LINKING OREGON DRIVER RECORDS AND CRASH DATA TO EVALUATE INTERVENTIONS AND MITIGATE DRIVER RISK

Final Report
TR-24-24-11-00

by

David Hurwitz, Ph.D., Professor
Hisham Jashami, Ph.D., RSP1, Assistant Professor (Sr Res)
Aiden Gray, Undergraduate Research Assistant
Oregon State University
101 Kearney Hall
Corvallis, OR 97331

for

Oregon Department of Transportation
Research Section
555 13th Street NE, Suite 1
Salem OR 97301

and

Federal Highway Administration
1200 New Jersey Avenue SE
Washington, DC 20590

February 2026

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
OR-RD-26-01		

4. Title and Subtitle	5. Report Date
Linking Oregon Driver Records and Crash Data to Evaluate Interventions and Mitigate Driver Risk	February 2026
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
David Hurwitz 0000-0001-8450-6516 Hisham Jashami 0000-0002-5511-7543 Aiden Gray 0009-0002-3900-8504	TR-24-24-11-00

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
Oregon Department of Transportation Research Section 555 13th Street NE Salem, OR 97301	11. Contract or Grant No. 30530, 24-07

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
National Highway Traffic Safety Administration 1200 New Jersey Ave., SE, West Building Washington, DC 20590	Final Report
	14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

ODOT seeks to reduce the number of risky drivers on the road through driver improvement programs, but the efficacy of these programs are rarely assessed. This report details the methodology for creating a linked database combining driver, verdict, accident, and crash data from ODOT to evaluate program performance and trends in risky driver behaviors. The linked database was used to generate visualizations of citation and crash data between the years 1995-2024, which were used to identify demographic trends and shifts in administrative procedures. In addition to visualization, several models were built to predict the probability of different crash types based on a variety of demographic factors. These models can help identify which groups are the most likely to engage in risky driving, knowledge which can tailor future intervention strategies.

17. Key Words	18. Distribution Statement
Crash, Accident, Verdict, License, Merged Crash Data, Risky Driver	Copies available from NTIS, and online at www.oregon.gov/ODOT/TD/TP_RES/

19. Security Classification (of this report)	20. Security Classification (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	81	

SI* (MODERN METRIC) CONVERSION FACTORS

APPROXIMATE CONVERSIONS TO SI UNITS

Symbol	When You Know	Multiply By	To Find	Symbol
		LENGTH		
in	inches	25.4	millimeters	mm
ft	feet	0.305	meters	m
yd	yards	0.914	meters	m
mi	miles	1.61	kilometers	km
		AREA		
in^2	square inches	645.2	millimeters squared	mm^2
ft^2	square feet	0.093	meters squared	m^2
yd^2	square yards	0.836	meters squared	m^2
ac	acres	0.405	hectares	ha
mi^2	square miles	2.59	kilometers squared	km^2
		VOLUME		
fl oz	fluid ounces	29.57	milliliters	ml
gal	gallons	3.785	liters	L
ft^3	cubic feet	0.028	meters cubed	m^3
yd^3	cubic yards	0.765	meters cubed	m^3
~NOTE: Volumes greater than 1000 L shall be shown in m^3.				
		MASS		
oz	ounces	28.35	grams	g
lb	pounds	0.454	kilograms	kg
T	short tons (2000 lb)	0.907	megagrams	Mg
		TEMPERATURE (exact)		
°F	Fahrenheit	(F-32)/1.8	Celsius	°C

APPROXIMATE CONVERSIONS FROM SI UNITS

Symbol	When You Know	Multiply By	To Find	Symbol
		LENGTH		
mm	millimeters	0.039	inches	in
m	meters	3.28	feet	ft
m	meters	1.09	yards	yd
km	kilometers	0.621	miles	mi
		AREA		
mm^2	millimeters squared	0.0016	square inches	in^2
m^2	meters squared	10.764	square feet	ft^2
m^2	meters squared	1.196	square yards	yd^2
ha	hectares	2.47	acres	ac
km^2	kilometers squared	0.386	square miles	mi^2
		VOLUME		
ml	milliliters	0.034	fluid ounces	fl oz
L	liters	0.264	gallons	gal
m^3	meters cubed	35.315	cubic feet	ft^3
m^3	meters cubed	1.308	cubic yards	yd^3
		MASS		
g	grams	0.035	ounces	oz
kg	kilograms	2.205	pounds	lb
Mg	megagrams	1.102	short tons (2000 lb)	T
		TEMPERATURE (exact)		
°C	Celsius	1.8C+32	Fahrenheit	°F

*SI is the symbol for the International System of Measurement

ACKNOWLEDGEMENTS

The authors thank the Oregon Department of Transportation (ODOT) and the Federal Highway Administration (FHWA) for funding this research project. The authors would also like to thank Josh Roll, ODOT Research Coordinator, and the ODOT Technical Advisory Committee (Stephanie Milton, Johnathan Munson, Tracy Pearl, Jody Raska, Ryan Stone, Vanessa Churchill, Tiana Tozer, Karen Ofearna, Zijia Zhong, Nicole Charlson, Christina McDaniel-Wilson, Peter Geissert) for providing valuable input throughout the project. The authors also thank Oregon State University undergraduate students Mahde Abusaleh, Clarence Fernando, Charles Tuckfield, Brandon Walker, for their assistance with a variety of validation tasks.

DISCLAIMER

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

In 2022, there were an estimated 45,070 reported collisions in the state of Oregon, with 554 of those representing fatal collisions resulting in 603 persons killed and 36,950 persons injured (ODOT, 2024). Risky driving behavior is widely recognized as a contributor to fatal collisions, which is why the reduction of risky driver behavior is one of the near-term emphasis areas within the ODOT Transportation Safety Action Plan (ODOT, 2021).

To this end, many states have implemented programs that identify drivers who display heightened levels of collision risk from the roadway and intervene through state mandated programs. These programs often include a blend of license restriction and education requirements; however, the efficacy of these programs is studied infrequently. In the state of Oregon, it has been decades since the last intervention program study.

This report documents the process of building a dataset linking driver information and collision outcomes with the goal of providing the Oregon DMV a tool that is capable of evaluating traffic offender programs and providing insight on the profiles of risky drivers that can be used to improve safety outcomes.

2.0 LITERATURE REVIEW

This chapter contains an overview of published literature and policy relating to habitual traffic offender programs currently operating in the United States. It contains a history and description of the current program in the state of Oregon, a comparison of Oregon's program to similar state-level programs in close proximity to Oregon, the efficacy of existing habitual offender programs, descriptions of existing traffic safety data in Oregon, and a synthesis of methodologies for combining disparate sources of traffic safety data.

2.1 OREGON'S RISKY DRIVER DIVERSION PROGRAMS

Risky driver diversion programs are a class of programs in Oregon that aim to reduce the frequency of crashes caused by repeat traffic offenders by removing them from the road until they are deemed safe to return. The purpose of this type of program is to deter unsafe behavior and ultimately save lives. This report focuses on four of the larger programs in the state of Oregon: the Habitual Traffic Offender Program, Driver Improvement Program, DUII Diversion Program, and the At-Risk Driver Program.

2.1.1 History of the Oregon Habitual Traffic Offender Program

The HTO program for the state of Oregon was started in 1974, and the last known evaluation of the program occurred approximately a decade later in 1986 (Jones, 1986). During the period of time from 1974 - 1984, the penalty following a prosecution was a 10-year license revocation. However, the original program was plagued by prosecution issues, as local authorities charged with administering the program did not apply the law uniformly.

In 1984, there was an amendment to Oregon law that reduced the penalty of the program to a 5-year revocation and transferred the revocation authority from local counties to the Oregon Motor Vehicles Division. This reduced variability in the application of the law and allowed the Motor Vehicles Division to issue revocations directly. The primary critique of the new system involved the low rates of delivery for revocation notices. In 1986, it was found that there existed a delivery rate of 47%, which was suspected of reducing the efficacy of the program (Jones, 1986). However, the study conducted by Dr. Jones in 1986 concluded that the program contributed to the prevention of crashes despite the low rates of delivery.

At the time of the last evaluation, an HTO was defined as someone who accumulated three major traffic offenses within a five-year time period. Major traffic offenses included "DUII, driving while suspended or revoked, reckless driving, "hit-and-run", eluding, and assorted violations involving assault, manslaughter or murder with a motor vehicle" (Jones, 1986). After the second offense, the driver was sent a letter warning them that another offense would designate them a HTO as well as resources such as advisory meetings and educational programs that were available.

2.1.2 Current HTO Program

The current habitual traffic offender program in Oregon is run by the Oregon Driver and Motor Vehicles Services (DMV). A habitual traffic offender is defined by Oregon law as anyone convicted of three or more of the following outlined offenses or more than twenty traffic violations in the span of five years (Oregon DMV).

The current offenses itemized by the Oregon DMV include:

- Any degree of murder, manslaughter, criminally negligent homicide, assault, recklessly endangering another person, menacing or criminal mischief resulting from the operation of a motor vehicle,
- Driving while under the influence of intoxicants,
- Driving while your license is suspended or revoked,
- Reckless driving,
- Failure to perform the duties of a driver after a collision that results in injury, and
- Fleeing or attempting to elude a police officer.

Traffic violations as defined by the Oregon DMV can be found in OAR 735-064-0220(2)(3) (Oregon DMV). Examples include, but are not limited to, abandoning a vehicle, careless driving, failure to drive on right, passing in a no-passing zone, failure to yield right-of-way, violating a speed limit, unsafe passing, and unlawful stop.

The penalty of the program is a five-year revocation, which is dispersed by a notice mailed to the address on file. The notice contains the effective date and time of the revocation, as well as instructions for surrendering the revoked license, the HTO's right to a hearing, and eligibility to apply for hardship permits.

For certain offenses, the length of revocation is greater than that outlined in the HTO program. One offense includes failure to perform the duties of a driver in the case of a fatality, which earns the driver a minimum revocation of five years. For offenses such as aggravated vehicular homicide, criminally negligent homicide, manslaughter to the 1st and 2nd degree, and murder to any degree, the punishment is permanent revocation (ODOT 2022).

2.1.3 The Oregon Driver Improvement Program

Parallel to the current HTO program, the state of Oregon also manages the Driver Improvement Program (DIP) which is also aimed at removing unsafe drivers from the road. This program is split into two divisions, the Provisional DIP and the Adult DIP, as shown in Table 2.1.

Table 2.1 Summary of Driver Intervention Programs in Oregon (excluding DUII programs)

Program	Age	Interventions	Duration of Intervention
HTO	No limit	Revocation	5 years
Provisional DIP	>14 years, <18 years	Suspension and restriction	90 day restriction 6 month to 1 year suspension
Adult DIP	18 years and older	Suspension and restriction	30 day restriction 30+ day suspension

2.1.3.1 Provisional DIP

The Provisional DIP involves drivers between the ages of 14 and 18 years old. In addition to the restrictions of holding a provisional license, the DMV will restrict a driver's license to only work-related travel for 90 days if they accumulate the following:

- Two convictions,
- Two preventable accidents, and
- A combination of one conviction and one preventable accident (Oregon DMV).

A conviction is defined as "determination of guilt by a court of law upon a plea, verdict, finding, or unvacated bail forfeiture," and a preventable accident is defined as "a traffic accident reported by a police officer that indicates a driver failed to do everything a driver reasonably could have done to prevent the accident" in OAR 735-072-0020(5) (Oregon DMV).

If the driver violates these restrictions once, they are at risk of license suspension. If the driver accumulates a third conviction or preventable accident while in the program, they automatically receive a six-month suspension. And for every conviction after, the driver will receive a six-month suspension. If the driver commits any of the offenses outlined in the HTO program, the DMV will suspend driving privileges for a year (ODOT 2022).

2.1.3.2 Adult DIP

The Adult DIP involves drivers over the age of 18. The main interventions of the adult DIP are still restriction and suspension, not revocation. This distinction separates it from the HTO program.

The DMV will place a restriction on a license for 30 days if a driver has accumulated the following offenses over a two-year period: three convictions, three preventable accidents, or any combination of the three.

Restrictions involve not allowing the driver to drive between the hours of 12:00 am and 5:00 am unless driving to a place of employment or residence. The restrictions begin 30 days from the date the notice was received, and the driver must carry their restriction letter in their vehicle at all times (OAR 735-072-0027)

The DMV will suspend a driver's license for 30 days if they accumulate five offenses as outlined in OAR 735-072-0041 within a two-year period. For every additional violation past five within two years, the suspension is extended by an additional 30 days. Drivers are eligible for hardship permits during the suspension period, which allows adult drivers to continue to drive to work, to seek medical care, and to fulfill essential functions such as grocery shopping (ODOT 2022).

2.1.4 The Oregon Driving Under the Influence of Intoxicants Program

There are two ways that a driver can have their license suspended for driving under the influence of intoxicants (DUII) in the state of Oregon. Oregon DUII Offenses and Interventions are summarized in Table 2.2. The first is through the application of the Implied Consent law, and the other is through a court conviction for DUII. Intoxicants are defined in Oregon law as intoxicating liquor, cannabis, psilocybin, a controlled substance, an inhalant, or any drug, as defined in ORS 475.005 that, when used either alone or in combination with intoxicating liquor, an inhalant, psilocybin, cannabis or a controlled substance, adversely affects a person's mental or physical faculties to a noticeable or perceptible degree" (ORS Section 801.321). The legal limit for blood alcohol content (BAC) in the state of Oregon is set at 0.08%. A BAC above 0.08% indicates that a motorist is driving while legally intoxicated.

Table 2.2 Summary of DUII Offenses and Interventions

Program	Offense	Intervention	Duration of Intervention
Implied Consent	Failing a breath test	Suspension	90 days or 1 year
Implied Consent	Refusing to take a breath test	Suspension	1 year or 3 years
Implied Consent	Refusing to take a urine test	Suspension	1 year or 3 years
Conviction	Class A Misdemeanor *Second offense	Jail Suspension IID Installation	48 hours Minimum 1 year, *3 years 1 year, *2 years
Conviction	Class C Felony	Incarceration Revocation IID Installation	Minimum term of 90 days Permanent license revocation 5 years

2.1.4.1 The Implied Consent Law

According to Oregon law, all drivers consent to a breath, blood, or urine test when they choose to operate a motorized vehicle. If a police officer requests one of the above tests, the driver is legally obligated to provide a sample or participate in a breathalyzer test. Drivers over the age of 21 years old fail the test if their BAC is greater than 0.08%. For any drivers under the age of 21 years old, the test is considered a failure if any quantity of alcohol is found in the blood. In the case that a test is failed, the attending police officer will physically take the driver's license and issue a 30-day temporary driving permit. After the expiration of the temporary permit, the suspension period begins. Suspension is either 90 days or 1 year for a failed test, and refusal to take a test results in a 1 year or 3-year suspension (ODOT). This suspension is separate from any suspension due to a court conviction.

2.1.4.2 DUII Convictions

According to Oregon state law, DUII is either a Class A misdemeanor or a Class C felony, depending on the circumstances. Unless specific circumstances apply, most DUII convictions are classified as a Class A misdemeanor. The punishment for a Class A misdemeanor is a 1-year license suspension for a first-time offense. The driver may also be sentenced to 48 hours in jail or 80 hours of community service (McBreen 2020). For second offenses that occur in a five-year period, the suspension length increases to 3 years. A DUII conviction shall be a Class C felony if three times in the past 10 years, a driver has been convicted of a DUII while operating a vehicle, boat, or aircraft. Other situations in which a DUII may be felony include cases that involve vehicular manslaughter or criminally negligent homicide by operation of a motor vehicle while under the influence. Convictions in other jurisdictions are included in this number. After conviction of a Class C felony, any subsequent DUII conviction shall be considered a Class C felony. If convicted of a felony for DUII, the driver faces incarceration for a minimum of 90 days (ORS Section 813.011) and permanent license revocation after a third misdemeanor DUII (ORS Section 809.235).

2.1.4.3 DUII Diversion Program

There are a few non-revocation interventions currently employed by the state of Oregon to combat DUII convictions in what is known as the DUII diversion program. The first are treatment programs aimed at addressing the underlying behavioral issues that may lead to DUII. This may be an option if the driver completes a screening interview with an Alcohol and Other Drug Screening Specialist. For first time offenders determined to not have a dependency, drivers are referred to a DUII education program. For previous DUII offenders or those found to have dependency on intoxicating substances, the driver must complete a DUII rehabilitation program. If a defendant signs a diversion agreement, is compliant with all driving regulations throughout the program, and shows the court proof of successful treatment, the court may dismiss the conviction permanently. However, there are several reasons outlined in ORS Section 813.215 that make a defendant ineligible for these programs, such as felony convictions related to DUII or operating a commercial vehicle while under the influence.

Included in the Oregon DUII Diversion program is a separate program for Ignition Interlock Devices (IID). These devices are installed in vehicles to prevent drivers from starting the vehicle while impaired by testing their breath for intoxicants. Often, the installation of these devices is court mandated if a driver is going through a diversion program as a measure of program compliance. An IID is a requirement for hardship permits that allow motorists to drive during the period of DUII suspension. It is also a requirement for participants in diversion programs who are allowed to drive during a probationary period. IID are also used as a measure of compliance by the court. Following conviction of a DUII misdemeanor, drivers must use an IID for a year following the end of the suspension period for a first conviction, and two years for a second conviction. In the case of a felony or third misdemeanor DUII, an IID is required for five years following the end of the suspension or revocation period.

2.1.5 The At-Risk Driver Program

The At-Risk Driver Program run by the Oregon DMV aims at removing drivers with physical and cognitive impairments that present a danger to themselves and others on the roadway. The At-Risk Driver program provides another way for the state to identify drivers that struggle with substance abuse disorders and provide treatment through interventions, such as medical monitoring. Unlike the standard DUII diversion program, which operates exclusively through law enforcement and the court system, drivers may be entered into the At-Risk Driver program by relatives, health care providers, friends, law enforcement, court, or DMV administrative staff. This provides an additional layer of detection and intervention to the DMV DUII intervention system. Table 2.3 below summarizes Oregon At-Risk programs and related components

Table 2.3 Summary of At-Risk Driver Program Components

Stakeholder	Description	Responsibility Within Program
Mandatory Reporters	Licensed physicians, primary health care providers, and certified health care practitioners (ex: nurse practitioner, mental health providers, physical therapists, etc.).	Mandatory reporters must fill out a Mandatory Impairment Referral to the DMV including: Patient name, address, sex, DOB Driver impairment Description of how impairment affects driving ability Provider information, license number, and signature If the report is accepted, the DMV MDO will send a Driver Medical Record for the provider to complete.
Non-mandatory Reporters	Friends and family members of the driver, other citizens	Non-mandatory reporters may submit a Driver Evaluation Request to the DMV containing: Reporter name and signature Driver name and DOB Description of why drivers is suspected of being unable to drive safely
Law Enforcement	Law enforcement officers at all levels	Law enforcement officers may submit a Driver Evaluation Request including: Reporter name, law enforcement agency, and signature Driver name, DOB, and ODL Description of why the driver is suspected of having an impairment Documentation of contact: including citations or crash reports Law enforcement officers at not called upon to testify at at-risk driver hearings.
Oregon DMV	Driver Specialty Services Department and Medical Determination Officer (MDO)	Driver Specialty Services reviews all reports to the agency and issues notification of acceptance or rejection by mail to reporter. If the report is accepted, the suspension timeline varies depending on the type of report. Mandatory Report – immediate suspension of license with five days' notice Non-mandatory Report – immediate suspension only if there is reason to believe the driver is an immediate danger. Otherwise, the driver has 60 days' notice before suspension with the opportunity to be granted a 30-day extension to gather documents and take additional tests
Driver	The driver reported to the DMV through any of the above channels	The role of the driver is to comply with all DMV instructions and provide further documentation and testing as requested by the DMV, which may result in the reinstatement of driving privileges if deemed appropriate.

