another relationship exists:

$$N_{TP} > N'_{TP} \tag{8.7}$$

Comparison

Based on these two inequalities (8.6 and 8.7), other relationships can be derived from equations 8.2, 8.3, 8.4, and 8.5:

$$D1_P < D1'_P,\ D2_P > D2'_P \tag{8.8}$$

$$D1_T > D1'_T,\ D2_T < D2'_T \tag{8.9}$$

Inequalities 8.8 and 8.9 indicate that using quasi-induced exposure where substantial speed variation exists will result in relative exposure being overestimated for passenger cars and underestimated for trucks. Then, replacing the denominator and numerator in equation 8.1 with inequalities 8.8 and 8.9, there is,

$$IR_P = \frac{D1_P}{D2_P} < IR'_P = \frac{D1'_P}{D2'_P} \tag{8.10}$$

$$IR_T = \frac{D1_T}{D2_T} > IR'_T = \frac{D1'_T}{D2'_T} \tag{8.11}$$

Inequalities 8.10 and 8.11 suggest that quasi-induced exposure results in the overall estimation of relative accident involvement ratio for passenger cars being comparatively lower and trucks being relatively higher, where considerable speed variation is apparent.

Theoretically, it has been demonstrated that there are problems with quasi-induced exposure when there is speed variation. In the following section the same issue is explored from the practical perspective using Michigan accident data from 2001. The basic idea is to calculate D1s, D2s, and IRs for fast- and slow-moving vehicles and identify the patterns for these three variables with different speed limits. An essential assumption behind the idea is that with a speed limit increases, the speed variation between slowing- and fast-moving vehicles increases. Arguably, the assumption is valid,

because increasing the speed limit opens up more opportunities for fast-moving vehicles to overtake slow-moving vehicles.

The current Michigan vehicle type classification is regrouped into two general categories: fast- and slow-moving vehicles. The former includes passenger cars and pickup trucks; the latter consists of a variety of heavy trucks and passenger vehicles with 3 or more axles. In this exercise, only two-vehicle accidents containing fast- and slow-moving vehicles are selected. The systematic procedures developed in chapter 4 are employed to clean the accident data and assign the responsibility. With the aid of another variable—speed limit of the roadway where accident occurred (*speedlmt*), the accidents are further classified. D1s, D2s, and IRs for fast- and slow-moving vehicles under different speed limits are presented in table 8.1, with the quasi-induced exposure approach.

Table 8.1. D1s, D2s, and IRs for fast- and slow-moving vehicles

speed limit	fast-moving vehicles			slow-moving vehicles		
	D1 (%)	D2 (%)	IR	D1 (%)	D2 (%)	IR
<40	94.019	94.890	0.991	5.981	5.110	1.170
40	95.021	95.209	0.998	4.979	4.791	1.039
45	94.423	94.130	1.003	5.577	5.870	0.950
50	93.315	92.728	1.006	6.685	7.272	0.919
55	92.289	90.874	1.016	7.711	9.126	0.845
>55	91.719	89.868	1.021	8.281	10.132	0.817

The important information shown in table 8.1 is that for fast-moving vehicles, IRs go up as the speed limits go up; while for slow-moving vehicles, it is the opposite. For fast-moving vehicles, the result coincides with inequality 8.10—with lower speed limits (where speed variation is less obvious), the values of IRs are relatively smaller; with higher speed limits (where speed variation is more obvious), the values of IRs become

larger. Therefore, this phenomenon coincides with the results concluded from the theoretical exploration.

Other information displayed in table 8.1 is that the IR of slow-moving vehicles is sensitive to the change of D2s—IRs change from 1.17 to 0.82 with 5% change of D2s. It suggests that the percentage of slow-moving vehicles in the traffic stream (supposedly D2s for heavy trucks) is another essential factor affecting the IR, in addition to the speed variation. Based on the last two columns of table 8.1, it is predicted that when the percentage of slow-moving vehicles reaches 30%, the IR will be significantly reduced. This might help to explain the significant difference between the D2 and W2 data on I-94, which is particularly favored by heavy truck traffic; for I-96 and I-75 with less truck traffic composition, the difference is insignificant.

However, there is a caveat in comparing the patterns of D1s and D2s with inequalities 8.8 and 8.9. The problem is that the whole driving population is changed due to different roadways with different speed limits. Specifically, the traffic stream has a large percentage of heavy trucks (relatively slow-moving vehicles) as the speed limits go up (i.e., from local surface roads to freeways). Therefore, it is expected that D2 percentages for fast-moving vehicles will decrease due to the change in the driving population. It is hard to tell whether speed variation, change of driving population, or both contribute to the D2 decreasing for fast-moving vehicles as presented in table 8.1.

It has been shown both theoretically and practically that the speed variation between vehicles affects the relative accident involvement ratio. For vehicle types that typically travel faster, the IRs will be underrepresented; while for vehicle types with slower traveling speeds, the IRs will be overrepresented. Whether the IRs are

significantly biased by the speed variation between vehicles also hinges on the percentage of slow-moving vehicles in the traffic composition. The bias will become larger as the percentage increases. This is also intuitive: when heavy trucks account for 1% of the traffic, the errors introduced by the speed variation are negligible; when heavy trucks reach 30% (such as I-94) of the traffic, the errors are significant.

In summary, when the sample size is sufficiently large, quasi-induced exposure is a generally legitimate approach to measure the exposure on freeway links, although it is admitted that substantial speed variance between different vehicles and percentage of slow-moving vehicles compromise the measurement. When the composition of slow-moving vehicles reaches certain amount (say, 30%), significant amount of biases will be introduced to D1, D2, and IRs and thus, the quasi-induced exposure approach becomes inappropriate and other exposure measurements should be pursued.