2.2 SIMILAR HTO PROGRAMS IN THE WESTERN UNITED STATES

In the United States, there are "at least 25 [states that] have enacted legislation regarding HTO's" (NCSL 2022). Figure 1.1 shows the current distribution across the country.

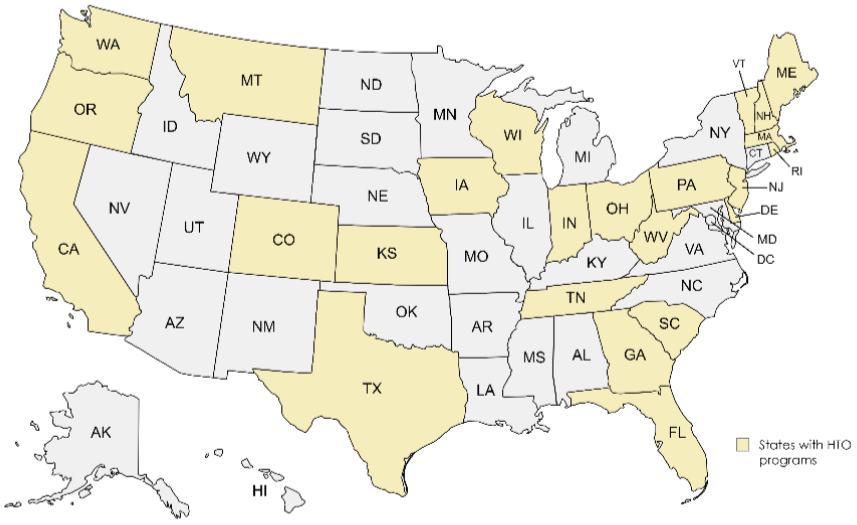


Figure 2.1 Map of the States with HTO Laws

The states that this study has chosen for comparative analysis are California, Washington, and Montana due to their geographic proximity to Oregon. By limiting analysis to the Western continental United States, the types of driving behavior present in those states may be more consistent. Thus, different policy initiatives will show varying approaches towards solving similar problems as opposed to varying regional problems.

2.2.1 Programs in California, Montana, and Washington

Using information gathered from the National Conference of Legislatures, a brief summary of each of the programs in the Western states has been compiled.

2.2.1.1 California

In California, an HTO is defined as a person who has accumulated a driving history while their license was revoked or suspended. A driving history can consist of two or more convictions with a point count of two within twelve months, three or more convictions with a point count of one within twelve months, three or more reported crashes within a twelve-month period, or any combination that results in a point count of three or above in a twelve-month period. Within 30 days of receiving a court or driving record that designates a driver as a HTO, the department will send notice to the district attorney responsible for the district the driver resides within. And, within 30 days of receiving the

notice, the district attorney will inform the department if the HTO will be prosecuted. The first conviction carries a punishment of a fine of $1,000 and imprisonment in jail for 30 days. A second conviction within seven years of the first carries the punishment of a fine of $2,000 and imprisonment in jail for 180 days. If a HTO is caught driving while their license is revoked, the punishment is a $2,000 fine and imprisonment in jail for 180 days (FindLaw, 2023).

2.2.1.2 Montana

An HTO is defined by Montana law as "any person who within a 3-year period accumulates 30 or more conviction points" (NCSL 2022). The point system that Montana uses is defined as follows:

- Deliberate homicide by operation of a motor vehicle: 15 points
- Mitigated deliberate homicide: 12 points
- Any offense punishable as a felony: 12 points
- Driving while under the influence: 10 points
- Failure to stop at a scene where another driver was injured or killed: 8 points
- Driving while license is suspended or revoked: 6 points
- Reckless driving: 5 points
- Illegal Drag Racing: 5 points
- Any vehicle liability protection offenses: 5 points
- Failure to stop at a scene where damage to property was inflicted: 4 points
- Speeding: 3 points
- All other moving violations: 2 points

When a driver has been designated an HTO, they may not be issued a license until three years have passed from the date of designation. One year into the three-year revocation, they may participate in a driver improvement program to obtain a restricted probationary license (MCA, 2023).

2.2.1.3 Washington

The definition of an HTO, according to Washington law, is "any person, resident or nonresident, who has accumulated convictions or findings that the person committed a traffic infraction as defined in RCW 46.20.270" (NCSL, 2022). To qualify, the driver must accumulate three or more convictions in the span of five years. Convictions include

vehicular homicide, vehicular assault, driving while under the influence, driving with a suspended or revoked license, failure to stop at the scene of an accident resulting in fatality, injury, or property damage, reckless driving, or eluding a police officer. In the case where more than one infraction is committed within six hours of the other, the department will treat it as one conviction but only the first time. Alternatively, if a driver is convicted of twenty or more reported traffic offenses (excluding driving without a permit), they may also be designated a HTO.

2.2.2 Comparison

Oregon does not have a demerit point system like the systems in California and Montana. Traffic violations are not weighted based on severity, but rather, there is a defined class of severe violations that are associated with the program. In addition to this severe class, there is a separate class of moving violations for which drivers can accumulate more convictions before reaching the threshold required to trigger admittance into the program. This is very similar to the HTO Program, in Washington, which also does not weigh traffic violations using a point system. In California and Montana, however, actions are given weights in a point system, and a threshold of points instead of convictions is set. These programmatic differences make it difficult to compare Oregon's program to those in California and Montana.

Another similarity between the Washington and Oregon HTO programs is the span during which convictions are counted. Both programs use three convictions within five years as the benchmark for the determination of HTO status. Of the four western states considered, the HTO definition in Washington is the most similar to the definition in Oregon. In the states of Montana and California, the programs have profound differences. The HTO program in California starts post-revocation. The program is only aimed at deterring drivers who have already had a license suspended, whereas Oregon and Washington use license suspension as an intervention within the program itself. And in Montana, the demerit point system and revocation period are different enough to warrant caution when making comparisons. Table 2.4 summarizing the main similarities and differences between the programs in relation to Oregon.

Table 2.4 Summary of HTO Program Parameters

State	Point System	Revocation Duration	Number of Convictions	Number of Traffic Offenses
Oregon	No	5 years	3 convictions	20 traffic violations
California	Yes	N/A	3 demerit points	N/A
Montana	Yes	3 years	30 demerit points	Moving violations: 2 points each
Washington	No	At least 4 years	3 convictions	20 traffic violations

2.3 OUTCOMES OF THE HTO PROGRAM

The primary concern of the HTO in Oregon is whether it succeeds in its goal to remove unsafe motorists from the road thereby preventing crashes. The primary intervention that the program uses to advance this goal is the threat of license revocation in response to non-compliance. To accurately gauge how well the program is working, the intervention of license revocation must

be assessed. In this section, a review of literature evaluating the efficacy of HTO Programs and license revocation is presented to establish the current body of knowledge.

2.3.1 Efficacy of HTO Programs

A major element of the HTO program is license revocation. If a motorist is found to be non-compliant with the program in the state of Oregon, violating the program restrictions by continuing to drive, they risk the penalty of a five-year revocation. Thus, one foundational question is does the tool work as a significant deterrent against future risky driving behavior? Studies have shown that license revocation, as opposed to license suspension, which involves a shorter length of time, has significant deterrent effects on the likelihood of recidivism in traffic offenders (Lee et al., 2018). Lee et al. conducted their study in South Korea, where the penalty for revocation was also a maximum of five years, which is the same as the state of Oregon. The penalty for suspension was a few months pause from driving, which is more severe than the Oregon DIP, in which suspension lasts for one month. The study concluded that the longer penalty of revocation resulted in less recidivism and increased the duration of compliance after the first revocation. This was done by utilizing Cox's proportional hazard model which allows for multivariate analysis.

2.3.2 Possible Opportunities for Improvement

Research has shown that using a demerit point system like the programs in California and Montana can have a positive deterrent effect on HTOs. It was found "that one demerit point reduced about 11.6% of the violation hazard for prior infringers" (Lee et al., 2018). The interaction between the effects of the demerit system and revocation are intertwined in systems where they are used concurrently. It was found that "for the limit of both suspension and revocation, the compliance duration of traffic law infringers can be extended when imposing a penalty of point accumulation" (Lee et al., 2018). Another study conducted by researchers Sagberg and Ingebrigsten in Norway focused on the impact of demerit points directly. The penalty point system in Norway sets the cap of demerit points at eight, and the accumulation of eight points within a three-year period results in a six-month revocation. The study found that "that driver at risk of losing their license tend to change their driving behavior so that they avoid further penalty points" (Sagberg & Ingebrigsten, 2018). This supports the conclusion that penalty (demerit) point systems have a positive deterrent effect on repeat traffic offenders.

Another possible avenue for improvement is increasing the real or perceived risk of being caught while driving with a revoked license. A study carried out by the California DMV looked into the rate at which drivers with revoked or suspended licenses were able to pass license checkpoints in the state of California. They found that 41% of suspended or revoked drivers were able to make it through checkpoints undetected and that the primary offenders were those who had not complied with mailing in their expired licenses to the DMV (Parrish and Masten, 2014). Due to the lack of electronic equipment used at the checkpoint sites, law enforcement officials were unable to determine which licenses were suspended. The study suggests that the low rate of capture could create significant safety issues, as unsafe drivers may perceive the risk of being caught to be low and drive illegally with greater frequency. The study identified a major weakness of the program, the ability to identify when HTO's continue to drive illegally. A few suggestions from the study include electronic card readers at checkpoints that can alert law

enforcement and the increased use of certified mail when sending license revocation notices to increase the DMV's certainty that the driver was correctly notified of their revocation.

There also exists an alternative form of habitual traffic offender treatment that focuses on rehabilitation over penalties. A form of this rehabilitation exists in Germany, where repeat traffic offenders are required to pass a "medical psychological assessment (MPA)" before having their license reinstated, as opposed to a stated timeline (Glitsch & Knuth, 2016). The idea behind this system is that behavior change is heavily influenced by the importance a traffic offender places on the change. Financial and penal barriers only address the outcome of a repeat offender, while rehabilitation addresses the root behavioral causes leading to less recidivism. The tool used to assess rehabilitation in Germany is the MPA, which consists of three parts: "a medical examination, computer-based performance tests, and psychological assessment. The results are summarized in a final overall assessment of the person's fitness to drive" (Glitsch & Knuth, 2016). This tool has been shown to help predict whether a person is likely to relapse into old behavioral patterns, with "only 6.5% [recidivism] in a 3-year period". The study also showed that giving information to repeat offenders early in the rehabilitation increases the likelihood of their success in the program from 37.1 to 81% (Glitsch & Knuth, 2016). The proposed amendments to traffic offender programs that came out of the study were increased guidance from certified MPA counselors, individualized rehabilitation plans, and an instructional booklet containing the terms of the program written in plain language.

2.4 BEST PRACTICES FOR TRAFFIC DATA LINKAGE

This report analyzes traffic data from varying archives maintained by Oregon Driver & Motor Vehicle Services. This section presents an overview of the dataset and examples of multi-source traffic data analysis that have been used to analyze similar datasets in past studies.

2.4.1 Traffic Data in the State of Oregon

2.4.1.1 DMV Data

Historically, researchers and traffic safety analysts have utilized DMV data to study licensing patterns, identify demographic factors that influence driving behavior, and evaluate the effectiveness of driver education programs. For instance, they might conduct studies to analyze age-related driving trends or the impact of gender on traffic violations. Typically, such analysis may involve linking DMV data with other sources, such as crash reports or adjudication outcomes, to explore correlations or causal relationships.

Data collected by the DMV primarily includes personal information and specific details of drivers registered within the state. This dataset contains various pieces of information, such as Name, Driver's License ID (ID), Date of Birth (DOB), and Sex, which are listed in Table 1.3. These details can be used for a wide range of analytical and operational applications, like ensuring legal compliance in vehicle operation, assisting law enforcement, and enhancing road safety through behavioral analysis.

Table 2.5 Example of DMV Data Summary

Variable Name	Data Description
Name	The full legal name of the driver.
ID (Driver's License ID)	A unique identifier assigned to each licensed driver, typically a combination of letters and numbers.
DOB	Date of Birth; indicates the driver's age and is crucial for eligibility and demographic analyses.
Sex	Gender of the driver; categories include Male, Female, Non-Binary, among others to accommodate diversity.

To that end, DMV data is crucial for various transportation research and policy development objectives. It provides essential details about drivers, such as their demographics, license status, and history of violations or suspensions. This data plays a pivotal role in identifying risky populations, i.e., drivers, assessing the effectiveness of driver education programs, and monitoring the impact of licensing policies on road safety. Several previous studies have used this data for various applications.

- *Driver Education and Training:* Studies such as the one conducted by Mayhew et al. (1996) have used DMV data to assess the outcomes of graduated licensing systems, demonstrating their effectiveness in reducing crashes among novice drivers.

- *Repetition:* Other research analyzed the likelihood of re-offending among drivers with prior offenses, often relying on DMV records to track individuals' driving history over time, informing interventions aimed at reducing repeat offenses.

- *Demographic Analyses:* DMV data has been used to study the impact of gender or age on driving behavior, e.g., helping to develop targeted policies for older drivers to maintain their mobility and safety.

2.4.1.2 Adjudication Data

Adjudication data, also known as verdict data, provides detailed information about the outcomes of traffic violations and legal records related to driving offenses. This dataset is filled with valuable information, including Name, Address, ID (Driver's License ID), Violation State, Violation Jurisdiction, Verdict ID, Violation Code, Violation Description, Citation Date, and Verdict, as illustrated in Table 2.6. It is essential for understanding the legal consequences of traffic violations and for monitoring the enforcement of traffic laws.

Adjudication data is frequently used in research to measure the effectiveness of traffic law enforcement strategies, analyze traffic violation patterns, and assess the impact of legal penalties on reducing traffic offenses. Linking adjudication data with DMV and crash data could provide valuable insights into identifying repeated offenses, the efficiency of penalty systems, and demographic trends in traffic law violations.

Table 2.6 Example of Adjudication Data Summary

Variable Name	Data Description
Name	The full legal name of the individual involved in the traffic violation case.
Address	The residential address of the individual, which could be used for correspondence or legal purposes.
ID (Driver's License ID)	A unique identifier for the individual, often used to link their driving records and violations.
Violation State	The state in which the traffic violation occurred, indicating jurisdiction.
Violation Jurisdiction	More specific location within the state, like county or city, detailing where the offense took place.
Verdict ID	A unique identifier for the legal outcome of the violation case.
Violation Code	A specific code assigned to the violation, categorizing the nature of the offense according to legal standards.
Violation Description	Detailed description of the traffic violation, providing insights into the nature of the offense.
Citation Date	The date on which the traffic citation was issued, important for legal proceedings and records.
Verdict	The outcome date of the adjudication process, such as Guilty, Not Guilty, Fined, Warning, etc., reflecting date of the legal decision.

Verdict data also plays a significant role in policy evaluation, particularly in assessing the impact of changes in traffic law on driver behavior and safety. For example, in a recent report titled Strategies to Improve State Traffic Citation and Adjudication Outcomes that was published by the Behavioral Traffic Safety Cooperative Research Program (BTSCRP) and the National Academies of Sciences, Engineering, and Medicine (2023), the authors discussed the importance of tracking citation and adjudication data for identifying risky drivers and suggests strategies for improving the citation-adjudication process. The analysis provided insights into how different penalties impact driver behavior and the importance of data for policy analysis. To that end, the data can be used for various purposes, such as:

- *Efficacy of Legal Penalties:* Research leveraging adjudication data has explored the gradual effects of various penalties on future traffic violations, offering insights into how different types of actions (e.g., fines, license suspensions) impact driver behavior.

- *Linkage with Crash Data:* By linking adjudication data with crash records, studies have examined patterns in post-violation crashes, identifying trends that suggest areas for intervention to prevent future incidents. For example, a study might investigate whether drivers who receive specific types of penalties for DUI offenses are less likely to be involved in subsequent crashes.

- *Policy Analysis:* Analyses of adjudication outcomes can inform policymakers about the real-world impacts of laws aimed at reducing distracted driving, speeding, and other risky behaviors. This has implications for refining legal approaches to enhancing road safety.

2.4.1.3 Crash Data

Each crash record has details about the location and time of occurrence along with crash severity and several other driver, roadway, and environmental related factors such as weather, driver sobriety, any changes to roadway at the time of crash such as construction, etc., as shown in Table 2.7. State Departments of Transportation (DOTs) play a crucial role in collecting and managing crash data in the United States. The data typically originates from detailed reports compiled by law enforcement officers who are often the first to arrive at the scene of a traffic crash. This information can help provide a more detailed understanding of the factors contributing to road crashes. It is also vital in identifying hazardous locations, evaluating the effectiveness of road safety measures, and guiding policy and infrastructure changes, which are all aimed at reducing crashes. After compiling the detailed crash reports, they are submitted to the respective State DOT.

State DOTs are responsible for aggregating, managing, and analyzing crash data to identify patterns, trends, and areas of concern related to road safety. This is a critical process that helps develop effective traffic safety measures, inform road design improvements, and shape traffic enforcement policies. The data collected also supports various state traffic safety programs, engineering projects, legislative initiatives, and law enforcement efforts, aimed at reducing traffic crashes and enhancing the safety of all road users. The collaboration between law enforcement agencies and State DOTs ensures that crash data is not only systematically collected across the country but also used to inform and improve traffic safety strategies at both the state and national levels. Due to its importance, crash data is collected carefully, updated regularly, and analyzed thoroughly. This ongoing process ensures that road safety measures remain relevant and effective, adapting to changing conditions and emerging challenges.

That said, the significance of crash data lies in its detailed records, such as the date and time of crashes, locations, types of vehicles involved, crash circumstances, and outcomes. This enables a multidimensional analysis of traffic safety, allowing for targeted interventions. Thus, crash data analysis has been essential for evaluating the impact of environmental factors, vehicle technologies, and driver behaviors on road safety.

Table 2.7 Example of Crash Data Summary

Variable Name	Description
Location	Specific site of the crash, providing spatial context.
Date and Time	When the crash occurred, providing temporal context.
Severity	The severity of the crash impacts, from minor injuries to fatalities.
Crash Type	Description of the crash, providing insights into potential causes and preventive measures.
Injury Type	Specific type of injuries sustained by parties involved in the crash.
Number of Vehicles Involved	The total number of vehicles involved in the crash.
Vehicle Type	Details about the vehicles involved, which can correlate with DMV data.
Driver Behavior	Noted behaviors leading to the crash, like impaired driving, which can be analyzed with adjudication data.
Alcohol Involvement	Indicates whether alcohol was a contributing factor in the crash. which can be integrated with adjudication data.
Hit and Run	Indicates if the crash was a hit-and-run, crucial for legal adjudications.
Weather Condition	Environmental factors at the time of the crash.
Road Condition	State of the road, which can influence the occurrence and severity of crashes.
Traffic Control Devices	Presence and type of traffic control at the crash site.
Traffic Signal Status	Status of traffic signals at the crash site (e.g., green, red, malfunctioning).