As a closing note to this section, it needs to be pointed out that the research results can also be applied to the phenomenon whenever there are speed disparities in the traffic stream. For example, the speed discrepancy is caused by the different drivers (e.g., younger and faster drivers versus older and slower drivers).

Chapter 9

CONCLUSIONS

This research has attempted to address two basic issues: 1) validation of the assumptions of the quasi-induced exposure technique, essentially, that the non-responsible driver-vehicle combinations involved in two-vehicle accidents are representative of the driving population; and if validated, 2) present general rules/guidelines how and when to use quasi-induced exposure.

In the attempts to validate the underlying assumptions, two techniques have been developed, by comparing D2s in two-vehicle accidents with three types of "true" exposure data and the non-responsible drivers in three-or-more-vehicle accidents. The exposure data include VMT, safety-belt use data, and truck volume data (W-2).

VMT data by driver age and gender were generally approximated from NHTS in combination with census data for a certain state. The results show that quasi-induced exposure estimates are operationally and statistically different from VMT estimates for Michigan, Utah, and California. The VMT estimates are derived from NHTS in combination with census data. The NHTS is a national survey and collects data from a nationally representative sample of households at the national and state levels. Therefore, the travel estimates should be statistically reliable and representative of individual states. However, the VMT distributions are calculated based on the census data instead of the

driving population data (i.e., traffic volume data). The data are actually indicative of the distribution of census-population for each state. The assumption that driving population is characteristic of the census population at each state must not hold. Since the VMT data for each individual state, supposedly as an exposure "truth," can not be justified, the validity of D2 data can not be determined. Therefore, it is unconvincing that quasi-induced exposure is a legitimate approach to measure the relative exposure on the state-maintained routes as presented by the accident data. Further research efforts are necessary to derive more dependable and accurate VMT estimates.

Safety-belt data obtained from UMTRI were compared with D2 data at three levels: state, stratum, and county. Although the sampling plan for the safety-belt data collection was implemented at the statewide level, driver age and vehicle type distributions in the safety-belt data differ significantly from the D2 data. Data errors are introduced to the driver age in the safety-belt data, since it is relatively difficult to identify accurate driver age during the field observations, especially age group 30-59; for vehicle type, there might be some errors due to the mismatched classification—the passenger cars in accident data might include some sport utility vehicles. Practically, the comparisons between the safety-belt and D2 data demonstrate a fairly good agreement of driver gender distributions both at the statewide and intersections levels, from the operational perspective. At the state level, quasi-induced exposure is argued to be a generally good technique in terms of driver gender; however, further justifications are needed for driver age and vehicle type, since there are subject to data observation errors and mismatched vehicle classification.

At the stratum and county levels, the discussions have illustrated that there is a disagreement in driver-vehicle characteristic estimates between the safety-belt and D2 data for three key characteristics. Notwithstanding that, it has been observed that with the safety-belt and accident data at the more finely-disaggregated levels, the variation of distributions by three D2 characteristics becomes more significant. The D2 estimates yielded by quasi-induced exposure are relatively stable, while the distributions given by the safety-belt data vary considerably. Given that, quasi-induced exposure seems less sensitive to the data disaggregation, at the levels such as state, statewide intersections, stratum, and stratum-intersections.

W-2 data by vehicle types are compared with D2 data in terms of different freeway routes. Analytical results consistently demonstrate that there are significant differences between the W-2 and D2 data (with five vehicle type categories) for the five freeway routes examined. The sample size seems to play an important role in the results. With the sample size increased by regrouping the vehicle types into two categories, the results become mixed regarding whether the D2 data fit the W-2 data. For I-96 and I-75 there is no significant difference between the D2 and W-2 data; for US23 and US12 the difference is shown statistically insignificant; but for I-94 the difference is both statistically and operationally significant. With using the multiple years of data, notwithstanding that the difference for vehicle type distribution remains operationally and statistically significant for I-94 and I-75, the discrepancy is much smaller with large sample size. In general, the results suggest that quasi-induced exposure is sensitive to the sample size (data aggregation) when it is used for freeway routes.

While the sample size is an issue, a question is also raised regarding potential theoretical difficulties associated with the quasi-induced exposure technique, especially on freeways where there is heavy truck traffic and substantial speed variance. It has been theoretically and practically proved that the speed variation between vehicles affects the IRs. For vehicle types that typically travel faster, the IRs will be underrepresented; while for vehicle types with slower traveling speeds, the IRs will be overrepresented. Whether the IRs are significantly biased by the speed variation between vehicles also hinges on the percentage of slower vehicles in the traffic composition. The bias will become larger as the percentage of slower vehicles increases. Using I-94 as an example, the results show that the IR of slower vehicles is sensitive to the change of D2s (table 8.1). Based on the last two columns of table 8.1, it is predicted that when the percentage of slow-moving vehicles reaches 30%, the IR will be significantly reduced. This might help to explain the significant difference between the D2 and W2 data on I-94, which is particularly favored by heavy truck traffic; however, for I-96 and I-75 with less truck traffic composition, the difference is insignificant.

The other technique to validate the underlying assumptions of quasi-induced exposure is to use three-or-more-vehicle accidents. Accident data from California, Maine, Michigan, and Utah have demonstrated generally good agreements of D2 distributions between two- and three-or-more-vehicle accidents. A similar comparison is developed in three-or-more-vehicle accidents between D2's (the first non-responsible drivers) and D3s (the rest of the non-responsible drivers). The results also consistently show that there is agreement between D2's and D3s, which indicates that the responsible drivers in three-or-more-vehicle accidents "randomly select" the non-responsible drivers. Deductively, it

suggests that D2s in two-vehicle accidents are randomly impacted by D1s on the road. Overall, these observations imply that at the state level, the underlying assumptions of quasi-induced exposure are supported from the operational perspective. That is, quasi-induced exposure is a good technique to measure the exposure of specific driver-vehicle groups of interest at the state level.

Based upon the conducted comparisons, the guidelines are summarized as follows in terms of how and when to use quasi-induced exposure:

- Practically speaking, at the state level quasi-induced exposure is a generally good technique to measure the relative exposure for three driver-vehicle characteristics considered.
- In determining whether quasi-induced exposure can be used in a typical application, two comparisons with accident data can be conducted:
    - comparing D2 distributions between two-vehicle and three-or-more-vehicle accidents; and
    - comparing D2' and D3 distributions in three-or-more-vehicle accidents.
- If validated, two procedures can be employed to make accident data usable by quasi-induced exposure:
    - the procedure to clean accident data; and
    - the procedure to assign responsibility for accident faults.

In the meanwhile, the research also reveals findings and difficulties associated with quasi-induced exposure:

- Quasi-induced exposure is sensitive to the sample size (data aggregation). With data are disaggregated at more finely-disaggregated levels, the variation of

distributions by three D2 characteristics becomes more conspicuous/significant both from the operational and statistical senses.

- Under the circumstances when substantial speed variation exists between different vehicles, quasi-induced exposure results in the estimation of relative accident involvement ratio for passenger cars being comparatively lower and trucks being relatively higher.

Overall, the empirical approach to validate the underlying assumptions of quasi-induced exposure has shown some potential. However, during the validation process using a variety of accident and exposure data, it has been demonstrated that there are a number of problems inherently associated with the data themselves. Suggestions for further research in this field include:

- Require more sophisticated accident collection efforts to improve the quality of accident data. The research has shown that a significant amount of accident data is eliminated due to the data errors, when efforts are directed to clean the accident data. This can be improved with the aid of advanced data collection equipments on the accident scene (e.g., laptops equipped with GPS and GIS devices), systematic police training, and a more informative accident form.
- Improve the reliability and availability of exposure data. This research has been greatly compromised due to the quality of the exposure data, for example, the driver age in the safety-belt data. It would be desirable to conduct a comprehensive statewide traveler survey to collect vehicle miles traveled data by three key driver-vehicle characteristics under different circumstances (e.g., roadway type, time-of-day)

- Explore the effect of speed variation between young versus old drivers on the results of quasi-induced exposure. This could be the extension work of the discussion on the speed variation between passenger cars and trucks (chapter 5).

APPENDICES

APPENDIX A

C++ CODE TO MANIPULATE HSIS DATA

```
#include <stdlib.h>
#include <stdio.h>
#include <string.h>

static const int line_size = 512;

int getdelim (char **lineptr, size_t *n, int delim, FILE *stream)
{
 int indx = 0;
 int c;
 /* Sanity checks. */
 if (lineptr == NULL || n == NULL || stream == NULL)
  return -1;

 /* Allocate the line the first time. */
 if (*lineptr == NULL) {
  *lineptr = (char *) malloc (line_size);
  if (*lineptr == NULL)
   return -1;
        *n = line_size;
 }

 /* Clear the line. */
 memset (*lineptr, '\0', *n);
 do {
        c = fgetc (stream);
        if (c > 128 || c == 0) continue;
        if (c == EOF) {
                break;
        }
  /* Check if more memory is needed. */
  if (indx >= (*n) ) {
   *lineptr = (char *) realloc (*lineptr, *n + line_size);
   if (*lineptr == NULL) {
                return -1;
   }
```

```
    /* Clear the rest of the line. */
    memset(*lineptr + *n, '\0', line_size);
    *n += line_size;
        }
        /* Push the result in the line. */
   (*lineptr)[indx++] = c;
   /* Bail out. */
   if (c == delim) {
    break;
        }
 } while (1);
 return (c == EOF) ? -1 : indx;
}

int getline (char **lineptr, size_t *n, FILE *stream)
{
 return getdelim (lineptr, n, '\n', stream);
}

int main(int argn, char *argv[])
{
 FILE *inpfile, *oupfile1, *oupfile2;
 char *oldline = NULL, *newline = NULL;
 char oldline_copy[4096], newline_copy[4096], saveline[40][4096];
 size_t len;
 char delim[2], delim2[2];
 int dup = 0, i, j;
 char oupf1[32], oupf2[32];
 int read, ll;

 delim[0] = ' ';
 delim[1] = '\t';
 delim2[0] = '\n';
 delim2[0] = '\n';

 if (argn < 2) {
  printf ("input file name is needed\n");
  exit(1);
 }

 strcpy (oupf1, argv[1]);
 strcat (oupf1, ".out1");
 strcpy (oupf2, argv[1]);
 strcat (oupf2, ".out2");

 inpfile = fopen(argv[1], "r");
```

```
oupfile1 = fopen(oupf1, "w");
oupfile2 = fopen(oupf2, "w");

if (inpfile == NULL) {
 printf ("input file could not be opened\n");
 return 1;
}

getline (&oldline, &len, inpfile);
strcpy (oldline_copy, oldline);
strtok (oldline,delim);

while ((read = getline (&newline, &len, inpfile)) != -1) {
        ll = strlen (newline);
 strcpy (newline_copy, newline);
 strtok (newline,delim);

        if (dup == 0)
                strcpy (saveline[0], oldline_copy);

 if (strcmp (oldline, newline) == 0) {
         dup ++;
  strcpy (saveline[dup], newline_copy);
 }
 else {
                if (dup == 1){
                        for (i=0; i<=dup; i++){
                                for (j=0; j<strlen(saveline[i])-1; j++)
                                        fputc (saveline[i][j], oupfile1);
                                fputs ("\t", oupfile1);
                        }
                        fputs ("\n", oupfile1);
                }
                if (dup > 1) {
                        if (dup > 10)
                                printf ("%s\n", newline);
                        for (i=0; i<=dup; i++){
                                for (j=0; j<strlen(saveline[i])-1; j++)
                                        fputc (saveline[i][j], oupfile2);
                                fputs ("\t", oupfile2);
                        }
                        fputs ("\n", oupfile2);
                }
                dup = 0;
        }
```

```
        strcpy (oldline, newline);
        strcpy (oldline_copy, newline_copy);

  }

  fclose (inpfile);
  fclose (oupfile1);
  fclose (oupfile2);
}
```