In Oregon, as in many other regions, it is possible to integrate crash data with DMV and adjudication records to obtain a more comprehensive understanding of traffic safety issues. This integration can provide a broader perspective on driver behaviors, compliance, and the effectiveness of traffic laws and enforcement practices. This linking method can help:

- *Identify High-Risk Groups:* By analyzing crash involvement in conjunction with driver histories and adjudication outcomes, strategies can be tailored to specific demographics or driver types.

- *Evaluate Policy Impact:* The effectiveness of road safety policies and interventions can be assessed by observing changes in crash patterns before and after their implementation.

- *Enhance Driver Education:* Insights from linked data analyses can be used to update and improve driver education programs, focusing on areas of greatest need identified through empirical evidence.

The use of crash data analysis is essential for creating comprehensive strategies to improve road safety. This includes making infrastructure enhancements, implementing effective law enforcement strategies, and launching public awareness campaigns. Crash data analysis is not just limited to statistical analysis, as it also plays a role in designing safer roadways, developing effective traffic laws, and planning education protocols. By identifying patterns and trends in crash data, researchers and policymakers can take proactive measures to mitigate future risks rather than only reacting to crashes that have already happened.

2.4.2 Multi-source Traffic Data Analysis

2.4.2.1 Data Cleaning in Preparation for Linkage

The quality of data linkage depends on the quality of the data used to generate matches. Frequent issues that appear when linking data sets include empty cells, empty rows, spelling errors, and duplicate cells. Additionally, when handling datasets of varying sizes, it can be difficult to achieve data agreement. Because incomplete or incorrect data introduces significant error in data linkage, data cleaning is a vital step to the success of merging independent traffic data sets in preparation for analysis.

Null or empty data fields are a prevalent issue in data linkage, as missing or incomplete data can contribute to matching errors. There are a few ways to mitigate this issue. One such tool is multiple imputation, the process of generating "imputed values that are representative of the original data" (Karimi et al., 2024). This creates a set of representative data that can be linked without the error created by empty data fields; however, in these cases, the analysis quality is only as good as that of the imputed data. An example of this approach is using linear interpolation to fill gaps in speed data if those gaps do not exceed a significant period (Bamney et al., 2022). An alternative approach to improving data integrity is that of linking the dataset with missing fields to a supplemental dataset that contains the missing information. An example of this can be seen in a study conducted by researchers for the Michigan Department of Transportation (MDOT). During the data cleaning process, they identified gaps in traffic volume data on divided highways. To compensate, they used MDOT's sufficiency file, which overlapped with their collected files and contained no missing segments (Savolainen et al., 2022). This ensured increased data integrity throughout the analysis.

When linking data sets involves strings such as names and locations, many small errors can arise due to the non-uniformity of data collection. Issues such as case sensitivity, suffix disagreement, nicknames, and the inclusion of middle initials are a few examples

of the numerous ways in which data fields can differ. Because these fields are often used as unique identifiers to link data, it is important that there is consistency within and between datasets. Otherwise, computer software may not recognize these unique keys as matching. There are numerous documented techniques to address this issue, including character uncertainty comparisons and string matching algorithms (Karimi et al., 2024).

The last common problem with linking large traffic datasets is that of varying size and dates of collection. Linking datasets of disproportionate sizes can introduce errors from multiple facets such as duplicate cells and variation between data collections. An example of this can be found in another study performed for MDOT in 2022, in which two types of data, free-flow speed data, and vehicle probe data, were collected and aggregated in different manners. To compensate for the increased variation of the probe vehicle data, it was segregated into groups based on season and time of day. This reduced time-related variation and allowed for better integration with coarse aggregated free-flow data (Savolainen et al., 2022a; 2022b).

2.4.2.2 Deterministic Linkage

Deterministic methods involve generating matches between datasets by looking for agreement between unique identifiers in the data. Put more simply, it is joining two datasets based on a common attribute. However, this attribute must be distinct so there is not notable overlap with other items. This allows deterministic linkage to achieve "a high level of linkage specificity" (Auguste & Pawelzik, 2024). Because of how straightforward the procedure is, it has been used successfully in various transportation studies where unique data fields are present. An example of a basic deterministic linkage procedure can be seen in Figure 2.2.

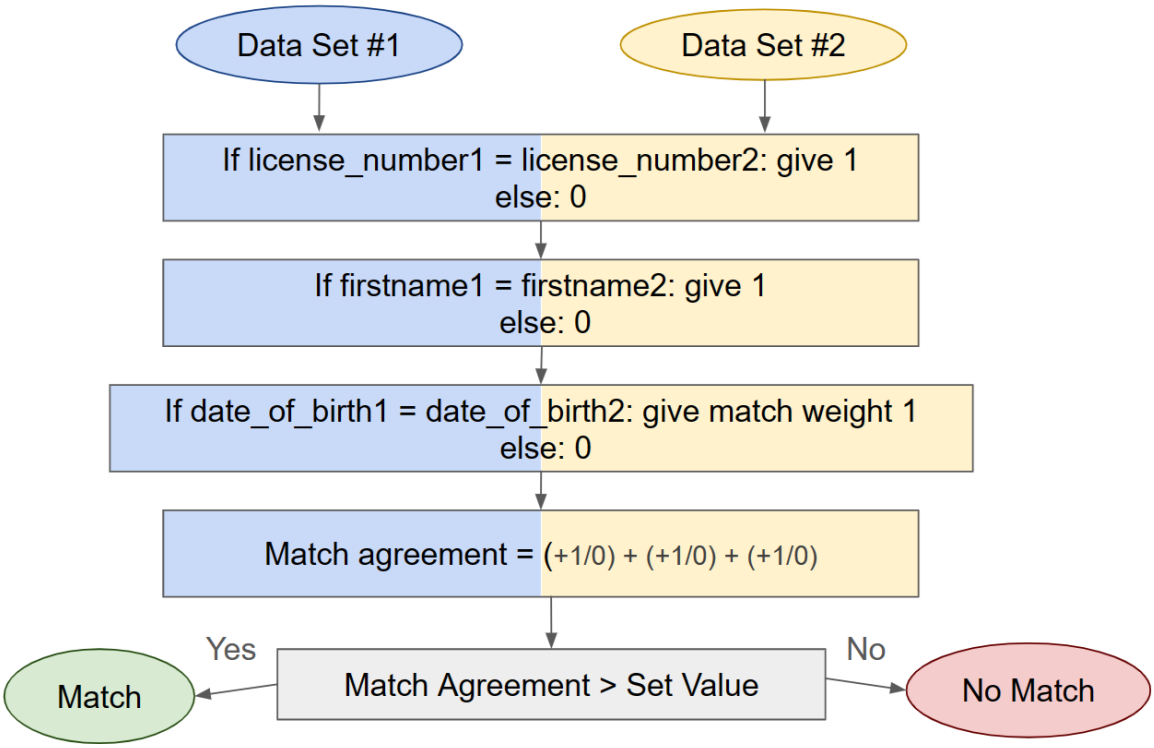


Figure 2.2 Deterministic Linkage Model Example

It can be hard to visualize what this workflow may look like in practice. To help illustrate a practical example of deterministic linkage, two fictitious sample data sets were created (Table 2.8 and Table 2.9). Then, using the unique identifiers of first names and last names, the two data sets were linked using deterministic linkage (

Table 2.10). The match agreement value was set at two in this specific example for the purposes of illustration.

Table 2.8 Example Data Set #1

ID	First Name	Last Name	DOB	Sex	License Number
1	Bob	Jacobs	1/24/2006	Male	A123456
2	Ryan	Green	5/9/1978	Male	A789101
3	Samantha	Sanders	11/21/1992	Female	A111213

Table 2.9 Example Data Set #2

ID	First Name	Last Name	Middle Name	Sex	Phone number
1	Bob	Jacobs	Jonathon	Male	403-392-3928
2	Ryan	Greene	Lee	Male	394-018-1640
3	Sam	Sanders	Molly	Female	283-497-9384

Table 2.10 Example of the Linkage of Data Sets #1 and #2

ID	First Name	Last Name	First Name	Last Name	Middle Name	DOB	Sex	Sex	Phone number	License Number	First Name Match	Last Name Match	Match
1	Bob	Jacobs	Bob	Jacobs	Jonathon	1/24/2006	M	M	403-392-3928	A123456	Yes	Yes	Yes
2	Ryan	Green	Ryan	Greene	Lee	5/9/1978	M	M	394-018-1640	A789101	Yes	No	No
3	Samantha	Sanders	Sam	Sanders	Molly	11/21/1992	F	F	283-497-9384	A111213	No	Yes	No

The simplicity of the deterministic model lends itself well to traffic data. One example of a study that used deterministic linkage was conducted by researchers at the Connecticut Transportation Institute. Auguste and Pawelzik, used several rounds of deterministic linkage to integrate breathalyzer data with police-reported crash data. This was achieved by replicating multiple rounds of matching, changing the variable qualifiers each time. The idea behind this methodology was to generate "numerous matching possibilities while still maintaining high data integrity" (Auguste & Pawelzik, 2024). After these matches were created, a match score was assigned based on the weight of the variables used to create it. From here, any case with a low match threshold was reviewed on a case-by-case basis. The result was 5,634 linked records with a false positive match proportion of 0.1% and a true match proportion of 84.7% (Auguste & Pawelzik, 2024). To evaluate the linked dataset, the researchers compared proportions such as sex, age, and injury severity from the original datasets to the proportions of the linked dataset. It was found that proportions from the linked dataset were consistent with expectations when compared to the original Driving Under the Influence (DUI) crash data, and "in the cases where there were significant changes in proportions, most, if not all, [could] be attributed to things outside of the linkage process" (Auguste & Pawelzik, 2024).

Another study carried out by researchers at the University of Massachusetts Amherst used deterministic linkage to combine data from the Massachusetts Crash Data System. In particular, the study focused on linking police-reported crash data (CDS) and EMS data documented through the Massachusetts Ambulance Trip Record Information System (MATRIS) (Tainter et al., 2020). The researchers used MATRIS as the base dataset to reduce the scope of cases and linked the CDS data to it. The linkage relied on incident date, incident location, patient date of birth, patient home zip code, and patient gender (Tainter et al., 2020). It was found that over 95% of matched records were true matches when the data was verified by the Massachusetts Department of Public Health. From the matched data set, researchers pulled several key fields such as chief complaint anatomic location, injury severity, and manner of collision to develop insight into injury trends pertaining to emphasis areas in the Highways Safety Improvement Program such as lane-departure and speeding crashes. The linked dataset allowed for a more comprehensive analysis of injury causation than the data found in CDS or MATRIS alone.

These two cases illustrate the analytic potential of deterministic linkage; however, the method does not come without limitations. Because deterministic linkage depends on high-quality, unique keys, it can be nearly impossible in some cases to link datasets together using this method. If there is a lack of unique keys or data errors, researchers will often create "decision rules" that govern which variables are given more weight when comparing matches than others. However, this can introduce human error as the choice of decision rules is up to the discretion of the researchers conducting the study (Doidge & Harron, 2018). Additionally, deterministic linkage requires highly polished datasets to reduce the error associated with the methodology. The intensity of data cleaning makes sense as deterministic linkage "faces constraints when the available data does not have unique identifiers or contains incomplete or wrong information" (Karimi et al., 2024). Thus, while data cleaning is relevant to all forms of linkage, it is especially pertinent when making direct matches. The last major limitation of deterministic linkage is that it does not handle confidential data well. The method relies on unique identifiers

such as those found in personal identifying information (PII), e.g., name, date of birth, or license number, and that data is often well protected and subject to privacy policies. Thus, whether or not deterministic linkage is the appropriate tool for traffic linkage depends on multiple factors that must be evaluated on a case-by-case basis.

2.4.2.3 Probabilistic Linkage

Probabilistic data linkage methods are used when there is a lack of strong, unique identifiers to merge datasets. This is pertinent to cases where there is a lack of consistency across report types or when aggregate data is used to avoid leaking PII. Without unique fields, the certainty of true matches decreases; however, probabilistic linkage is a powerful tool that can bypass some of the common limitations of traffic safety datasets. The fundamental idea behind probabilistic linkage is that it involves creating probabilistic models for two (or more) datasets that need to be linked. The models are then compared using quasi-unique fields (such as gender or last name) to generate a match score. This allows researchers to extrapolate conclusions without precise matches by observing the agreement patterns across datasets. Another way of describing probabilistic linkage is looking at an array of attributes that each narrow down the list of possible matches until the most likely match is found. The drawback of probabilistic linkage is that it "heavily depends on the quality and relevance of the chosen variables and the accuracy of the underlying probabilistic model" (Karimi et al., 2024). However, when used carefully, it allows for the analysis of complicated datasets through the use of statistical modeling. A good way to approach the difference between deterministic and probabilistic linkage is that deterministic linkage is based on rules, and probabilistic linkage is based on weights or scores (Doidge & Harron, 2018). Neither method is inherently superior or more accurate than the other. Rather, it is the conditions imposed by the data that decide which method is best suited for a particular linkage task.

A basic example of the workflow of probabilistic linkage is shown in Figure 2.3. The x, y, and z variables represent match weights generated by the underlying prediction model. These would determine how heavily the variables sex, first name, and age are used to predict if two data entries are a match. The match probability cutoff value would also be determined by statistical means based on the model chosen.

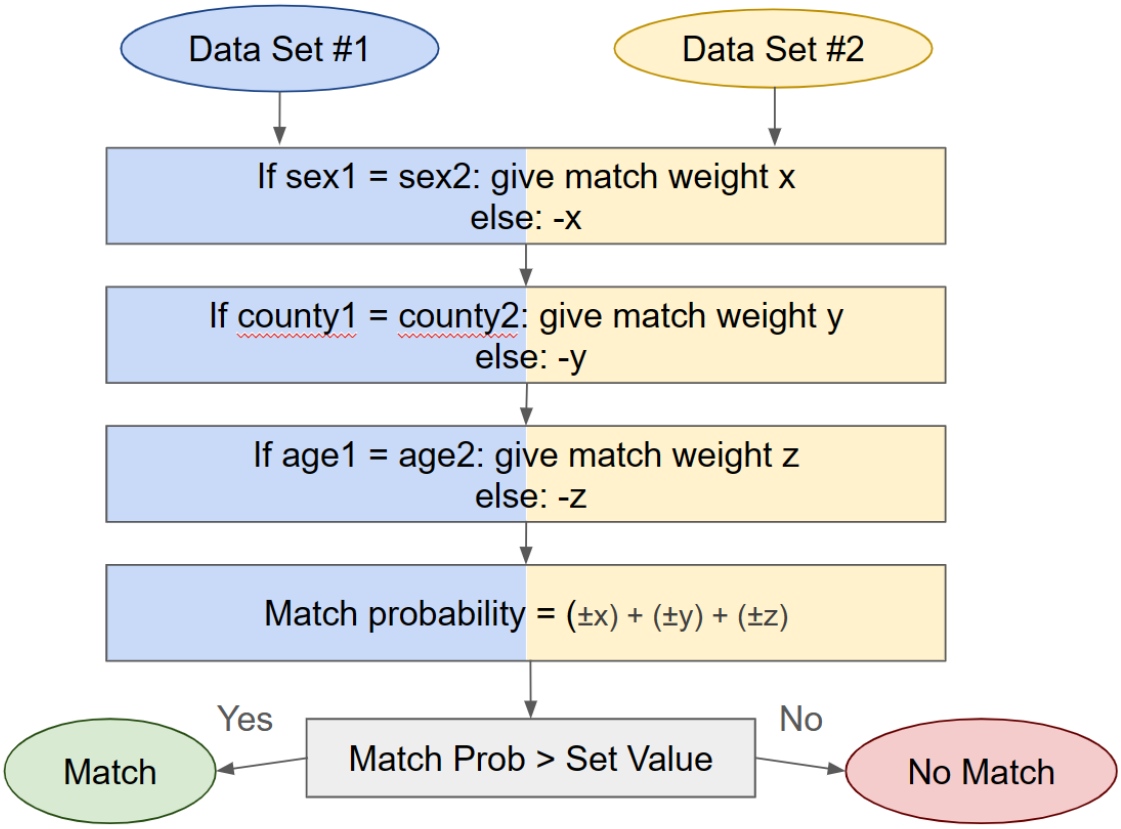


Figure 2.3 Probabilistic Linkage Model Example

To illustrate the power of probabilistic linkage, a third fictitious data set was created (Table 2.11) and linked to data set one in Table 2.8. The match probability threshold was arbitrarily set at 0.40 for demonstrative purposes and the match weights were randomly generated (Table 2.12).

Table 2.11 Example Data Set #3

ID	First Name	Sex	Age
1	Bob	Male	18
2	Ryan	Null	46
3	Sam	Female	32

Table 2.12 Probabilistic Linkage of Data Sets #1 and #3

ID	First Name	Last Name	First Name	Sex	Sex	DOB	Age	License Number	First Name Match	Sex Match	Match Prob Total	Match
1	Bob	Jacobs	Bob	Male	Male	1/24/2006	18	A123456	Yes: 0.23	Yes: 0.18	0.41	Yes
2	Ryan	Green	Ryan	Male	Null	5/9/1978	46	A789101	Yes: 0.23	No: -1.15	-0.92	No
3	Samantha	Sanders	Sam	Female	Female	11/21/1992	32	A111213	No: -1.24	Yes: 0.18	-1.06	No

The probabilistic linkage models presented so far have been theoretical, but there is an already developed probabilistic method for linking police crash databases and medical databases referred to as the Crash Outcomes Data Evaluation System (CODES) (Kweon, 2011). CODES is run by the National Highway Traffic Safety Administration (NHTSA). The system is special in that it uses aggregates to avoid identifying individuals, and it is used on both the national and state levels. NHTSA publishes reports from states that have adopted such program, but unfortunately, the state of Oregon is not one of them. These reports include a detailed methodology for mapping and analyzing traffic datasets using advanced software. One such report published by researchers at the University of Utah provides a comprehensive summary of CODES methodology and applications using the General Use Model (GUM), which contains a standardized list of common traffic safety data elements. Using data from eleven states, researchers were able to provide four sample analyses "designed to demonstrate the utility of the GUM" (Cook et al., 2015).

The Model Minimum Uniform Crash Criteria Guideline or MMUCC Guideline dedicates chapter 10 to describing the best practices for traffic data integration and linkage. The chapter analyzes various state data collection agencies to determine which elements can be linked to state-level crash data. Examples of agencies include the American Association of Motor Vehicle Administrators, the Commercial Driver's License Information System, state citation and adjudication databases, traffic court records systems, the National Emergency Medical Services Information System, and the National Trauma Data Bank.

2.4.2.4 Geospatial Coordinate Linkage

As tools such as ArcGIS and other geospatial software have become more advanced, there has been an increase in their use for traffic safety analysis. This is due in part to the implementation of large-scale geocoded coordinate datasets. These datasets can link attributes such as the details of vehicular crashes and road geometry to geographic points, allowing for the integration of data through geospatial analysis.

An example of this type of analysis can be seen in a report sponsored by the MDOT, wherein the researchers used ArcGIS to merge datasets with mile lane inventory, road classifications, and traffic volumes using spatial analysis tools. Data cleaning was also performed in ArcGIS to ensure geographical continuity and correctness. This allowed researchers to build a decision-support tool that recommends treatments for pedestrian

and bike safety based on context-sensitive data such as speed limits and road geometry (Savolainen et al., 2022). This example showcases how geographic data can be combined to create a more complete picture of how geography and safety intersect.

Another application of geospatial analysis is spatial regression modeling. Various geospatial software tools are capable of statistical modeling within the program itself, allowing for integrated analysis. A study performed in the city of Baltimore, Maryland, used the programs QGIS and GeoDa to generate spatial autoregressive lag (SAR) models. These SAR models were used to regress socioeconomic indicators and crash proportions in the city of Baltimore (Dezman et al., 2016). It was concluded that socioeconomic factors were not associated with the crash distribution of the city of Baltimore, but that knowledge allowed researchers to pursue other approaches to better predict behavior which is in itself powerful. These built-in tools allow links between demographic and geospatial data that is critical to understanding underlying patterns that directly impact safety.

The limitation of geospatial linkage is that is requires precise, quality geocoordinates. In some cases, this data may not be available which prohibits the use of this linkage method. And even if geospatial data is available, it may be unreliable. Geospatial data is notorious for needing intensive data cleaning and processing, just like attribute data. For example, it may be necessary to screen for crash data points that are not aligned with a road or highway. If a crash data point shows up 200 ft from the road of interest, its geocoordinates may be invalid. And these rogue data points can have a measurable influence on data integrity and analysis. Thus, secondary data cleaning specifically for geospatial data is necessary to use this method effectively.

2.5 TRAFFIC DATA LINKAGE CASE STUDIES

Because large-scale traffic data linkage is an emerging field of data science, there is comparatively less pervious academic work than other more established research practices. However, the work that has been done shows promise of the benefits that having large-scale, integrated crash datasets can have. This section will review a few cases in which linked crash datasets produced robust research outcomes.

2.5.1 New Jersey Safety and Health Outcomes Data Warehouse

The Children's Hospital of Philadelphia (CHOP) houses the New Jersey Safety and Health Outcomes (JS-SHO) Center for Integrated Data. This center has recently been dedicated to building out the JS-SHO Data Warehouse: a crash dataset with integrated citation, driver's license, birth certificate, EMS, and hospital data (Carey, 2020). The database integrates administrative datasets that encompass the entire state of New Jersey. Due to the large size of the database, probabilistic linkage was used to combine the administrative datasets. The methodology involved an iterative algorithm designed to link all data sources independently. If two sources agreed, they were grouped in a set under an individual. This maximized connections and prevented the algorithm from allowing minor disagreements to interfere with the matching process (Curry et al., 2021). In addition to administrative data, equity indicators were also integrated within the database. This was a major focus during development to distinguish the tool

from other integrated datasets by allowing more complex sociological analysis. Another new addition was the use of widespread geospatial data integration. The software engine ArcGIS was used to geocode the residential addresses of all individuals with a New Jersey address. This was done to assist in determining the distance between residential addresses and crash locations for future analysis. The rich depth of data combined with the individual case linkage approach allows the data warehouse to support both broad transportation investigations, as well as specialized study into subgroups identified within the base. The usefulness of the JS-SHO Data Warehouse can be found in the numerous studies that were conducted using the integrated data.

2.5.1.1 Child Safety and Young Driver Programs

The use of specialized restraints for young children in vehicles has been a quickly progressing science over the past few decades. A study carried out using the NJ-SHO database examined the injury and driver characteristic trends of children involved in collisions where the restraint status of the child was identified. It was found that young children were more likely to be injured if restrained using a vehicle belt instead of a booster seat. Children were also more likely to be improperly restrained if the driver was not wearing proper restraints, had evidence of alcohol abuse, was at fault for the crash, or was outside the age of 21-34 years old (Myers et al., 2022). This suggests that continued effort regarding child restraint interventions and further research into child restraint injury patterns will most likely be needed to resolve gaps in restraint use.

CHOP also sponsors the Young Driver Safety program, which has used the NJ-SHO database to investigate trends in the driving behavior of teenagers and young adults. Many studies in this area of interest center on the overlap between driving behavior and mental health. One such study looked at the relationship between mood disorders and the rate of licensure and crashes in young adult drivers. It was found that youths with mood disorders were 30% less likely to acquire a license compared with youths without a mood disorder and rates of moving violations among drivers with mood disorders were greater than among those without mood disorders (Gaw et al., 2024). Additionally, neurology is also considered when attempting to further understand young driving behavior. It was found that among young drivers, those with ADHD were more likely to crash multiple times and were determined to be at fault for a higher proportion of their crashes than their non-ADHD counterparts within 24 months (Curry et al., 2022). This opens up opportunities to understand how the brains of young adult drivers differ from those of adult drivers, and how to tailor driving interventions particular to those demographics.

2.5.1.2 Studies on the Impact of Advanced Age on Driving

Advanced age has a significant impact on both cognitive and physical ability, both of which are important to safe driving behavior. Thus, it is imperative to study exactly how advanced age impacts crash rates. For example, it found that the overwhelming majority (95%) of crashes occurred within 25 miles of the driver's residence (Joyce et al., 2022). Thus, distance-based restrictions for older drivers are likely to be ineffective. Additionally, the database has been used to determine the risks associated with older drivers. It was found that older licensed drivers have lower crash rates than middle-aged drivers; however, their rate of being involved as a driver in a fatal crash is 30% to 50%

higher (Palumbo, 2019). These findings can be used to determine effective interventions for older drivers that help them retain their autonomy while ensuring safety.

2.5.1.3 Equity in Transportation

Using the NJ-SHO database, a new program known as the Bayesian Improved Surname Geocoding (BISG) algorithm was created to estimate ethnic and racial demographic information. BISG works by combining census information on surnames and racial/ethnic composition to produce the probability that an individual belongs to one of six groups: White, Hispanic, Black, Asian/Pacific Islander, Multiracial, and American Indian/Alaska Native. It was demonstrated that it is possible to calculate BISG race/ethnicity probabilities for 98.9% of drivers using surname and residential address, two fields commonly available in licensing and crash data (Sartin et al., 2021). This is just one tool being implemented alongside the database. Many more studies are emerging on other vulnerable populations. For example, it was found that those living in lower-income areas were much less likely to be driving safe vehicles, a pattern that was particularly strong among the youngest drivers (Metzger et al., 2020). As the database grows and becomes more integrated, the sample size of marginalized populations will grow as well. This will allow transportation professionals and policy makers to analyze macroscopic trends and increase the safety of the roads for all users.

2.5.2 Crash Outcomes Data Evaluation System (CODES)

The probabilistic linkage models presented so far have been theoretical, but there is an already developed probabilistic method for linking police crash databases and medical databases referred to as the Crash Outcomes Data Evaluation System (CODES) (Kweon, 2011). CODES was run by the National Highway Traffic Safety Administration (NHTSA) from 1992 to 2013 with the intent of linking vehicular crashes with the associated medical and financial outcomes. This was done to create a more comprehensive understanding of crash outcomes. In 2013, NHTSA transferred control of the program to individual states, but this was only if they chose to adopt the CODES program. NHTSA still publishes reports from states that have adopted such a program, but the state of Oregon is not one of them. These reports include a detailed methodology for mapping and analyzing traffic datasets using advanced software. One such report published by researchers at the University of Utah provides a comprehensive summary of CODES methodology and applications using the General Use Model (GUM), which contains a standardized list of common traffic safety data elements. Using data from eleven states, researchers were able to provide four sample analyses “designed to demonstrate the utility of the GUM” (Cook et al., 2015). This provides an example for how such a system might work. It also illustrates an important feature of the CODES program; it can be used on both the national and state level. Due to the aggregation of data, CODES methodology is capable of linking massive datasets. This makes it a prime candidate for statewide or multistate transportation studies.

2.5.2.1 NHTSA Studies

Using the CODES database, the injury outcomes for crashes involving motorcyclists were evaluated across 18 states. The purpose of the study was to determine if helmet use had a significant impact on head and face injuries. It was found that wearing a helmet

reduced motorcyclist head and facial injuries. Helmets are 40 percent effective at preventing moderate to severe head or facial injuries in single-vehicle crashes and 22 percent effective at preventing moderate to severe head or facial injuries in multiple-vehicle crashes. It also significantly reduced the likelihood of a traumatic brain injury (TBI). It was estimated that the effectiveness of motorcycle helmets at preventing TBI was 41 percent for single-vehicle crashes and 25 percent for multiple-vehicle crashes. Thus, while the study only encompassed 18 states, it can be concluded that mandatory helmet laws for motorcyclists have the potential to lower the quantity of head injuries that occur in crashes involving motorcyclists (Cook 2009).

Another study performed by NHTSA analyzed the importance of seatbelts in reducing morbidity (the occurrence of injury) and mortality using CODES data from seven states. It was found that seatbelt use was highly effective at reducing both. For another form of comparison, inpatient charge for unbelted passenger vehicle drivers admitted to an inpatient facility as a result of a crash injury was more than 55 percent greater than the average charge for those who were belted. This provides quantitative consequences to limited seatbelt use. One of the limitations of the study was that seat belt use was found to be overreported in police crash reports, which is in accordance with other studies performed by NHTSA. However, when the values were adjusted, the difference was still significant (Johnson et al., 1996).

2.5.2.2 State-run Studies

A study conducted by the Kentucky Transportation Center (KTC) demonstrates the use of CODES at a state-wide level. Researchers set out to identify the impact of Cable Median Barriers (CMB) on the injury severity of crashes. CODES was used as a preexisting integrated database containing both injury and crash data. From there, databases containing highway geometry were combined with CODES to identify crashes that occurred on sections of highway involving CMB. This was then compared to highways that use concrete medians and those that do not have a median to understand the impact of CMB on crash injuries. It was found that compared to road segments with no median barrier, occupants in median-involved crashes on a road segment with a CMB were 72% less likely to have a police-reported injury. Interestingly, the study was inconclusive on the difference between CMB and concrete median barriers due to conflicting results (Singleton et al., 2018). This type of study addresses how CODES can be used to evaluate transportation technology in its local context by providing a base that can be built upon to fit the specific need of the study. In this case, using both the CODES database and a database containing CMB road features allowed the KTC to evaluate the use of different medians types efficiently.

Another study conducted in South Carolina used the preexisting CODES program to compare non-fatal crashes among teen drivers to non-fatal crashes among adult drivers. By combining crash data and injury data from hospital records, injury-inducing crashes involving teen drivers were able to be analyzed in context. It was found that teen drivers ages 15–17 in South Carolina had 2.5 times the single vehicle nonfatal crash rate per licensed driver and 11 times the rate per vehicle mile traveled (Shults et al., 2019). The study also provided valuable insights into teen driving behavior. For example, all

passengers were greater than 5 times as likely to be restrained in a crash if the teen driver was restrained which illustrates the importance of seatbelt use among teens. Speeding was also found to be of concern as teen drivers were cited as speeding at the time of the crash nearly twice as frequently as adult drivers and 60 percent of teen driver crashes involved speeding. Thus, programs tailored to reducing crashes among teen drivers should focus on interventions that target seat belt use and speeding violations.

2.6 SUMMARY

Habitual Traffic Offender (HTO) status in the state of Oregon is clearly defined by the Oregon legislature as an individual who has been convicted of at least three traffic offenses or at least 20 moving violations in the span of five years. License revocation is the primary intervention used by the program, and the maximum penalty is a five-year license revocation. However, the last known evaluation of the program occurred in 1986 under Dr. Jones at the Oregon DMV, and while the conclusion reached was that the HTO program was effective, driving habits may have changed in the past four decades, as well as improvements in crash data availability, statistical methods, and software. These shifts in behavior and methods warrant additional evaluation of the program as it functions today. Currently, there are many gaps in our understanding of how the Oregon HTO program influences driver behavior, ultimately making roads safer for the traveling public. A few examples of these include the efficacy of license revocation as an intervention tool, the efficacy of HTO status notification through certified mail, the efficacy of driver improvement courses offered during the program, and the role that other intervention programs like the Oregon Driver Improvement Program (DIP) play.

Other states in the Western United States also have HTO programs, but there are differences in the way the programs are structured, which may prevent direct comparisons. For example, while there have been studies done about revoked licenses in California, the HTO program in California is based on a demerit point system. This introduces uncertainty when trying to relate the resulting driver behavior in California to the resulting driver behavior within Oregon. The state with the most similar program is Washington; however, this does not lend any insights into the efficacy of Oregon's program due to a similar lack of scientific literature surrounding the subject. The absence of research on the efficacy of conviction based HTO programs in the Pacific Northwest is another gap that has been identified by this literature review. To compensate for the lack of recent research in Oregon surrounding the HTO program, literature pertaining to other programs was explored. It was found that in South Korea, license revocation was an effective deterrent against future traffic offenses (Lee et al., 2018). This does support the use of the tool, but it is hard to make direct comparisons due to the large cultural differences between South Korea and the United States. Driving behavior is linked to cultural attitudes, and there is no way to control for this discrepancy when looking at past studies. This also supports the need for research conducted on license revocation within the Pacific Northwest specifically.

The three data sets being used to test the HTO program in the state of Oregon are crash, DMV, and adjudication data. The likely methods to be used for linkage of these three datasets are a combination of probabilistic and deterministic linkage. Deterministic linkage relies on using personal identifying information (PII) to directly compare cases and see if they are a match. Probabilistic linkage involves creating a statistical model that compares the likelihood that two cases are a match. These methods make it possible to combine large quantities of data into a

complete set on which statistical tests can be performed. It is important to have complete data sets when doing traffic safety analysis because the more information is available about the intersection of demographic and traffic data, the better complex behavioral trends can be analyzed to support improved decision making. A commitment to the methods of data linkage to be used for this project will be robustly documented as part of Task 3. The integration of DMV and adjudication data, particularly when linked with crash records, provides a rich dataset for understanding and improving road safety. Through detailed visualizations and statistical analyses, researchers can reveal patterns and trends that inform more effective policies, driver education programs, and enforcement strategies, ultimately aiming to reduce traffic-related injuries and fatalities.

3.0 RESEARCH METHODOLOGY

The Linking of Oregon Driver Records and Crash Data to Evaluate Interventions and Mitigate Driver Risk incorporated various methods to collect, validate, link, and analyze required data sets. One data set was provided by the Transportation Data Section Crash Analysis Reporting Unit (i.e., crash data) and three data set were provided by the Oregon Driver and Motor Vehicle Services Division (i.e., accident data, verdict data, and driver record data).

The chapter includes the identification and description of required data sets, variables, and the specification of the analysis techniques. It also documents the proposed data sets for collection, including data availability and data quality. The methodology describes the kinds of deterministic and probabilistic techniques used to link items across different data sets.

3.1 DATA SETS

Four primary datasets were used to accomplish the overarching research goal. These included Oregon crash data, accident data, verdict data, and driver record data. The general characteristics of these datasets are described in the following subsections.

3.1.1 Crash Data

The crash data was provided through the Transportation Data Section Crash Analysis Reporting Unit, which contains a comprehensive source of traffic crash records spanning from 2002 to 2022. This dataset, stored in a Microsoft Access database, encompasses an array of information detailing the situation surrounding each crash. Each year of the dataset contains 48 tables of factors and two sources of data with 89,000 rows and 50,000 rows, respectively. This includes factors such as location, time, severity, and contributing environmental elements. The crash data covers the entire state of Oregon at the regional and district levels, as shown in Figure 2.1. With its detailed coverage, this dataset is a valuable resource for understanding traffic safety trends and identifying risk factors associated with driver behavior and road crashes. By leveraging this data, the project conducted in-depth analyses to reveal patterns, correlations, and causal relationships that can inform the development of targeted safety interventions and regulatory policies. Moreover, efficient data and analysis were facilitated by the structured database format, allowing researchers to extract valuable insights for evidence-based decision-making.

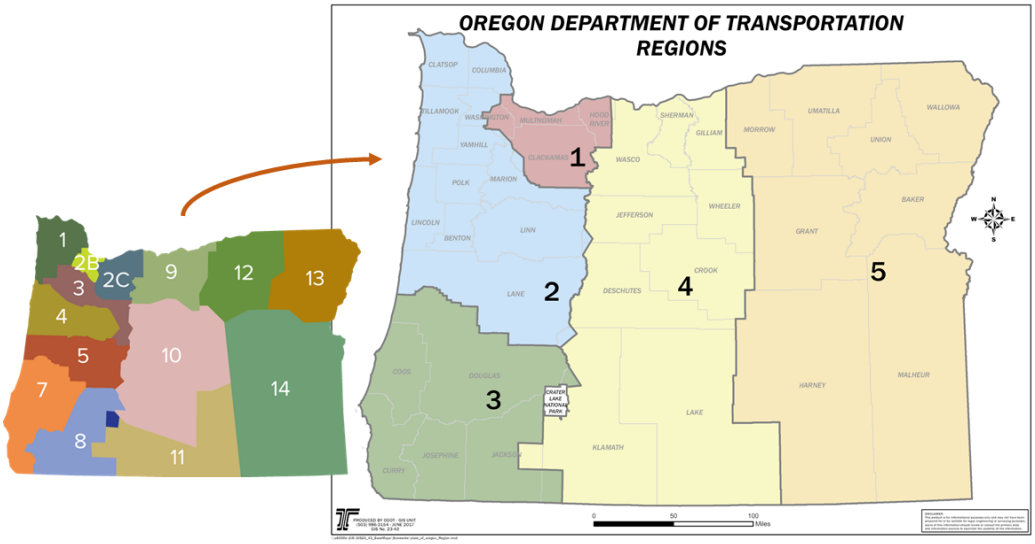


Figure 3.1 Crash Coverage across ODOT's Five Regions and 14 Maintenance Districts

3.1.2 Accident Data

The accident data, sourced from the Oregon Driver and Motor Vehicle Services, offers the potential for additional insights into traffic incidents, in addition to providing the bridge for linking driver records to crash and verdict records. It is worth noting that crash is the preferred term for a traffic incident, rather than accident. In this report, the word "accident" is only used to describe data from the accident dataset to avoid confusion with data from the crash dataset. Covering the period from 2013 to 2023, this dataset is contained within a single Excel file comprising approximately 900,000 rows of useful recorded information. It includes demographic profiles of drivers, detailed descriptions of accident types, dates, and jurisdictions, as well as outcomes such as insurance details. T As this data is maintained by the Oregon DMV, it could serve as a complement to the information provided by the verdict data and driver record data. Together, these datasets enhance the overall data framework, providing a comprehensive basis for evaluating interventions and mitigating driver risk. By integrating this accident data with verdict and driver information, researchers will gain a multi-dimensional view of traffic incidents, which is crucial for developing targeted safety measures and effective risk reduction strategies.

3.1.3 Verdict Data

Verdict data was obtained from the Oregon Driver and Motor Vehicle Services, which covers a comprehensive set of legal outcomes associated with traffic violations of Oregonian drivers. Spanning all legal adjudications related to traffic offenses, this dataset is distributed across five Excel files, each containing approximately 1 million rows of detailed case information. This division into multiple files is necessary because a single Excel sheet cannot hold more than 1 million rows of data. Excel's row capacity limitation requires that extensive datasets be segmented to ensure that all data is accommodated without loss of information. This segmentation allows for comprehensive management and analysis of each subset of data within its respective file, facilitating more efficient data processing and recovery. From violation codes

and descriptions to adjudication outcomes and citation dates, this dataset offers a comprehensive view of the legal processes surrounding traffic violations. By analyzing these legal outcomes, the project aims to assess the effectiveness of legal penalties and enforcement strategies in modifying driver behavior and reducing instances of traffic violations. Furthermore, the dataset provides valuable insights into the deterrent effects of various penalties on repeat offenses and serious traffic crashes, thereby informing policy adjustments and enhancements to the legal framework aimed at promoting road safety and compliance with traffic laws.