APPENDIX B

COMPARISON BETWEEN D2 AND SEAT-BELT DATA AT STRATUM LEVEL (INTERSECTIONS)

Table B.1. Seat-belt versus D2 distributions for stratum 1—intersection

characteristics		seat-belt		D2		difference (%)
		N	%	N	%	
gender	male	1265	53.9	6269	50.4	3.5
	female	1080	46.1	6171	49.6	-3.5
age	16-29	652	27.8	3853	30.9	-3.1
	30-59	1474	62.8	7275	58.5	4.3
	60+	219	9.4	1312	10.6	-1.2
vehicle type	passenger car	1196	71.7	9124	83.8	-12.1
	pickup truck	473	28.3	1761	16.2	12.1

Table B.2. Seat-belt versus D2 distributions for stratum 2—intersection

characteristics		seat-belt		D2		difference (%)
		N	%	N	%	
gender	male	886	59.7	6290	52.7	7.0
	female	598	40.3	5640	47.3	-7.0
age	16-29	373	25.2	3777	31.6	-6.4
	30-59	948	63.8	6691	56.1	7.7
	60+	163	10.9	1462	12.3	-1.4
vehicle type	passenger car	748	71.3	8289	79.8	-8.5
	pickup truck	301	28.7	2103	20.2	8.5

Table B.3. Seat-belt versus D2 distributions for stratum 3—intersection

characteristics		seat-belt		D2		difference (%)
		N	%	N	%	
gender	male	552	56.0	4337	50.9	5.2
	female	433	44.0	4188	49.1	-5.2
age	16-29	245	24.9	2669	31.3	-6.4
	30-59	604	61.4	4709	55.2	6.2
	60+	135	13.7	1147	13.5	0.2
vehicle type	passenger car	446	60.1	5864	78.2	-18.1
	pickup truck	296	39.9	1637	21.8	18.1

Table B.4. Seat-belt versus D2 distributions for stratum 4—intersection

characteristics		seat-belt		D2		difference (%)
		N	%	N	%	
gender	male	2146	58.5	4616	51.6	6.9
	female	1524	41.5	4326	48.4	-6.9
age	16-29	1029	28.0	2704	30.2	-2.2
	30-59	2272	61.9	5065	56.6	5.2
	60+	368	10.1	1173	13.2	-3.1
vehicle type	passenger car	2180	81.3	6627	85.6	-4.3
	pickup truck	502	18.7	1111	14.4	4.3

APPENDIX C

COMPARISON BETWEEN D2 AND SEAT-BELT DATA AT COUNTY LEVEL

Table C.1. Seat-belt versus D2 distributions for county—male drivers

county	male driver				difference (%)
	seat-belt		D2		
	N	%	N	%	
Allegan	141	73.4	406	56.2	17.3
Bay	136	69.7	583	52.9	16.8
Berrien	157	51	606	50.9	0.1
Calhoun	28	45.9	571	49.7	-3.8
Eaton	89	66.9	458	48.2	18.7
Genesee	123	51.9	2091	49.9	2.0
Grand Traverse	47	57.3	591	52.4	4.9
Ingham	51	40.2	1937	50.8	-10.6
Jackson	149	58.2	715	52.9	5.3
Kalamazoo	634	57.4	1414	50.8	6.6
Kent	288	53.4	3503	53.3	0.1
Lapeer	117	58.8	355	56.0	2.8
Lenawee	47	68.1	401	53.6	14.5
Livingston	150	70.8	670	52.6	18.3
Macomb	299	57	4233	53.7	3.3
Monroe	85	65.4	466	49.5	15.9
Muskegon	186	62.2	800	51.5	10.7
Oakland	616	58.3	8040	52.1	6.2
Ottawa	29	76.3	1224	54.3	22.0
Saginaw	17	63	1169	52.1	10.9
Shiawassee	8	50	247	54.9	-4.9
St. Clair	3	30	669	52.5	-22.5
St. Joseph	60	72.3	165	50.5	21.8
Van Buren	160	58.6	241	53.4	5.2
Washtenaw	669	55.8	1793	52.2	3.7
Wayne	2686	59.8	9177	52.7	7.1