3.1.4 Driver Record

The driver information dataset was provided by the Oregon Driver and Motor Vehicle Services, which constitutes a comprehensive source of demographic and driving history data for all Oregonian drivers holding valid licenses. Driver records were organized into six distinct files, each containing approximately 1 million rows of detailed driver profiles. This dataset encompasses a wide array of personal and driving-related attributes, including the full legal name, Driver's License ID, Date of Birth, and Sex. These attributes are crucial for constructing detailed driver risk profiles and identifying demographic factors that influence driving behaviors. Integrating this data with the verdict data enriches the dataset with additional factors, providing a more robust framework for developing targeted educational and regulatory interventions tailored to specific driver risks. Consequently, this approach promotes responsible driving practices and enhances the effectiveness of safety measures. Additionally, the dataset enables longitudinal studies to assess the long-term impacts of interventions on driving behaviors and road safety outcomes, facilitating evidence-based policy decisions and interventions to improve overall traffic safety.

3.2 DATA HANDLING AND LINKAGE TECHNIQUES

To ensure privacy and confidentiality, the principal investigator and associate investigator replaced all personal identifying information with unique anonymous IDs before data mining and analysis began. This means that no actual names or other PII were visible to any of the research assistants involved in the project. The tasks described in the following subsections were performed before any further analysis was conducted.

3.2.1 Data Cleaning

In this crucial preliminary phase, our research team applied careful data-cleaning techniques across all four Oregon datasets (i.e., crash data, accident data, verdict data, and driver record data). Each dataset underwent a detailed cleaning process where null values are eliminated, duplicates resolved, and inconsistencies corrected to ensure the highest data integrity. Automated scripts in Excel, Python, and R Studio were extensively deployed to detect and correct spelling errors and other common data entry inconsistencies. For more complex discrepancies, such as conflicting data entries across different datasets, manual reviews were conducted. Additionally, the research team trained research assistants to perform these cleaning tasks, ensuring thorough preparation of the data. These methods were applied consistently across all data sources. This comprehensive approach not only prepared the data for effective linkage but also ensured the reliability of subsequent analyses.

3.2.2 Quality Check

After the initial data cleaning, a comprehensive quality check was essential to validate the effectiveness of the cleaning processes. This phase was managed by a different group of research assistants than those involved in the data cleaning, ensuring the validity of the results. These students were trained by the research team to use Excel software for conducting quality control checks. Thorough validation techniques, including range checks for numerical data and consistency checks for categorical data, were implemented. For instance, the verification of the verdict data involved thorough cross-validation with the driver's license and name, ensuring that all legal outcomes are correctly captured and accurately represented in the dataset. This thorough validation process was pivotal in maintaining data integrity, providing confidence in the reliability of the data before proceeding to the linking phase.

3.2.3 Deterministic Approach

In the deterministic linkage phase, unique identifiers such as driver's license IDs, DOB, and names were used to link datasets with a high level of precision. This task was conducted exclusively by research assistants and senior researchers in conjunction using Python to combine all data. This method involved matching records across datasets where exact matches of identifiers were found, ensuring the reliability of the linkages. For example, the deterministic linkage between the driver information dataset and the verdict data was facilitated by the precise matching of unique identifiers using the DOB, name, and driver's license ID. This combined data was then incorporated with additional attributes such as gender and date of birth from the driver information data, and was assigned a specific name (e.g., *combined-1*) for clarity and ease of further processing. Once this initial linkage was complete, the data was broken down into Excel sheets to facilitate subsequent QA/QC processes.

This thorough approach allows for the seamless integration of data from different sources, which was crucial for comprehensive analyses that relied on different data inputs. Afterward, the processed data set (*combined-1*), which now included detailed demographic and identification data, was linked with the accident data using identifiers such as DOB, gender, name, and driver's license ID. The more identifiers included, the more accurate the resulting dataset. This merged dataset was again given a new name (e.g., *combined-2*). Finally, QA/QC was performed on this integrated dataset by research assistants before pivoting to the analysis phase, ensuring the data's accuracy and completeness.

3.3 ANALYSIS PLAN

3.3.1 Visualization

In the visualization strategy, a variety of dynamic visuals were developed by the research team using Python to facilitate deep insights and easy communication of findings. These visuals were complemented by time series graphs created in R, which were used to locate trends in traffic crashes over time, revealing seasonal patterns or long-term changes in crash rates. Bar charts and pie charts were employed to depict categorical data such as gender, or types of crashes and their outcomes, providing a clear view of proportions and comparisons. These varied forms of visualization will not only make the data more accessible but also enhance the decision-making

process by presenting complex data in an engaging and understandable way. It also helped validate some of the subsequent analyses.

3.3.2 Descriptive Statistics

Descriptive statistics were calculated using Python, provided a detailed summary of the datasets, adding depth to the insights gained from visualizations. Measures of central tendency, such as mean, median, and mode, were calculated to summarize typical crash characteristics and identify outliers. Measures of variability, including the range, variance, and standard deviation, were used to understand the dispersion and consistency within the crash data, which is crucial for assessing the reliability of the findings. Frequency distributions counted the occurrence of specific crash types, highlighting common risks and hazards. Through cross-tabulations, relationships between categorical variables, such as sex and DUI, were explored, uncovering patterns that informed preventive measures. Additionally, percentiles and quartiles were calculated to segment data into meaningful groups, such as identifying particularly age group. Together, these descriptive statistics helped us understand the data, providing a robust basis for comprehensive analysis, and facilitating the next steps when conducting advanced statistical modeling.

3.3.3 Statistical analysis

The statistical analysis used a comprehensive approach to dive deeply into the various components of data sources collected across Oregon, employing two distinct types of analysis to capture a broad spectrum of insights:

Basic Regression Analysis:
The research team applied multiple linear regression analysis to explore potential relationships across various traffic-related variables. This analysis helped identify patterns that might not be immediately obvious, such as the impact of road conditions, causes of crashes, and temporal factors on crash occurrences. Alongside regression, a set of additional statistical tests were employed, including ANOVA to compare means across multiple groups, chi-square tests to examine relationships between categorical variables, and both parametric and non-parametric tests to assess data properties without assuming a normal distribution. Furthermore, correlation matrices were constructed to visualize and quantify the strength and direction of relationships between variables. Using R, known for its robustness and flexibility, the team systematically analyzed data, providing a robust framework for identifying key factors that could influence traffic safety and policy decisions.

4.0 DATA SECURITY PROTOCOLS

The data management and protection protocols for this project are designed to meet the standards of security set by Oregon State, following best practices. This chapter describes the data management measures used to maintain the integrity and confidentiality of the data, ensuring secure and responsible handling throughout the research process.

4.1 PROJECT DATA SECURITY PLAN

The Office of Information Security (OIS) at Oregon State University oversees all data management practices within the university. As part of this project, a data security plan was created with the support of Tom Ordeman, the Governance, Risk, and Compliance Manager of OIS. The data collected for this project is classified as “confidential information”, as it includes personally identifying information (PII) that presents a serious risk to individuals if it is exposed. An example of confidential information used in this project are driver's license numbers, which can be used in conjunction with other identifying information to steal an individual's identity if leaked. As such, the data security plan was carefully crafted to follow data security best practices as recommended by OIS.

4.1.1 Data Security Plan and OIS Approval

4.1.1.1 Data Security Plan Standards

The data security plan lays out several key components of secure data handling: safe file sharing, storage, and access. Oregon State University Baseline Standards of Care were used as a guideline for the security plan. These standards outline requirements for elements such as the network monitoring, access restrictions, and file encryption necessary to fulfill the threshold of security for confidential data. The data security plan fulfills all of the requirements in the Baseline Standards of Care, as they represent best practices for data security.

After careful review to ensure compliance with the above standards, the data security plan was approved on August 23, 2024 by Max Simon, the Outreach and Awareness Coordinator for OIS, in conjunction with Tom Ordeman. The approval was confirmed by email with the project Principal Investigator.

4.1.1.2 Data Security Plan Compliance

In accordance with the data security plan, SharePoint was used to share and store the data files for this project. This was because SharePoint meets the minimum 128-bit symmetric-key algorithm encryption standards set by OIS. Encryption prevents files from being read by unwanted parties, which protects the confidentiality of the data. Due to the high level of risk associated with confidential information and PII, SharePoint encryption was used to prevent the possibility of dangerous third-party data leaks.

To increase security further, only the Principal Investigator and Co-principal Investigator had access to data containing PII. Access to the PII was controlled by a username assigned by OSU IT, an alpha numeric password that changes every 6-months, and dual authentication through a smartphone app and the inputting of a randomized three-digit code. The reduced level of access to PII protects it by lowering the number of times that the files could be compromised during transfer and analysis.

The work done by research assistants took place on datasets that have been anonymized. In addition to the fact that they will not be working with PII, all research assistants are certified by the Collaborative Institutional Training Initiative (CITI) in social and behavioral research. This training provides instruction on maintaining ethical conduct during research using human subjects. Research assistants also had to take continuing courses on data security through the university. When working with the datasets, student researchers used computers within the locked transportation lab. These computers use the Oregon State network that is continuously monitored by IT specialists for suspicious activity. By limiting which data student researchers can access and where they can access it, there are less chances for the data to be compromised through human error.

4.2 INSTITUTIONAL RESEARCH BOARD EXPERIMENT APPROVAL

The Institutional Research Board (IRB) at Oregon State University oversees all human subjects research within the university to protect the rights of subjects. This keeps all OSU experiments within the rules set forth by the Department of Health and Human Services in regards to human subject rights and treatment.

4.2.1 IRB Application

This study did not fulfill any of the existing IRB exemptions, so the full IRB application was proposed for review. This included a study overview, a methodology summary, participant information, proposed data management practices, proposed data security measures, and a record retention plan in addition to other smaller elements. These sections outlined every procedure the data underwent and the protections in place to ensure the integrity of the study. Because this study contains confidential information (including PII), additional information was requested in regards to data management and security to ensure that the study was protecting the identity of participants.

In addition to the full proposal, the IRB requested three additional documents. The first was the data security plan and its receipt of approval from OIS. The second was the scope of work provided by ODOT that authorized the use of confidential data. The third was an Excel file containing every data attribute used within the project. This last item was used to ensure that all direct participant identifiers were acknowledged in the data management section of the proposal. The combination of these three documents confirmed to the IRB that the data was being attained safely and legally.

4.2.2 Approval

The IRB application was certified by the PI on October 9th, 2024. It was approved the same day by IRB analyst Adeline Oka (See Appendix B).

5.0 DEVELOPMENT OF LINKED DATABASE

5.1 OVERVIEW AND OBJECTIVES

The integration of multiple datasets was an essential step in this task to create a comprehensive and reliable resource to achieve the goal of this project and for proper analysis in the later tasks. This task focused on merging four primary datasets: Driver, Verdict, Accident, and Crash data. Each dataset originated from separate sources and contained unique information, making them invaluable for understanding traffic safety and driver behavior. The merging process was designed to combine these datasets systematically while addressing data inconsistencies and redundancies.

The primary objective was to generate a unified dataset that could support in-depth analysis and provide insights into better understanding the most at-risk drivers. This would provide ODOT and DMV important information that could help inform strategies for reducing fatal and severe injury crashes. A systematic approach was followed to ensure that data integrity was maintained throughout the process, despite the challenges posed by inconsistencies in formatting, naming conventions, and incomplete entries. By integrating these datasets, the resulting resource offers a multidimensional view of traffic safety, ultimately aiding in policy development and targeted interventions.

5.2 DATASETS

Four datasets were used to execute this task. Each dataset had its unique challenges that had to be addressed during the merging process. The complexity of the data required a structured approach to ensure that no critical information was lost during integration.

Table 5.1 provides a brief description of each of the four datasets, as well as the challenges of each dataset.

Table 5.1 Summary of Datasets

Dataset	Owner	Description	Relevant Data Fields	Challenges
Driver Data	Oregon DMV	A comprehensive source of demographic and driving history data for all Oregonian drivers holding valid licenses. It is organized into six distinct files, each containing approximately 1 million rows of detailed driver profiles.	Full legal name Oregon Driver's License (ODL) Date of Birth Sex	Duplicate names Inconsistent use of unique ODLs Empty cells
Verdict Data	Oregon DMV	This dataset documented legal outcomes of traffic violations. This dataset was critical for understanding the enforcement and legal consequences of driver behavior.	Violation Codes Violation Descriptions Citation Date Verdict Date	Duplicates due to multiple citations Careful cleaning needed to ensure that records were not removed incorrectly
Accident Data	Oregon DMV	Documentation on specific traffic incidents that provided information on accident outcomes.	Involved parties Accident types Outcomes	Inconsistencies in name formatting Erroneous data entries
Crash Data	ODOT	This dataset offered more detailed records of crash events.	Environmental factors Roadway factors Crash severity Locations	Dataset required integration with other datasets to provide useful information

5.3 PROGRAMMING LANGUAGE

Python was selected for this task due to its efficiency and computational speed, particularly when managing large datasets. The amount of data in this project required a tool that could handle millions of rows across multiple datasets while performing complex operations accurately and quickly. Python provided the necessary environment for automation, enabling repetitive tasks such as duplicate detection, formatting check, and linking records to be executed systematically and consistently. The merging process also required advanced data cleaning techniques to address issues such as inconsistent naming conventions, missing data, and duplicate entries.

Real-time quality control was another critical requirement. Python enabled the implementation of automated QA/QC processes, which were essential for validating data integrity after each stage of merging. These processes included checks for duplicate records, formatting errors, and mismatched identifiers. By ensuring continuous validation throughout the merging process, Python helped maintain the reliability and accuracy of the final dataset. In addition to its technical capabilities, Python offered scalability and reproducibility. Custom scripts were developed to perform the merging process to the specific needs of this project, while also

ensuring that the steps could be replicated for future updates or analyses. Even with Python, and operating on significantly powerful computers, merging trials ran as long as 30 hours per attempt.

5.4 DATABASE CONSTRUCTION

5.4.1 Data Preparation

Data preparation was an essential step to clean and organize the datasets before merging. It involved removing duplicates, standardizing formats, and resolving inconsistencies to ensure everything was prepared correctly for the next steps.

5.4.1.1 Driver Data

The cleaning process for Driver Data involved addressing duplicates, ensuring proper formatting, and preparing the dataset for merging with other sources. The total observations initially received were 5,958,859, distributed across six Excel files. These files were converted to a CSV format for computationally faster reading and easier uploading and downloading in Python, and then concatenated into a single dataset before performing the following steps. However, several inconsistencies and duplications required attention before further analysis. The variables used included:

- Name
- ODL
- DOB
- Sex
- First Issued
- Latest Expiration
- Experience (years) = Event Date – First Issued

5.4.1.2 Duplicate Removal

The dataset passed through several quality checks to identify and handle duplicate records:

- **By Name:** Initial checks by name revealed 405,818 duplicate records. However, these were not dropped because some drivers shared the same names but had different ODLs, making them unique entries.
- **By DOB:** Similarly, duplicates could not be identified using only DOB, as some drivers shared the same date of birth.

- **By ODL:** Duplicate ODLs totaled 52,506 observations.

- **By Name, ODL, and DOB:** These fields combined confirmed 54,005 duplicates, aligning with the results based on ODL only. These duplicates were carefully reviewed, resulting in 27,343 entries being dropped while retaining 26,662 unique entries. The duplicates exist because some records share the exact same ODL, name, and DOB, but discrepancies appear in the gender field, where it is either different or empty. Additionally, some records are duplicated entirely.

- **By Name and DOB:** This check was necessary as it addressed cases where ODLs were duplicated. A total of 1,503 duplicate entries were identified. Of these,752 records had two ODLs for the same name and DOB, while one driver had three different ODLs. These duplicates were dropped, leaving a final dataset with 5,958,105 unique observations out of 5,958,859. They were dropped because they had either empty cells or different genders. It would also be hard to identify them if we keep both, especially when merging them later.

5.4.1.3 Formatting Adjustment

To ensure consistency and compatibility, the following formatting adjustments were made:

- **DOB Formatting:** The DOB formatting step involved converting the original date of birth (DOB) field in the dataset into a consistent and standardized date format. This process ensured that all DOB entries follow the same structure, making them easier to analyze and merge with other datasets. This was achieved through two coding steps:

    - **Parsing Dates:** Converting DOB entries into a proper date format (e.g., YYYY-MM-DD). This was important because dates may originally appear in various inconsistent formats, such as MM/DD/YYYY or text-based formats (e.g., "January 1, 2024").

    - **Error Handling:** This process ensured that invalid or incorrectly formatted dates were automatically converted to "NaT" (Not a Time) rather than causing errors in the process. This made it easier to identify and handle problematic entries later.

- **Name Formatting:** Names in the original dataset included extra spaces after the last name. These spaces were removed to prevent issues such as mismatches during merging with other datasets.

- **ODL Formatting:** ODLs were converted to string format to accommodate variations where some ODLs were purely numerical, while others included letters. This ensured compatibility with other datasets.

The cleaning process ensured that the Driver Data was both accurate and uniformly formatted, ready for integration with Verdict, Accident, and Crash data in subsequent steps. A total of 754 records still had missing values in the "Sex" field out of the final dataset (5,958,105 unique entries).

5.4.1.4 Verdict Data

The cleaning process for Verdict Data focused on addressing duplicates, ensuring proper formatting, and preparing the dataset for merging with other sources. A total of 22,856,683 observations were initially received, distributed across five Excel files. These files were converted to CSV format to enable faster computational processing and easier uploading and downloading in Python. They were then concatenated into a single dataset for subsequent examination and cleaning to address specific challenges and inconsistencies.

5.4.1.5 Duplicate Removal

Unlike Driver Data, checking for duplicates in Verdict Data by Name or ODL alone was not feasible because drivers could have multiple records for different violations. Instead, duplicates were identified based on a combination of variables. The variables used included:

- Name
- ODL
- DOB
- Sex
- Verdict ODL
- Citation Date
- Verdict
- Violation Code

Using this approach, 324,830 duplicate records were identified and out of those, 169,612 removed, reducing the dataset to 22,687,071 unique observations. This step ensured that each record represented a distinct violation while preserving the integrity of the dataset.

5.4.1.6 Formatting Adjustments

Consistent formatting was applied to ensure compatibility and accuracy across key fields:

- **Citation Date and Verdict Date Formatting:** The citation and verdict date fields were reformatted to a standardized date format (e.g., YYYY-MM-DD). This

process ensured consistency and allowed for accurate temporal analyses. Invalid or improperly formatted dates were converted to "NaT" (Not a Time) for further review.

- **Name and ODL Formatting:** Similar to Driver Data, names were cleaned to remove extra spaces, and ODLs were converted to string format. These adjustments ensured compatibility with other datasets where name and ODL formatting may vary.

The cleaning process reduced the total number of records to 22,687,071 by addressing duplicates and ensuring uniform formatting across all fields. These steps prepared the Verdict Data for smooth integration with Driver, Accident, and Crash data in subsequent steps of the analysis.

5.4.1.7 Accident Data

The cleaning process for Accident Data focused on handling duplicates, ensuring proper formatting, and preparing the dataset for integration with other sources. Initially, 872,731 observations were received in a single Excel file, which was later converted to CSV within the code itself, requiring careful cleaning to address challenges specific to this dataset.