Table C.2. Seat-belt versus D2 distributions for county—age

county	16-29			30-59			60+		
	seat-belt	D2	diff. (%)	seat-belt	D2	diff. (%)	seat-belt	D2	diff. (%)
Allegan	17.7	30.2	-12.5	70.8	56.7	14.1	11.5	13.1	-1.6
Bay	27.7	29.4	-1.7	62.6	55.3	7.3	9.7	15.3	-5.6
Berrien	13.6	27.5	-13.9	69.2	56.7	12.5	17.2	15.8	1.4
Calhoun	29.5	33.9	-4.4	57.4	52.9	4.5	13.1	13.1	0
Eaton	16.5	29.7	-13.2	69.9	59.8	10.1	13.5	10.5	3
Genesee	27	30.6	-3.6	61.2	57.8	3.4	11.8	11.6	0.2
Grand Traverse	19.5	29.7	-10.2	64.6	55.6	9	15.9	14.6	1.3
Ingham	33.1	36.4	-3.3	48.8	54.2	-5.4	18.1	9.4	8.7
Jackson	25	29.5	-4.5	57	57.4	-0.4	18	13.1	4.9
Kalamazoo	34.9	37.7	-2.8	57.1	51.5	5.6	8.0	10.8	-2.8
Kent	28	34.5	-6.5	64.4	56.2	8.2	7.6	9.3	-1.7
Lapeer	35.2	34.4	0.8	50.3	54.4	-4.1	14.1	11.2	2.9
Lenawee	24.6	31.8	-7.2	49.3	52.8	-3.5	26.1	15.4	10.7
Livingston	26.4	30.5	-4.1	64.2	58.4	5.8	9.4	11.1	-1.7
Macomb	31.8	29.8	2	59.8	57.9	1.9	8.4	12.3	-3.9
Monroe	27.7	33.6	-5.9	58.5	56.0	2.5	13.1	10.4	2.7
Muskegon	30.1	30.1	0	61.5	55.9	5.6	8.4	14	-5.6
Oakland	26.7	28.9	-2.2	64.8	61.2	3.6	8.5	9.9	-1.4
Ottawa	31.6	35.7	-4.1	52.6	52.9	-0.3	15.8	11.4	4.4
Saginaw	29.6	31.8	-2.2	55.6	54.4	1.2	14.8	13.8	1
Shiawassee	37.5	32.4	5.1	43.8	54.7	-10.9	18.8	12.9	5.9
St. Clair	30.0	31.4	-1.4	50.0	54.9	-4.9	20.0	13.7	6.3
St. Joseph	18.1	31.5	-13.4	77.1	53.8	23.3	4.8	14.7	-9.9
Van Buren	20.9	26.8	-5.9	69.2	57.0	12.2	9.5	16.2	-6.7
Washtenaw	30.3	33.0	-2.7	62.4	58.9	3.5	7.3	8.1	-0.8
Wayne	28	28.9	-0.9	62.6	58.6	4.0	9.4	12.5	-3.1

Table C.3. Seat-belt versus D2 distributions for county—passenger car

county	passenger car				difference (%)
	seat-belt		D2		
	N	%	N	%	
Allegan	86	72.3	450	75.6	-3.3
Bay	87	78.4	700	78.7	-0.3
Berrien	155	82.0	893	85.9	-3.9
Calhoun	40	88.9	789	83.1	5.8
Eaton	61	82.4	632	79.0	3.4
Genesee	112	82.4	2814	79.6	2.8
Grand Traverse	39	70.9	758	78.8	-7.9
Ingham	73	79.3	2749	85.6	-6.3
Jackson	133	80.6	866	77.7	2.9
Kalamazoo	615	83.6	2035	84.4	-0.8
Kent	313	80.7	4828	85.5	-4.8
Lapeer	82	82.0	370	69.0	13.0
Lenawee	29	87.9	470	76.7	11.2
Livingston	85	66.9	786	75.6	-8.7
Macomb	271	77.7	5319	81.6	-3.9
Monroe	49	76.6	604	75.7	0.9
Muskegon	164	88.2	1051	81.1	7.1
Oakland	541	72.1	11152	86.0	-13.9
Ottawa	17	85.0	1580	82.5	2.5
Saginaw	9	100.0	1617	83.4	16.6
Shiawassee	7	87.5	281	74.9	12.6
St. Clair	5	83.3	848	78.5	4.8
St. Joseph	34	72.3	199	74.3	-2.0
Van Buren	135	76.7	288	77.2	-0.5
Washtenaw	626	80.4	2422	85.3	-4.9
Wayne	2626	81.5	12620	86.7	-5.2

BIBLIOGRAPHY

Aldridge, Brian, Himmler, Meredith, et al. (1999). Impact of passengers on young driver safety. Transportation Research Record 1693: 25-30.

Baker, R.F. (1971). The Highway Risk Problem – Policy Issues on Highway Safety. Wiley, New York.

Baker, R. and Nedler, J. (1987). GLIM Manual. Royal Statistical Society, Oxford, England.

Bishop, Y., et al. (1975). Discrete multivariate analysis: theory and practice. MIT press, Cambridge, Mass.

Brodsky, H. and Hakkert, A.S. (1983). Highway accident rates and rural travel densities. Accident Analysis and Prevention, 15(1): 73-84.

Brog, W. and Kuffner, B. (1981). Relationship of accident frequency to travel exposure. Transportation Research, 808: 55-61.

Brown, I.D. (1982). Exposure and experience are a confounded nuisance in research on driver behavior. Accident Analysis and Prevention, 14(5): 345-352.

Campbell, R. and Filmer, R.J. (1981). Deaths from Road Accidents in Australian. 10th Conf. Of Economists, Canberra, Australia.

Carlson, W.L. (1970). Induced exposure revisited. HIT Lab Rep. University of Michigan, Ann Arbor.