5.4.1.8 Duplicate Removal

Checking for duplicates using Name, ODL, or DOB alone, or even combined, was not feasible because some drivers had multiple accident records. Instead, duplicates were identified by including the accident date along with other common identifiers. The variables used included:

- Name
- ODL
- DOB
- Accident Date

The Accident Date was used because it is unlikely for a driver to have multiple accidents on the same day. It is worth mentioning that the Accident ODL could not be utilized, as some observations were assigned different Accident ODLs despite occurring on the same day and in the same county. Using this approach, 1,892 duplicate records were identified, of which 966 were removed, leaving 926 retained as unique crash events. This step ensured that the dataset accurately represented distinct crash incidents while preserving essential data for analysis.

5.4.1.9 Formatting Adjustments

The following formatting issues were addressed to standardize the dataset:

- **Name Formatting:** Names in the Accident Data were consistent, with no extra spaces or irregularities, so no adjustments were necessary.

- **Boolean Variable Standardization:** The dataset used "False" and "True" to represent binary variables such as crash severity (e.g., whether the crash was fatal or not). These values were converted to "0" and "1," respectively, to facilitate subsequent analysis.

- **DOB Formatting:** The DOB field was reformatted to a consistent date format (e.g., YYYY-MM-DD), similar to the cleaning processes for Driver and Verdict Data. This ensured compatibility during the merging process.

The cleaning process reduced the dataset to 871,765 unique observations out of 872,731 by addressing duplicates and standardizing variable formats. These adjustments ensured the Accident Data was accurate, consistent, and ready for integration with Driver, Verdict, and Crash data in subsequent analyses.

5.4.2 Data Merging

5.4.2.1 Sequential Validation of Pairwise Merges

This phase started after cleaning all the datasets to ensure they were ready for integration. Pairwise merging was conducted systematically, combining two datasets at a time to maintain accuracy and consistency. Performing these incremental merges not only facilitated the validation of data at each step but also helped in identifying and resolving inconsistencies early in the process. This step-by-step approach acted as a QA/QC measure, allowing for bidirectional validation to ensure alignment between datasets before proceeding to the final merge. Such a methodical process ensured the reliability of the integrated data for subsequent analysis.

Driver and Verdict Data

The integration of Driver Data with Verdict Data was the first step in the merging process. Various combinations of linking criteria were employed to maximize the accuracy of matches while addressing data inconsistencies.

LINKING METHODS

- **Name only:** Addressing cases where ODLs were missing or inconsistent.

- **ODL only:** Focusing on individuals with consistent ODLs across datasets.

- **Name and ODL:** Capturing records where both Name and ODL matched.

DATA VARIATIONS

- **Similarities in Names Causing Inaccuracies:** Some records exhibited similarities in names, which is a common occurrence in the database. It is possible for two individuals to have identical names. For example, "Sarah Marie Johnson" in Driver Data with ODL 123456 could inaccurately match with "Sarah Marie

Johnson" in Verdict Data with ODL 12345. These similarities resulted in additional matches when merging by Name only.

OUTCOMES OF UNIQUE DRIVER MERGING:

- **Merging by Name only:** This method produced 5,915,850 matches, capturing cases where names were similar, but ODLs were inconsistent or missing. However, as mentioned earlier, this approach increased the number of matches, which required careful review to avoid mismatches. This approach resulted in 644,672 unmatched drivers in the verdict data (i.e., drivers in the verdict that could not be found in the driver Oregon data).

- **Merging by ODL only:** This method identified 5,956,931 matches by capturing similar ODLs without considering names. The downside of this approach was that some individuals from different states or regions shared similar ODL structures, or ODLs had minor inconsistencies, leading to inaccuracies.

- **Merging by Name, DOB, and ODL:** This method resulted in 5,956,931 matched records, which served as the final dataset for subsequent analyses. It provided the most reliable matches by ensuring both Name and ODL were consistent between the datasets before executing the final data merge. Note that the final observations were slightly fewer compared to merging by ODL only due to the presence of data from other states in the Verdict database, which matched with some drivers from Oregon. This approach resulted in 603,591 unmatched records in the verdict data (i.e., drivers in the verdict that could not be found in the driver Oregon data).

CHALLENGES WITH MERGING

- **ODL Constraints in Verdict Data:** The Verdict Data initially only contained ODLs for drivers whose ODL numbers started with "6." This limitation excluded many records, but it was resolved through additional data requests from the DMV.

This step established a foundation for subsequent mergers by addressing critical inconsistencies and ensuring reliable links between Driver and Verdict Data. The final dataset of 5,956,931 unique drivers and 24,166,575 records provided a robust basis for further integration and analysis.

Accident and Crash Data

The integration of Crash Data with Accident Data was chosen as the second merge pair due to the datasets having similar sizes. Various combinations of linking criteria were employed to maximize the accuracy of matches while addressing inconsistencies between the datasets. To reduce the size of the data files, datasets were separated and linked by year to facilitate easier data handling.

LINKING METHODS

The merging process utilized the following linking method:

- **Serial Number only:** Focused on using a unique identifier for each event to ensure consistency between datasets.

EXAMPLES OF DATA VARIATIONS

- **Formatting Discrepancies:** The CDS technicians used the serial number provided by the DMV but added leading zeros and numerical prefixes based on attributes such as duplicate entries. This required the use of coding logic to accurately match records based on the differences that could arise. A new, unique ID number was generated from the serial number and incident date to mitigate this issue, as well as the county number.

OUTCOMES OF MERGING

- **Merging by serial number and date of incident:** Resulted in 627,206 matched records. This method proved to be the most reliable, as IDs were generally consistent between datasets.

CHALLENGES WITH MERGING

- **Serial Number Formatting Discrepancies:** The ODOT CDS data used a different serial number format that builds off the DMV format. This reformatting process was not clear and required the use of the CDS Crash Manual to implement data cleaning before the files could be merged.

This step established a solid foundation for further integration with Verdict and Crash Data by addressing key inconsistencies and ensuring reliable links between Crash and Accident Data. The final dataset (627,206) provided a robust basis for subsequent merging and analyses.

5.4.2.2 Final Merge

The final merge that took place combined the merged Driver and Verdict Data with the merged Accident and Crash Data. This was done using identifiers such as legal name, DOB, ODL, and sex. The final merged dataset had 24,629,556 records and provided a more complete picture of crash data in the state.

5.4.3 Quality Assurance and Quality Control (QA/QC)

5.4.3.1 Manual QA/QC Workflow

A robust QA/QC process was implemented at each step of the cleaning and merging process to ensure data accuracy and integrity. After each merge: Driver with Verdict Data, Driver with Accident Data, and so on, random samples were selected for manual verification. Approximately 300 observations were randomly exported to Excel and reviewed line by line by research team members. This process involved highlighting inconsistencies, documenting observations, and adding notes directly to the Excel sheets.

The QA/QC process was not limited to individual reviews. To further ensure accuracy, data sets were swapped among researchers so that inter-rater reliability checks could be performed. Each researcher independently reviewed a subset of the data previously reviewed by another, and results were compared for consistency. This approach helps to minimize human error and reinforced the reliability of the manual verification process.

5.4.3.2 Tool-based QA/QC Workflow

The primary tool for the merging process was Python, which was used to automate the identification and resolution of common data issues. Automated checks were conducted at each step to address problems such as:

- **Duplicates:** Python scripts identified and flagged entries with the same Name, DOB, and ODL. For example, "John Thomas Smith" appearing multiple times in Driver Data with slight variations was identified and resolved.
- **Null Values:** Entries with missing DOBs or sex information, such as the 431 drivers with no sex information, were flagged for further review.
- **Formatting Inconsistencies:** Names with spaces in Driver Data but not in Accident Data, such as " John Thomas Smith" versus " John Thomas Smith’ ‘" were standardized to ensure accurate matches. The QA/QC process combined Python-based automation with manual reviews to achieve a high level of confidence in the merging results.

5.4.3.3 Data Validation and Reliability Checks

Validation of the merging process was conducted by comparing Python-generated results with manually reviewed samples. The 300 randomly selected observations were reviewed in Excel, and each entry was cross-verified with Python outputs. Notes and discrepancies were documented for every observation, and feedback loops ensured that corrections were incorporated into the Python scripts where needed.

After completing the manual validation, all merging processes demonstrated a 100% success rate, with the manually verified results matching exactly with the Python outputs. This consistency was further supported by inter-rater reliability checks, where researchers independently reviewed the same data sets and reached identical conclusions.

The validation process confirmed the accuracy and reliability of the Python-based merging process, ensuring that the final integrated dataset met the highest standards of data quality. These steps provided confidence in the dataset's appropriateness for further analysis and research applications.

5.4.4 Summary of Challenges

The development of the linked database presented several challenges due to inconsistencies and complexities within and across the datasets.

Table 5.2 summarizes the challenges encountered over all four datasets during the merging process.

Table 5.2 Challenges of Merging Datasets

Challenge	Description
Name Formatting Issues	Names in the datasets exhibited significant formatting discrepancies. For instance, Driver Data included spaces after last names, while Accident Data omitted these spaces. Additionally, middle names in Accident Data were often reduced to initials or omitted entirely, leading to a complicated name-based merging.
Duplicate Records	Duplicate entries were predominant across all datasets. Identifying and handling these duplicates required combining multiple fields such as Name, ODL, and DOB to distinguish true duplicates from legitimate multiple records for the same individual.
Date of Birth (DOB) Errors and Formatting	DOB fields in the datasets contained errors such as invalid dates (e.g., "06/01/9998" instead of "06/01/1998") or blank cells, which reduced the effectiveness of merges involving DOB. Furthermore, DOB formatting varied across datasets. Adjustments were necessary to standardize DOB entries to the MM/DD/YYYY format, ensuring consistency during integration.
Identifier Formatting	ODLs in Driver Data were treated as numeric fields, while Accident Data stored them as strings without spaces. This inconsistency required reformatting to ensure compatibility during merging.
Overlapping Identifiers Across States	Some ODLs in the datasets matched with records from other states, complicating merges that relied on ODL alone.
Name Changes Over Time	Drivers changing their last names, often due to marriage, or using middle names as first names or children of divorced parents alternating between either parent's last name, which created additional complexities when names were used as a linking criterion.
Incomplete Data Fields	Some critical fields in the datasets were incomplete. For instance, the "Sex" field in the Driver Data contained missing entries, and blank or invalid DOB cells were present in Accident Data, reducing the reliability of these attributes in merges.
Quality Assurance Limitations	Due to the large number of observations, with millions of rows in each dataset, it was impractical to perform manual QA/QC on every data record. Instead, hundreds of rows were randomly reviewed in each iteration. Although extensive QA/QC measures were implemented using Python, including bidirectional validation processes, the reliance on sampling rather than exhaustive review introduced a potential margin of error.

These challenges necessitated meticulous cleaning, formatting adjustments, and merging strategies to create a cohesive and reliable integrated database.

5.5 DATABASE CONSTRUCTION SUMMARY

The development of the linked database involved a systematic process of cleaning, formatting, and merging four primary datasets; Driver, Verdict, Accident, and Crash Data, to create a unified resource for analyzing traffic safety and driver behavior. Thorough data preparation was carried out and addressed challenges such as duplicate records, inconsistent identifiers, and formatting discrepancies, ensuring the datasets were accurate and compatible. The merging process was conducted sequentially, starting with pairs of datasets and ending in a comprehensive integration using ODLs and probabilistic methods. Despite limitations, such as incomplete fields, reliance on sampling during QA/QC, and the use of spatial and temporal methods for Crash Data integration, the final database successfully consolidated millions of records while maintaining data integrity. This linked dataset provides a multidimensional view of driver actions, legal outcomes, and crash events, establishing a robust foundation for further analysis and supporting strategies to mitigate driver risk and improve traffic safety.

6.0 RESULTS

Once the final dataset was complete, the focus shifted to identifying macroscopic trends in the data that could produce valuable insights for the DMV. This included looking at citation numbers across the years, demographic analysis, the relationship of speed involved citations on speed related crashes, the relationship of DUII involved citations on DUII crashes, the relationship of the number of citations on crashes in general, and which type of citations have the highest impact on crash severity levels.

6.1 INITIAL DATASET VISUALIZATION

6.1.1 Citations by Year

The first trend analyzed was the number of citations issued per year since 1995. Prior to 1995, citation volumes were below triple digits for unknown reasons, so those years were excluded from analysis. Figure 6.1 visualizes the number of citations, which illuminates several discontinuities in the data. There are several years in which the number of citations increases rapidly before stabilizing at a higher quantity. The most prominent shift occurred between the years of 2010-2012, where the number of citations increased by close to a million additional citations. This increase is unlikely to be caused solely by a behavioral shift, and it most likely due to change in administrative methodology. The Oregon State Police switched from handwritten citations to electronic records in the early 2000s, which may account for the increase in the volume of citations. However, there is not enough conclusive evidence to connect the two events.

The only shift in the graph that was likely caused by human behavior is the variation in the graph that occurred in 2020-2022 during the COVID-19 pandemic. The pandemic significantly reduced travel during 2020, which could account for the drop in citations during the year. Interestingly, citation numbers spiked back to pre-pandemic levels in 2021 before dropping in 2022. There are not enough years of citation data post-2022 to make any definitive conclusions about the impact of post-pandemic behavior on citation volumes.

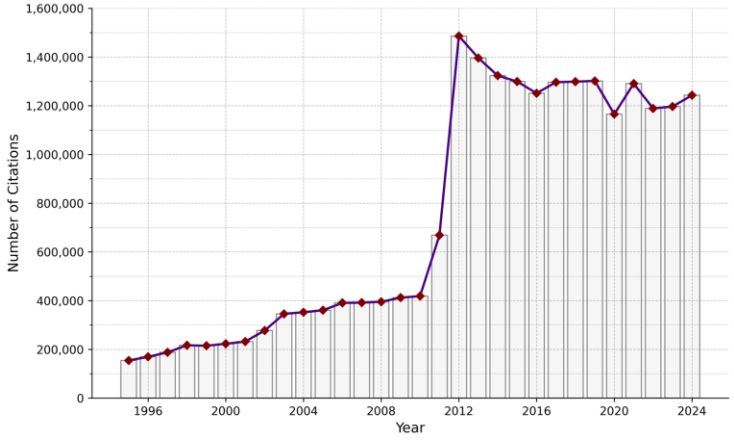


Figure 6.1 Number of Citations by Year

6.1.2 Citations by Type

This study focused on three categories of violations: DUII-related, speed-related, and habitual incident violations. Violations were sorted using DMV violation codes provided in the verdict data. Table 6.1 shows the top ten most common codes (by frequency of occurrence) that were considered in this analysis. These ten codes make up 75% of annual citation volumes, which is why this analysis focuses on only ten out of the many codes available.

Table 6.1. Ten Most Common Types of Citations

Code and Name	Description
Driver Improvement Violation	Incident Involved One or More Violations
Habitual Minor Incident	Incident representing verdicts counted as minor offenses in the habitual offender program
D56 – Failed to Answer Citation	Failure to answer a citation, pay fines, penalties and/or costs related to the original violation (detail sometimes required)
S92 – Speeding with Detail	Speeding - Regulated or posted speed limit and actual speed (detail required)
D36 – Failed to Maintain Liability Insurance	Failure to maintain required liability insurance
SR-22 Violation	Customer has let SR-22 go into suspense
B26 – Driving While License Suspended	Driving or operating a motor vehicle while license suspended
M14 – Failed to Obey Traffic Control Device	Failure to obey sign or traffic control device
D45 – Failed to for Trial/Court	Failure to appear for trial or court appearance (detail sometimes required)
Accident Uninsured	Accident Uninsured

The top ten citation types by frequency were plotted against time (Figure 6.2). For many of the top citation types, such as accident uninsured and driving while license suspended, the number of citations has stayed consistent since 2012. The citations relating to Driver Improvement Violations, Habitual Minor Incidents, and Speeding with Detail follow the same pattern from 2012 to 2024, with a dip in 2016 and a spike in 2021. This may be because speeding can be considered a habitual minor incident and may be recorded twice which influences the shape. The other trend of interest is that the citation numbers for Failed to Answer Citation and Failed to Appear for Trial or Court swap between 2019-2021. It is likely that the definition of the citation changed around 2019, and code D56 was replaced by D45. Even if this is the case, Failure to Appear for Trial or Court has a lesser volume of citations post 2020 than Failure to Answer Citation. It is unclear if this is a post-pandemic behavior shift or if the newer code was defined in a way than excludes previous forms of citations that occurred under the old definition.