Carr, B.R. (1969). A statistical analysis or rural Ontario traffic accidents using induced exposure data. Accident Analysis and Prevention, 1: 343-358.

Carreker, L.E. et al. (2000). Geographic information system procedures to improve speed and accuracy in locating crashes. Transportation Research Board, 1719: 215-218.

Carroll, P.S. (1971). The Meaning of Driving Exposure. HIT-lab Rep. (April). Highway Safety Research Institute, the University of Michigan, Ann Arbor.

Carroll, P. S. (1973). Classifications of driving exposure and accident rates for highway safety analysis. Accident Analysis and Prevention, 5: 81-94.

Ceder, A., et al. (1982a). Relationships between road accidents and hourly traffic flow – I, probabilistic approach. Accident Analysis and Prevention, 14(1): 20-34.

Ceder, A., et al. (1982b). Relationships between road accidents and hourly traffic flow – II, probabilistic approach. Accident Analysis and Prevention, 14(1): 35-44.

Cerrelli, E.C. (1973). Driver exposure – the indirect approach for obtaining relative measures. Accident Analysis and Prevention, 5: 147-156.

Chandraratna, S., et al. (2003). Problem driving maneuvers of older drivers. Presented in the 2003 Transportation Research Board Meeting, Washington, D.C.

Chang, Myung-Soon. (1982). Conceptual development of exposure measures for evaluating highway safety. Transportation Research Record, 847: 37-42.

Chapman, R.A. (1967). Traffic collision exposure. New Zealand Road Symp. 1:184-498.

Chapman, R.A. (1973). The concept of exposure. Accident Analysis and Prevention, 5(2): 95-110.

Chipman, M. (1983). Motor vehicle accident fatality statistics: an investigation of reliability. Canadian Journal of Public Health, 74(6): 333-349.

Chipman, M. L. et al. (1993). The role of exposure in comparisons of crash risk among different drivers and driving environments. Accident Analysis and Prevention, 25(2): 207-211.

Chira-Chavala, T. and Cleveland, D. (1986). Causal analysis of accident involvements for the nation's large trucks and combination vehicles. Transportation Research Record 1047: 56-64.

Cirillo, J.A. (1983). Limitation of the current NASS system as related to FHWA accident research. Transportation Research Circular, 256: 20-21.

Cooper, D.F. and Ferguson, N. (1976). Traffic studies at T-junctions. Traffic Engineering control, 17(7):306-309.

David, J., DeYong, Raymond C., et al. (1997). Estimating the exposure and fatal crash rates of suspended/revoked and unlicensed drivers in California. Accident Analysis and Prevention, 29(1): 17-23.

De Silva, H.R. (1942). Why We Have Automobile Accidents. Wiley, New York.

Dunlap and Associates. (1953). An Analysis of Risk and Exposure in Automobile Accident. Prepared for the Office of the Surgeon General, Department of the Army.

Erlander, S., Gustavsson, J., and Larusson, E. (1969). Some Investigations on the relationship between road accidents and estimated traffic. Accident Analysis and Prevention, 1(1): 17-65.

Erlbaum, N. S. (1989). Estimated county level vehicle miles of travel. Data Services Bureau Report. Planning Division, New York State DOT, Albany, April.

Federal Highway Administration. (1977). Evaluation of the highway related safety problem standards. U.S. Department of Transportation, Washington D.C.

Federal Highway Administration. (1997). Safety data: costs, quality, and strategies for improvement. Executive summary. FHWA-RD-96-027, 10p.

Ferlis, R. A. and Larry A. Bowman. (1981). Procedures for measuring regional VMT. Transportation Research Record 815: 1-6.

Foldvary, L.A. (1969a). Road accident rates by number of vehicles on the register and miles of travel. Australian Road Research 3(9): 72-89.

Foldvary, L.A. (1969b). Methodological lessons drawn from a vehicle-mile survey. Accident Analysis and Prevention, 1(3): 223-244.

Goeller, B.F. (1968). Modeling the traffic safety system. Accident Analysis and Prevention, 1: 167-204.

Goldstine, R. (1991). Influence of road width on accident rates by traffic volume. Transportation Research Record 1318: 64-69.

Golias, J. and Yannis, G. (2001). Dealing with lack of exposure data in road accident analysis". Proceedings of the 12th International Conference on Traffic Safety on Three Continents, Moscow, 2001.

Greeblatt, J., M. et al. (1981). National accident sampling system nonreported accident survey. U.S. Department of Transportation, National Highway Traffic Safety Administration DOT HS-806198.

Grossman, L. (1954). Accident-exposure index. Proceedings of Highway Research Board 33: 129-138.

Hadi, M.A., Aruldhas, J., Chow, L.F., Wattleworth, J.A. (1993). Estimating Safety Effects of Cross-Section Design for Various Highway Types Using Negative Binomial Regression. Transportation Research Record, 1500: 169–177.

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

Haight, F.A. (1971). Indirect methods for measuring exposure factors as related to the incidence of motor vehicle traffic accidents. Prepared for National Highway Traffic Safety Administration. Pennsylvania Transportation and Traffic Safety Center, University Park.

Haight, F.A. (1973). Induced exposure. Accident Analysis and Prevention, 5: 111-126.

Hakkert, A.S. and D. Mahelel. (1978). Estimating the number of accidents at intersections from a knowledge of the traffic flow on the approaches. Accident Analysis and Prevention, 10(1): 69-79.

Hakkert A. S., et al. (1978). The reliability of reporting of road accidents with casualties. RSC report 77/17. Road Safety Center, Israel.