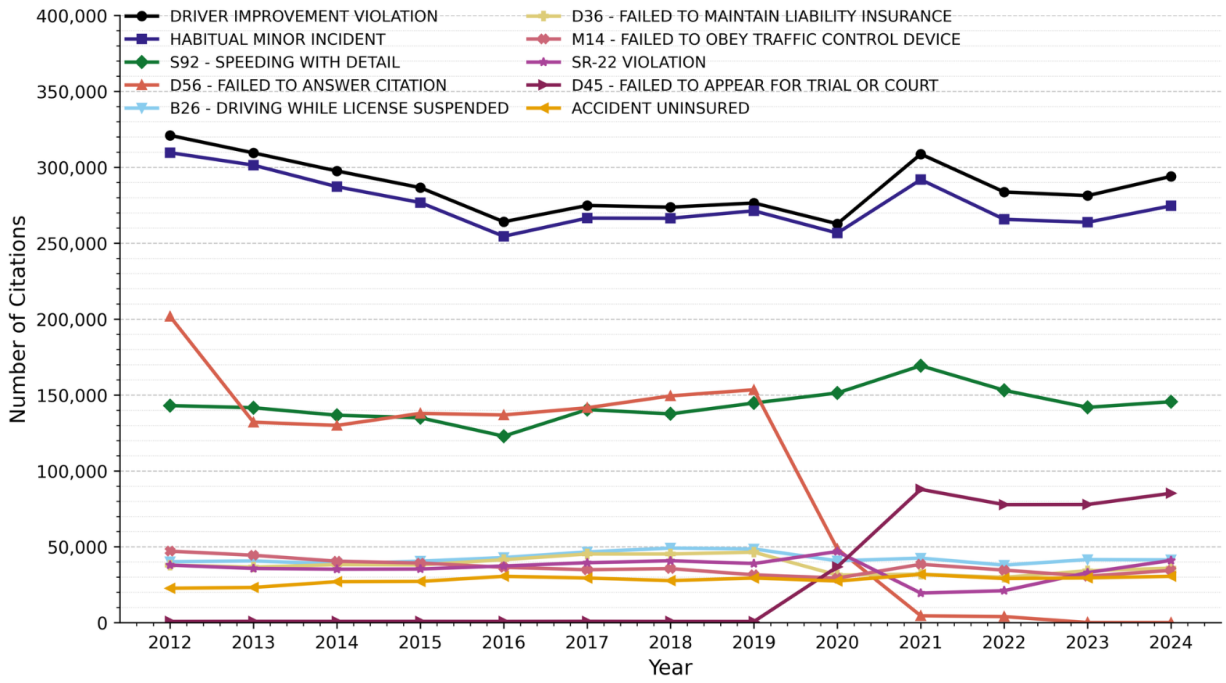
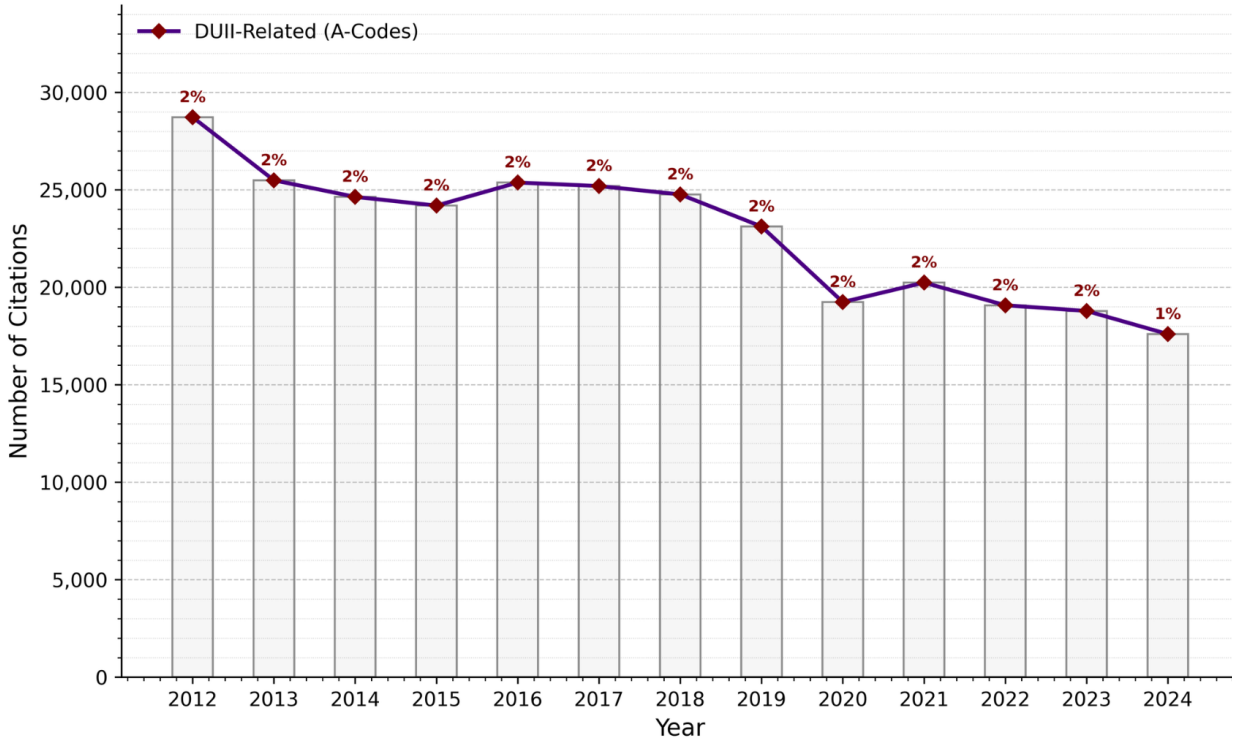
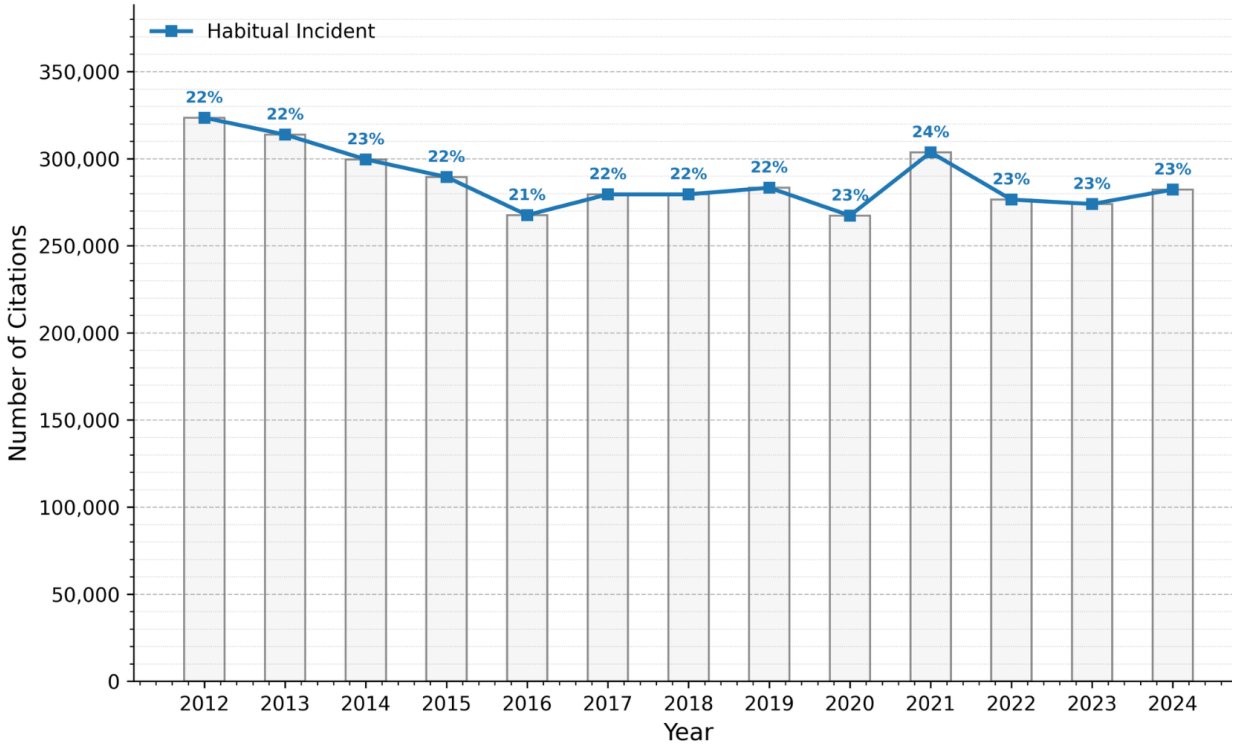


Figure 6.2 Top Ten Citation Types by Year

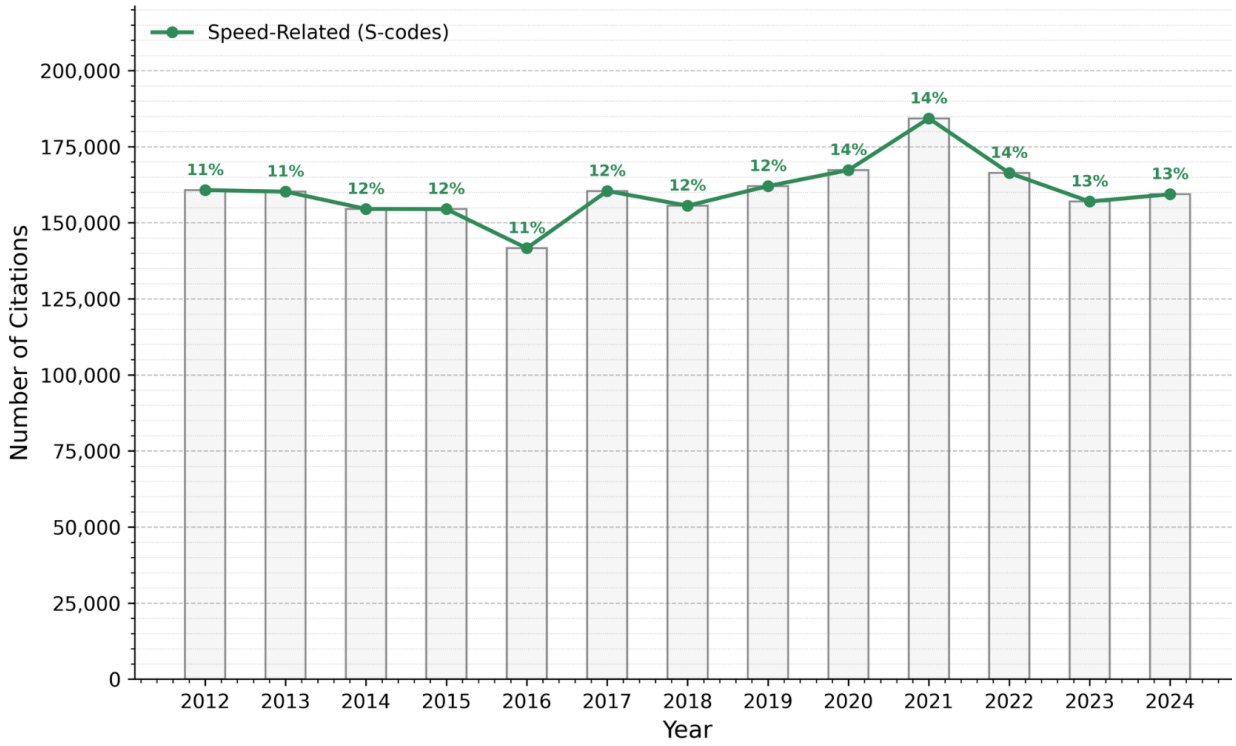
Citation types were then sorted into three major categories: DUII, Habitual Incidents, and Speeding. Using data between 2012 and 2024, the annual number of violations for each category was plotted. **Error! Reference source not found.**Figure 6.3 shows that DUII-related violations make up a much smaller proportion of violations compared to speed-related and habitual incident-related violations. In addition to be a small percentage of overall citations, DUII-related citations have decreased between 2012 and 2024. This may be attributed to post-Covid-19 law enforcement staffing shortages and the increased difficulty of DUII convictions due to changes in Oregon law. The number of habitual citations (which encompasses both minor and major incidents) has remained a steady percentage during the same time period. In contrast, the percentage of speeding citations has increased between 2012 and 2024, with a spike around 2021.


DUII-Related (A-Codes)
Number of Citations
Year
0
5,000
10,000
15,000
20,000
25,000
30,000
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2%
2%
2%
2%
2%
2%
2%
2%
2%
2%
2%
2%
1%

(A)


Habitual Incident
Number of Citations
Year
0
50,000
100,000
150,000
200,000
250,000
300,000
350,000
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
22%
22%
23%
22%
21%
22%
22%
22%
23%
24%
23%
23%
23%

(B)



(C)

Figure 6.3 Types of Citations by Year

There is an interesting trend where the speeding and habitual incident curves both follow the same trends, with a dip around 2016 and a spike around 2021.

6.1.3 Demographic Analysis

The demographics of drivers who commit violations provide insight into what groups may need additional targeted education. The two demographic variables available for analysis are age and sex.

6.1.3.1 Age-based Trends

Figure 6.4 shows the distribution of citations by both age and sex. The first major observation is that the data is skewed towards younger drivers. The number of citations peaks around 22-24 years of age before slowly tapering off at older ages. The difference between male and female drivers is the most pronounced during younger years. As drivers get older, the difference in the number of citations between the sexes slowly decreases.

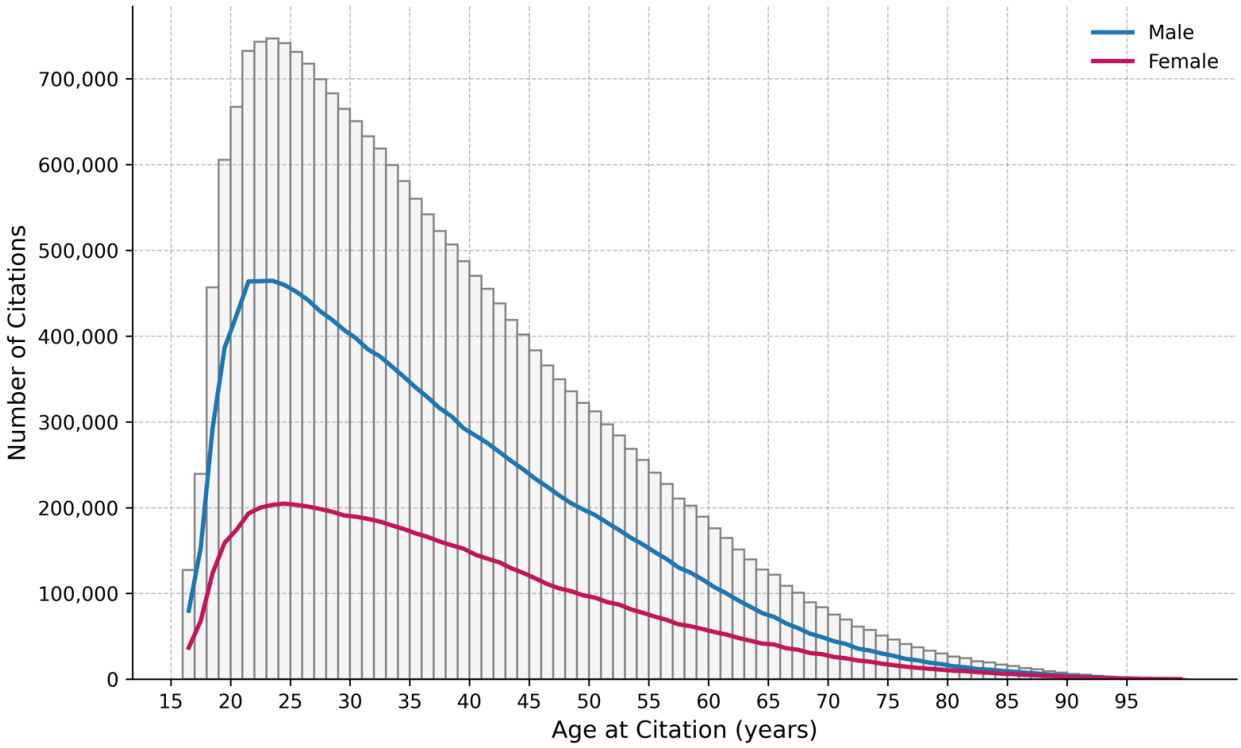


Figure 6.4 Citations by Sex and Age

6.1.3.2 Sex-based Trends

While younger male drivers do accrue more citations, the percentage of female drivers with citations has doubled since 1995. Figure 6.5 shows the gap in the percentage of yearly citations by sex, which slowly narrowed between the years of 1995 and 2012. Post-2012, the gap between the sexes stayed constant at a difference of around 25%. Interestingly, in 1995, the total share of male and female violations was only 73% of all violations. Since 2012, the number has increased to around 90% of violations with properly documented sex fields.

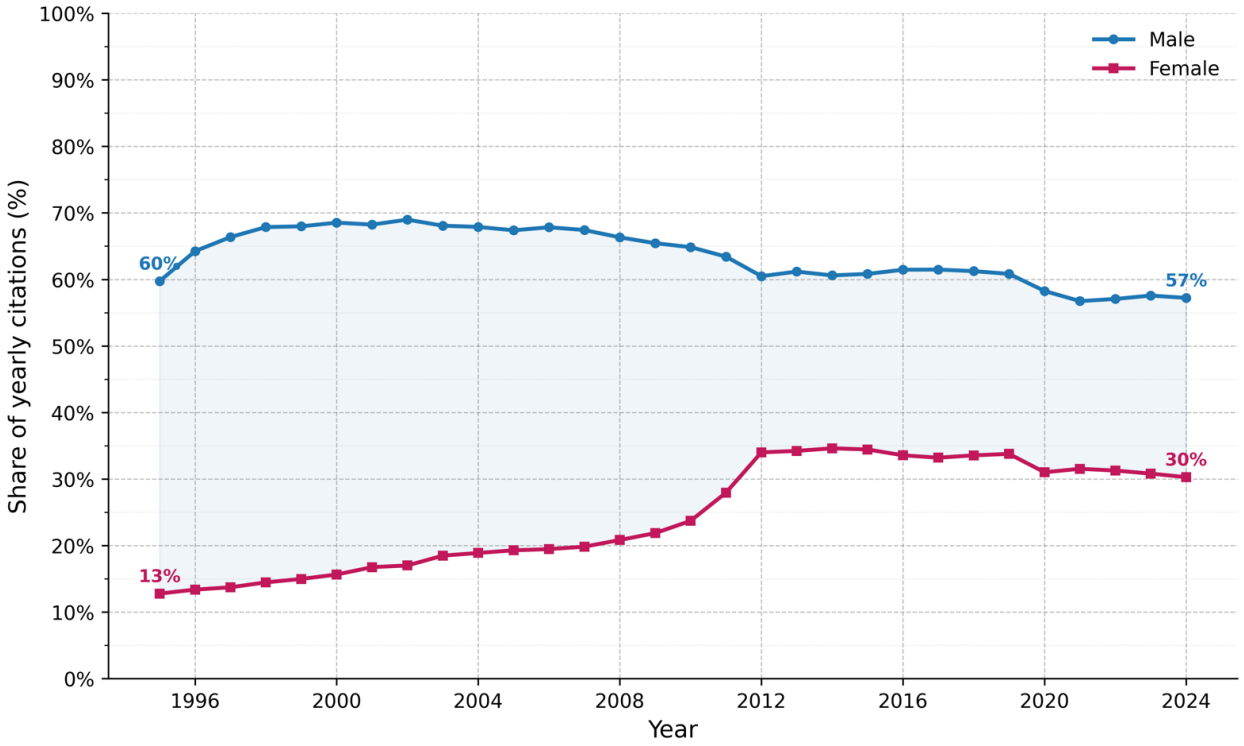


Figure 6.5 Share of Yearly Citations by Sex

Experience-based Trends

Using data about the date of birth and the date that a license was first issued, the experience of each driver in Oregon can be calculated in years. After determining the Oregon driving experience of each driver, drivers were sorted into various categories based on previous experiences with law enforcement while driving. Those four categories were defined as:

- 0 (No Citations, No Crashes),
- 1 (Citations, No Crashes),
- 2 (No Citations, Crashes), and
- 3 (Citations and Crashes).

Figure 6.6 shows the correlation between driver experience and experience with law enforcement while driving. The observation that younger drivers are overrepresented in the population with only citations is consistent with this observation. This trend occurs in both sexes, with no large differences between the sexes.