Hall, W.K. (1970). An empirical analysis of accident data using induced exposure. HIT Lab Rep. University of Michigan, Ann Arbor.

Hammer, C. G. (1969). Evaluation of minor improvements. Highway Research Record 286: 33-45.

Hampson, G. (1982). The theory of accident compensation and the introduction of compulsory seat belt legislation in new south Wales. Proc. A.R.R.B. Conf., 11(5): 135-140.

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

Hauer, E. (2001). Computing and interpreting accident rates for vehicle types or driver groups. Transportation Research Record 1746: 69-73.

Hauer, E. and A. S. Hakkert. (1988). Extent and some implication of incomplete accident reporting. Transportation Research Record 1185: 1-10.

Hautzinger, H., et al. (1985). Accuracy of official road accident statistics. Internationals Verkehrswesen, 37(4).

Heimbach, C. L., et al. (1980). The effect of reduced traffic lane width on traffic operations and safety for urban undivided arterials in North Carolina. Highway research program, North Carolina State University, Raleigh.

Highway Performance Monitoring System Field Manual. (1995). FHWA, U.S. Department of Transportation, February.

Hing, J. Y., et al. (2003). Evaluating the impact of passengers on the safety of older drivers. Journal of Safety Research, 34: 343-351.

Hodge, G. A. and A. J. Richardson. (1985). The role of accident exposure in transport system safety evaluation I: intersection and link site exposure. Journal of Advanced Transportation 19(2): 179-213.

Hodge, G. A. and Richardson, A. J. (1978). A study of intersection accident exposure. Proc. A.R.R.B. Conf., 9(5): 151-160.

Hodge, G.A. and Richardson, A. J. (1985). Role of accident exposure in transport system safety evaluations II: group exposure and induced exposure. Journal of Advanced Transportation, 19(2): 201-213.

http://www.michigan.gov/sos/1,1607,7-127-1627_8669_8998-22931--,00.html

Ivan, J.N., and O'Mara, P.J. (1997). Prediction of traffic accident rates using Poisson regression. Presented in the 1997 Transportation Research Board Meeting, Washington, D.C.

Jacobs, H.H. (1961). Conceptual and Methodological Problems in Accident Research. Behavioral Approaches to Accident Research. Association for the Aid of Crippled Children, New York.

Joksch, H. C. (1973). A pilot study of observed and induced exposure to traffic accident. Accident Analysis and Prevention: 5(2): 127-136.

Jovanis, P. P., et al. (1991). Exploratory analysis of motor carrier accident risk and daily driving patterns. Transportation Research Record 1322: 34-43.

Karlaftis, M.G. and Golias, I. (2002). Effects of road geometry and traffic volumes on rural roadway accident rates. Accident Analysis and Prevention, 34(3): 357-365.

Karlaftis, M.G., and Tarko, A. (1998). Heterogeneity considerations in accident modeling. Accident Analysis and Prevention 30 (4): 425–433.

Keall, M. D. (1995). Pedestrian exposure to risk of road accident in New Zealand. Accident Analysis and Prevention 27(5): 729-740.

Keller, H. H. (1985). Traffic records in Germany: traffic record systems forum. Pergamon Press, Elmsford, New York.

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

Kirk, Adam and Stamatiadis, N. (2001b). Evaluation of quasi-induced exposure. Final Report, College of Engineering, University of Kentucky.

Klein, D. and Waller, J.A. (1970). Causation, Culpability and Deterrence in Highway Crashes. Prepared for the Department of Transportation, Automobile Insurance and Compensation Study.

Knuiman, M. W., Council, F. M. and Reinfurt, D. W. (1993). The effect of median width on highway accident rates. Transportation Research Record 1401: 70-80.

Koonstra, M.J. (1973). A model for estimation of collective exposure and proneness from accident data. Accident Analysis and Prevention, 5: 157-173.

Kuroda, K., et. al. (1985). The impact of vehicle size on highway safety. Presented at the 85th Transportation Research Board Annual Meeting, January 1985.

Kuzmyak, J. R. (1981). Usefulness of existing travel data sets for improved VMT forecasting. Transportation Research Record, 815: 7-12.

Lalani, N. and Walker, D. (1981). Correlating accidents and volumes at intersections and on urban arterial street segments. Traffic Engineering and Control, 22:359-363.

Leong, J. W. (1975). Relationship between accidents and traffic volumes at urban intersections. Australian Road Research, 5(3): 72-99.

Lighthizer, Dale Reed. (1989). An empirical validation of quasi-induced exposure. Ph.d Dissertation, Department of Civil and Environmental Engineering, Michigan state University, E. Lansing, MI.

Lock, J. B. (1966). Age and sex in relation to road traffic accidents. Proc. Australian Road Research, 3(1): 623-637.

Lyles, R.W., et al. (1994). Quasi-induced exposure: to use or not to use? Presented at the 94th Transportation Research Board Annual Meeting, January 1994.

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

Lyles, R. W., W. C. Taylor, et. al. (1993). Methods for estimating truck exposure. Final report. Department of civil and environmental engineering, Michigan State University, East Lansing, Michigan.

Maleck, T. and Hummer, J.E. (1987). Driver age and highway safety. Transportation Research Record, 1059: 6-12.

Mengert, P. (1982). Literature review on induced exposure models. Transportation Systems Center, Cambridge, Mass., Report No. PM-223-U5-4A, 46 p.

McKnight , A. J. (1981). Accident data: a limited tool for evaluation. Transportation Research Circular, 233: 12-13.

Mountain, L., Fawaz, B. (1991). The accuracy of estimates of expected accident frequencies obtained using an empirical bayes approach. Traffic Engineering and Control, 32(5): 246-251.

Nembhard, David A., and Young, M.R. (1995). Parametric empirical bayes estimates of truck accident rates. Journal of Transportation Engineering, 121(4): 359-363.

O'Day, J. (1993). Accident data quality. NCHRP Synthesis of Highway Practice, 192, 54p.

Operations Research Inc. (1971). Development of Highway Safety Statistical Indicators. Prepared for U.S. Dept. of Transportation (DOT HS-800 472) by Operations Research Inc, 1400 Spring Street, Silver Spring (Contact FH-11-7505) MD 20910.

Perry, D.R. and Callaghan, B.E. (1980). Some comparisons of direct and indirect calculations of exposure and accident involvement rates. Proc. 10t A.R.R.B. Conf., 10(4): 263-274.

Philipson, L.L., et al. (1988). Statistical analysis of highway commercial vehicle accidents. Accident Analysis and Prevention, 13: 289-306.

Philipson, L.L., Rashti, P. and Fleischer, G.A. (1981). Statistical analysis of highway commercial vehicle accidents. Accident Analysis and Prevention, 13(4): 289-306.

Peleg, M. (1967). Evaluation of the conflict hazard of uncontrolled junctions. Traffic Engineering and Control: 346-347.

Raff, M. S. (1963). Interstate highway accident study. Highway Research Board Bulletin 74: 18-45.

Risk, A. and Shaoul, J.E. (1982). Exposure to risk and the risk of exposure. Accident Analysis and Prevention, 14(5): 353-357.

Saccomanno, F.F., Buyco, C. (1988). Generalized loglinear models of truck accident rates. Transportation Research Record 1172: 23-31.

Satterhwaite, S.P. (1981). A survey of research into relationships between traffic accidents and traffic volumes. T.R.R.L., S. R. 692.

Scott, R. E. and P. S. Carroll. (1971). Acquisition of information on exposure and on non-fatal crashes, volume II: accident data inaccuracies. Highway Safety Research Institute, University of Michigan, May.

Shaw, L. and Sichel, H.S. (1971). Accident Proneness. Pergamon Press, Oxford.

Shinar, D., J. R. Treat and S. T. McDonald. (1983). The validity of police reported accident data. Accident Analysis and Prevention, 15(3): 175-191.

Shinar, D., et al. (1983). The valid of police reported accident data. Accident Analysis and Prevention, 15(3): 175-191.

Smeed, R.J. (1955). Accident rates. Int. Road Safety Traffic Review, 3(2): 30-40.

Smeed, R.J. (1968). Variations in the Pattern of Accident Rates in Different Countries and Their Causes. General Report, Theme IV, Ninth Int. Study Week in Traffic and Safety Engineering. Munich.

Smith, R. N. (1966). The reporting level of California state highway accidents. Traffic Engineering Journal, June.

Squires, C.A., and Parsonson, P.S. (1989). Accident comparison of raised median and two-way left-turn lane median treatments. Transportation Research Record 1239: 30-40.

Stamatiadis, N., Deacon, J.A . (1997). Quasi-induced exposure: methodology and insight. Accident Analysis and Prevention, 29(1): 37-52.

Stamatiadis, N. and Puccini, Giovanni. (1998). Fatal crash rates in the southeastern united states – why are they higher? Transportation Research Record 1665: 118-124.

Steward, R.G. (1960). Driving exposure – what does it mean? How is it measured? Traffic Safety Res. Rev. 4(2): 9-11.

Surti, V. H. (1965). Accident exposure for at-grade intersections. Traffic Engineering 36(3): 26-27 and 53.

Tanner, J.C. (1953). Accidents at rural three-way junctions. J. Inst. Highw. Eng. 2(11): 56-67.

Taylor, W.C. and T. Maleck. (1987). Accident datafile accuracy. Michigan State University, College of Engineering, East Lansing, 10 p.

Thorpe, J.D. (1964). Calculating relative involvement rates in accident without determining exposure. Australian Road Research, 2: 25-36.

Thorpe, J.D. (1968). Accident rates at signalized intersections. Proc. A.R.R.B. Conf. 4(1).

Transportation Research Circular 231. (1981). Truck-accident data systems: state-of-the-art report. Transportation Research Board, 8 p.

Underwood, R.T. (1966). Low volume rural Y- junctions. Aust. Rd. Res. 2(8): 3-23.

University of California. (1975). Exposure/involvement rates. Appendix H, truck/bus accident study, December.

Vey, A.H. (1937). Relationship between daily traffic and accident rates. American City 52(9): 119.

Waller, P.F., Reifurt, D.W., Freeman, J.L. and Imrey, P.B. (1974). Method for measuring exposure to automobile accidents. Presented at the 101st Annual Meeting of the American Public Health Assoc.

Wass, C. (1977). Traffic Accident Exposure and Liability. ROC Runsted: Allerod.

Wilbur Smith and Associates. (1966). A report on the Washington area motor vehicle accident cost study. U.S. Bureau of Public Road report PB 173859.

Willett, P. (1977). Perth's experience with pelican crossings. Aust. Rd. Res. 7(4): 36.

Willis, C. O., et al. (1983). Evaluation of accident reporting histories. Transportation Research Record, 910: 19-26.

Wright, P. and Burnarn, A. (1985). Roadway features and truck safety. Transportation Planning and Technology, 9: 42-51.

Zegeer, C.V. (1982). NCHRP synthesis of highway practices 91: highway accident analysis systems. TRB, National Research Council, Washington D.C.