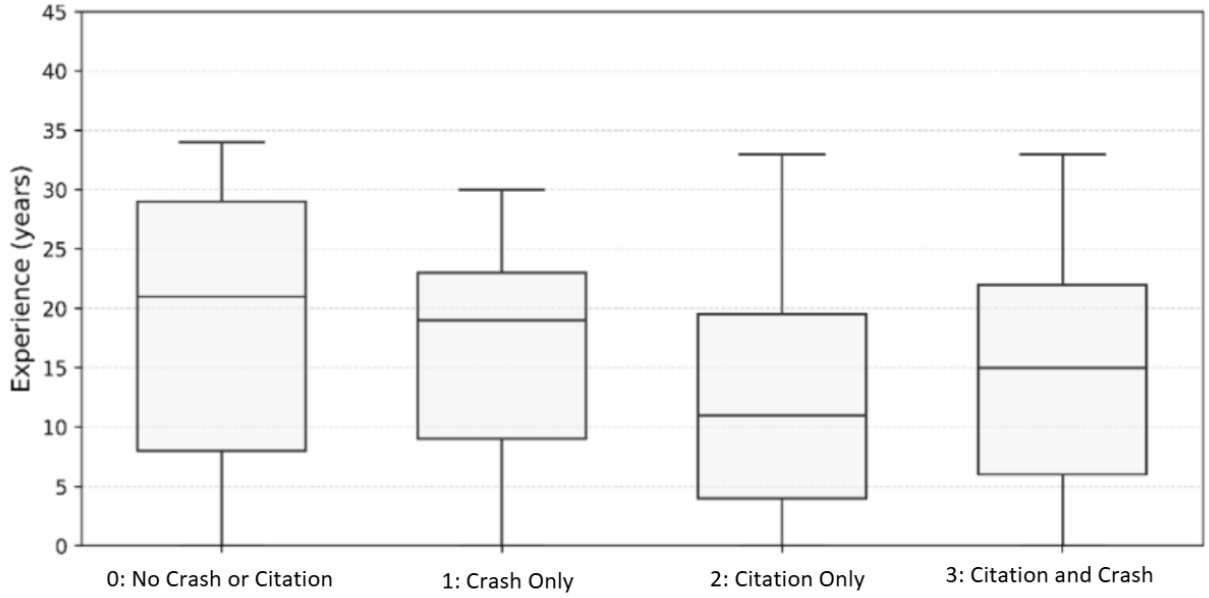


Figure 6.6 Experience (years) by Risk Category

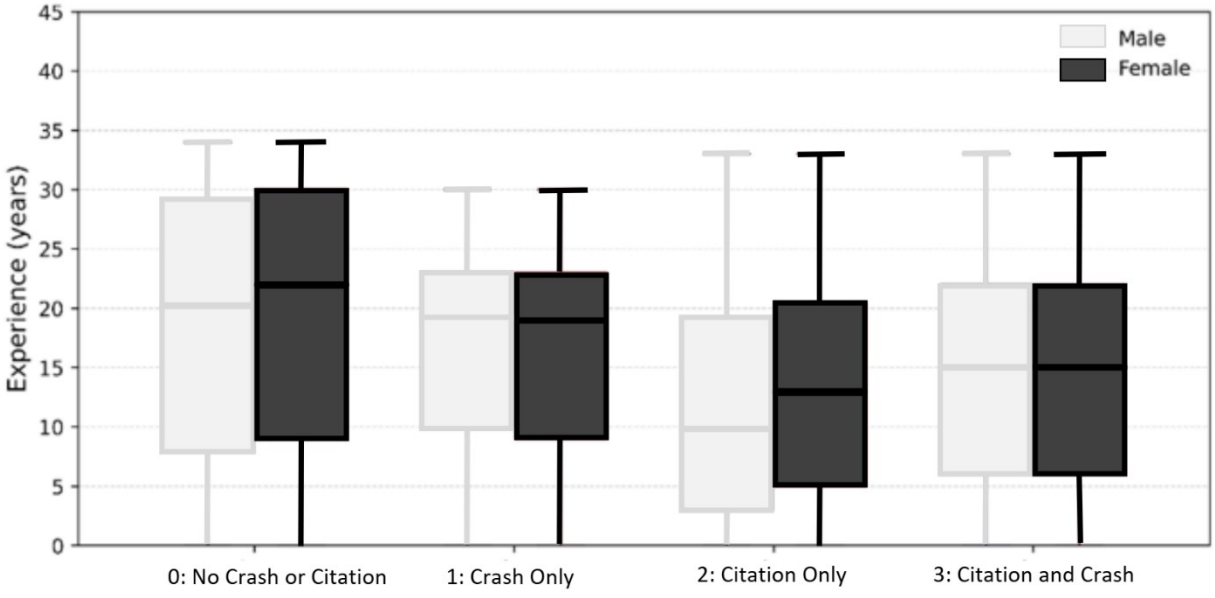


Figure 6.7 Experience (years) by Risk Category and Sex

6.2 ANALYSIS OF THE IMPACT OF CITATIONS ON CRASH RATES

To determine the impact of citations on crash rates, several graphs were produced to illustrate the relationship.

Figure 6.8 shows the visualization of the time difference in months between the issuance of the first citation and the first recorded crash. There is a large spike in the number of drivers getting

into a crash soon after their first citation. The rate of drivers getting into crashes after their first citation leveled out for two years before steadily decreasing. This shows that the two years after getting a citation are when most drivers, who have received a citation, are at their highest level of risk.

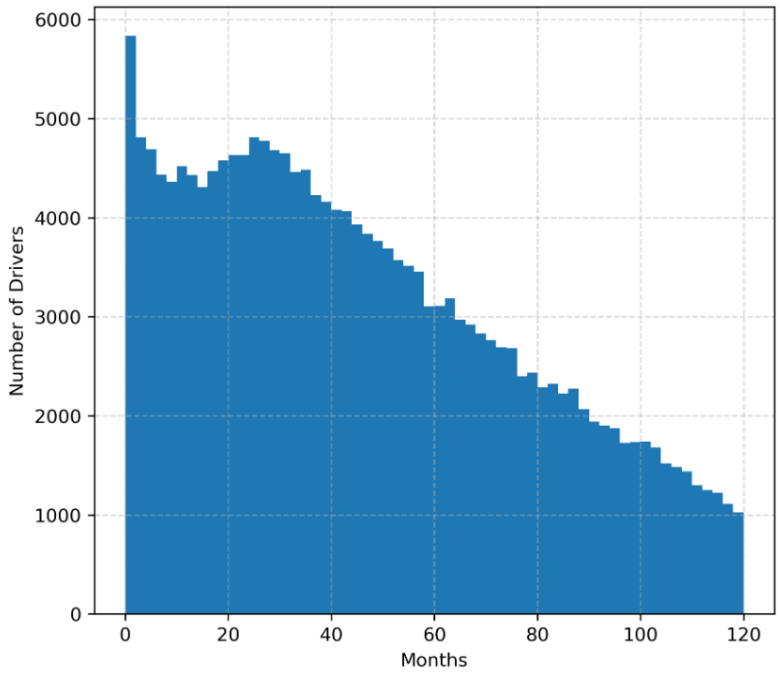


Figure 6.8 Months Between First Citation and First Crash (1994-2024)

6.2.1 Statistical Analysis

To better understand the data, a series of statistical modeling techniques was applied. Three binomial Generalized Linear Models (GLMs) with a logit link function were used to estimate the likelihood of a driver being involved in a speed-related crash, a DUII-related crash, or any crash based on prior citation history, age, and gender. Additionally, a multinomial logistic regression model was used to examine how these same factors influenced crash severity levels, comparing outcomes across PDO (property damage only), C (possible injury), B (non-incapacitating injury), and KA (fatal or incapacitating injury) categories. The analysis window was set to 1995–2024 for the verdict data and 2014-2021 for the crash data because statewide citations reporting stabilized after 1994, as indicated by Figure 6.1, which shows consistent citation volumes over time within this period. Note that, the regression coefficients in the tables are estimated on the log-odds scale and are not directly intuitive. For interpretation, these coefficients are transformed into odds ratios (ORs) by exponentiating them. An OR greater than 1 indicates higher odds of the outcome compared with the reference group, and an OR less than 1 indicates lower odds, holding the other variables constant.

6.2.1.1 Speed-related crashes

The speed-related crash model was estimated using GLM. The dependent variable indicated whether a driver had at least one crash coded as speed-involved. Predictors included indicator variables for having any prior speeding citation and any prior DUII citation (both measured strictly before the first crash), a sex indicator (male versus female), and driver age group, with the younger-than-21-years age group set as the baseline. Odds ratios were computed from the regression coefficients, and 95 percent confidence intervals were used to assess statistical significance.

Results of the model are presented in Table 6.2. Both prior speeding and prior DUII history were positively associated with the likelihood of being involved in a speed-related crash and were statistically significant ($P < 0.01$). Drivers with a prior speeding citation had approximately 20% higher odds of being involved in a speed-related crash compared with those without any prior documented speeding history. Male drivers had higher odds of speed-related crashes relative to female drivers. Age effects were also statistically significant; when compared with drivers younger than 21 years, older age groups showed higher odds, with the largest increases among drivers aged 25-34 years and 45-54 years. This effect began to decline after the age of 55, reflecting a reduced likelihood of speed-related crash involvement among older drivers.

Table 6.2. Logistic regression results for speed-related crash involvement

Variables	Coeff	Std Error	P-value	95% CI	Odds ratio
Intercept	-7.47	0.07	<0.001	(-7.62, -7.33)	0.01
Prior speeding citation	0.18	0.01	<0.001	(0.16, 0.19)	1.19
Prior DUII citation	0.42	0.01	<0.001	(0.40, 0.45)	1.53
Male (Baseline: Female)	0.43	0.01	<0.001	(0.42, 0.45)	1.54
Age less than 21 yrs	baseline	baseline	baseline	baseline	baseline
Age 21–24 yrs	2.80	0.08	<0.001	(2.65, 2.95)	16.48
Age 25–34 yrs	3.32	0.07	<0.001	(3.17, 3.46)	27.53
Age 35–44 yrs	2.73	0.07	<0.001	(2.58, 2.87)	15.26
Age 45–54 yrs	2.43	0.07	<0.001	(2.28, 2.57)	11.3
Age 55–64 yrs	2.43	0.07	<0.001	(2.28, 2.57)	11.31
Age 65–74 yrs	2.26	0.07	<0.001	(2.11, 2.40)	9.56
Age 75+ yrs	1.70	0.07	<0.001	(1.55, 1.85)	5.46

6.2.1.2 DUII-related crashes

A similar model was used to estimate the likelihood of DUII-related crashes. However, the dependent variable in this case indicated whether a driver had at least one crash coded as DUII-involved. Predictors were consistent with the previous model, including prior DUII and speeding citation history, sex, and age group, with drivers younger than 21 years serving as the baseline category. Odds ratios were derived from the regression

coefficients, and 95 percent confidence intervals were used to evaluate statistical significance.

The model output for DUII-related crashes is presented in Table 6.3. Prior DUII history was associated with about 270% higher odds of being involved in a DUII-related crash, which was statistically significant ($P < 0.001$). Additionally, prior speeding citations increased the odds by about 27% ($P < 0.001$). Male drivers exhibited about 67% higher odds of DUII-related crashes when compared to female drivers ($P < 0.001$). When compared to drivers younger than 21 years, the odds of being involved in a DUII-related crash increased substantially with age. The largest increase was observed among drivers aged 25-34 years, who had about 95 times higher odds compared with the youngest group. The large difference for those under 21 years is likely due to legal drinking restrictions that prohibit alcohol consumption, which naturally limit their exposure to DUII-related crashes. Unlike DUII, the odds from the speed-involved crashes model were higher across age groups, but the increases were much less pronounced when compared to those under 21 years old. This suggests that younger drivers tend to engage in speeding but are less likely to be involved in alcohol-related crashes due to the age limit.

Table 6.3. Logistic regression results for DUII-related crash involvement

Variables	Coeff	Std Error	P-value	95% CI	Odds ratio
Intercept	-9.87	0.24	<0.001	(-10.33, -9.40)	0.01
Prior speeding citation	0.24	0.01	<0.001	(0.21, 0.27)	1.27
Prior DUII citation	1.31	0.02	<0.001	(1.28, 1.34)	3.70
Male (Baseline: Female)	0.51	0.01	<0.001	(0.49, 0.54)	1.67
Age less than 21 yrs	Baseline	Baseline	Baseline	Baseline	Baseline
Age 21–24 yrs	3.33	0.24	<0.001	(2.86, 3.80)	28.06
Age 25–34 yrs	4.56	0.24	<0.001	(4.09, 5.02)	95.11
Age 35–44 yrs	4.23	0.24	<0.001	(3.77, 4.70)	68.99
Age 45–54 yrs	3.92	0.24	<0.001	(3.46, 4.38)	50.43
Age 55–64 yrs	3.90	0.24	<0.001	(3.44, 4.36)	49.37
Age 65–74 yrs	3.66	0.24	<0.001	(3.20, 4.12)	38.82
Age 75+ yrs	2.84	0.24	<0.001	(2.37, 3.30)	17.10

6.2.1.3 All type-related crashes

The third and final binomial GLM with a logit link estimated the likelihood of involvement in any crash type. The main independent variable was the lifetime number of citations a driver received, grouped as zero citations (baseline), one citation, two to five citations, six to ten citations, eleven to fifteen citations, and more than fifteen citations. Age group was also included, with drivers younger than 21 years as the baseline, like the prior models (Speed and DUII).

Results of the model are reported in Table 6.4. Crash odds increased steadily with the number of citations. When compared with drivers with zero citations, odds were about

47% higher with one citation, about 303% higher with two to five, about 387% higher with six to ten, about 603% higher with eleven to fifteen, and about 849% higher with more than fifteen citations (all $P < 0.001$). Age effects were also large; relative to drivers younger than 21 years, odds were higher across all older groups, with the largest increases observed for ages 25-34 and 65-74 years. Overall, the pattern indicates a steady increase in crash involvement as citation history accumulates, even after accounting for age.

Table 6.4. Logistic regression results for all type-related crash involvement

Variables	Coeff	Std Error	P-value	95% CI	Odds ratio
Intercept	-5.86	0.03	<0.001	(-5.92, -5.80)	0.01
Zero Citation	Baseline	Baseline	Baseline	Baseline	V
Cit1 (vs Cit0)	0.39	0.01	<0.001	(0.37, 0.40)	1.47
Cit2–5 (vs Cit0)	1.39	0.01	<0.001	(1.39, 1.40)	4.03
Cit6–10 (vs Cit0)	1.58	0.01	<0.001	(1.57, 1.59)	4.87
Cit11–15 (vs Cit0)	1.95	0.01	<0.001	(1.94, 1.96)	7.03
Cit15+ (vs Cit0)	2.25	0.01	<0.001	(2.24, 2.26)	9.49
Age less than 21 yrs	Baseline	Baseline	Baseline	Baseline	Baseline
Age 21–24 yrs	2.50	0.03	<0.001	(2.44, 2.57)	12.41
Age 25–34 yrs	2.95	0.03	<0.001	(2.89, 3.02)	19.16
Age 35–44 yrs	2.58	0.03	<0.001	(2.50, 2.62)	13.21
Age 45–54 yrs	2.46	0.03	<0.001	(2.39, 2.52)	11.67
Age 55–64 yrs	2.56	0.03	<0.001	(2.50, 2.62)	12.93
Age 65–74 yrs	2.67	0.03	<0.001	(2.60, 2.73)	14.39
Age 75+ yrs	2.57	0.03	<0.001	(2.50, 2.63)	13.02

To isolate the effect of citations from other independent variables, the estimated probability of crash involvement by citation category was plotted in Figure 6.9. Crash Probabilities increased from about 3.6% with zero citations, to 5.4% with one, 13.5% with two to five, 16.0% with six to ten, 21.6% with eleven to fifteen, and 27.2% with more than fifteen citations. The smooth rise in the line overlay aligns well with the regression findings. To that end, because the number and pattern of citations strongly correlated with crash involvement, the model can be used to flag higher-risk (aggressive) drivers earlier. Escalating citation counts, moving from one to two-five and beyond, provide a clear warning sign that a driver's crash risk is rising, and could support the need for targeted education, monitoring, or intervention before a crash happens.

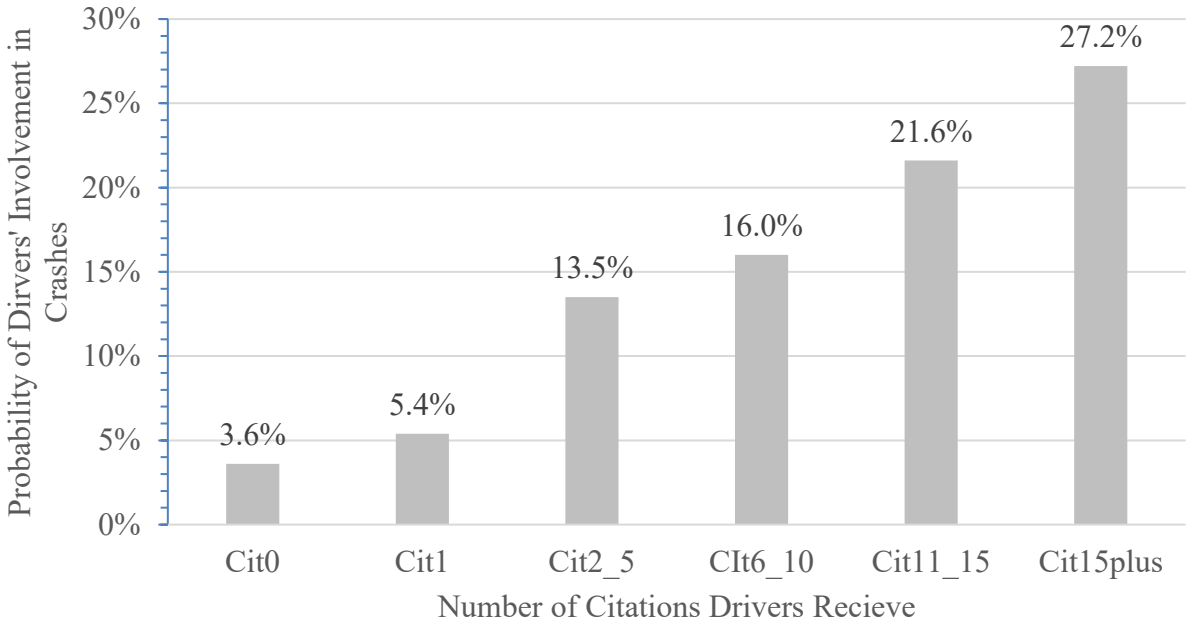


Figure 6.9 Probability of Crash Involvement by Citation History

6.2.1.4 Crash Severity Model

A multinomial logistic (ML) regression was used to compare crash severity categories relative to PDO. Severity levels B (non-incapacitating injury), C (possible injury), and KA (fatal or incapacitating injury) were modeled simultaneously. Severity levels 4 (A) and 5 (K) were collapsed into a single KA category to address class imbalance at the most severe end. Predictors included prior citation indicators for the top-10 violation codes measured strictly before the first crash, sex (male versus female), and age group, with drivers younger than 21 years as the baseline.

In terms of the results, the odds ratios from the ML model were graphically presented in Figure 6.10. It has three contrasts that are overlaid for each predictor: C in light gray, B in light blue, and KA in dark red. Points mark the odds ratios, and horizontal bars show the 95 percent confidence intervals, with a vertical reference line at OR = 1. Predictors are arranged on the y-axis in logical blocks (age groups, then violation codes, then male). In this figure, the right side indicates higher odds than PDO, and the left side indicates lower odds.

Results showed that the age profile differed across severity levels. Compared with PDO, the odds for C increased with age (e.g., 25–34: OR = 1.82; 45–54: OR = 2.26), while B decreased in older groups (e.g., 25–34: OR = 0.67; 45–54: OR = 0.56). For KA, odds increased again at older ages (e.g., 65–74: OR = 1.35; 75+: OR =1.47), showing that the most serious outcomes are concentrated among older drivers. Prior administrative and DUII-linked indicators were the clearest severity markers: implied consent (KA OR = 1.31), SR-22 (KA OR = 1.23), no insurance (D36) (KA OR ≈ 1.26), and failed to answer citation (D56) (KA OR = 1.23). Routine moving violations showed weak or negative

associations with higher severity; for example, signal violation (M14) was below one for B and KA. The male coefficient was below one for C and B (C OR = 0.58, B OR = 0.73) and slightly above one for KA (OR = 1.14), indicating lower odds for minor categories but a small positive association at the most severe level.

To that end, these trends suggest a two-part finding. As age increases, crashes are more often recorded as “possible injury” rather than “minor injury”, and the odds of KA increase again in older groups. Independent of age, prior DUII/administrative non-compliance signals are most predictive of serious outcomes, whereas common moving violations carry little information about injury severity once a crash occurs.

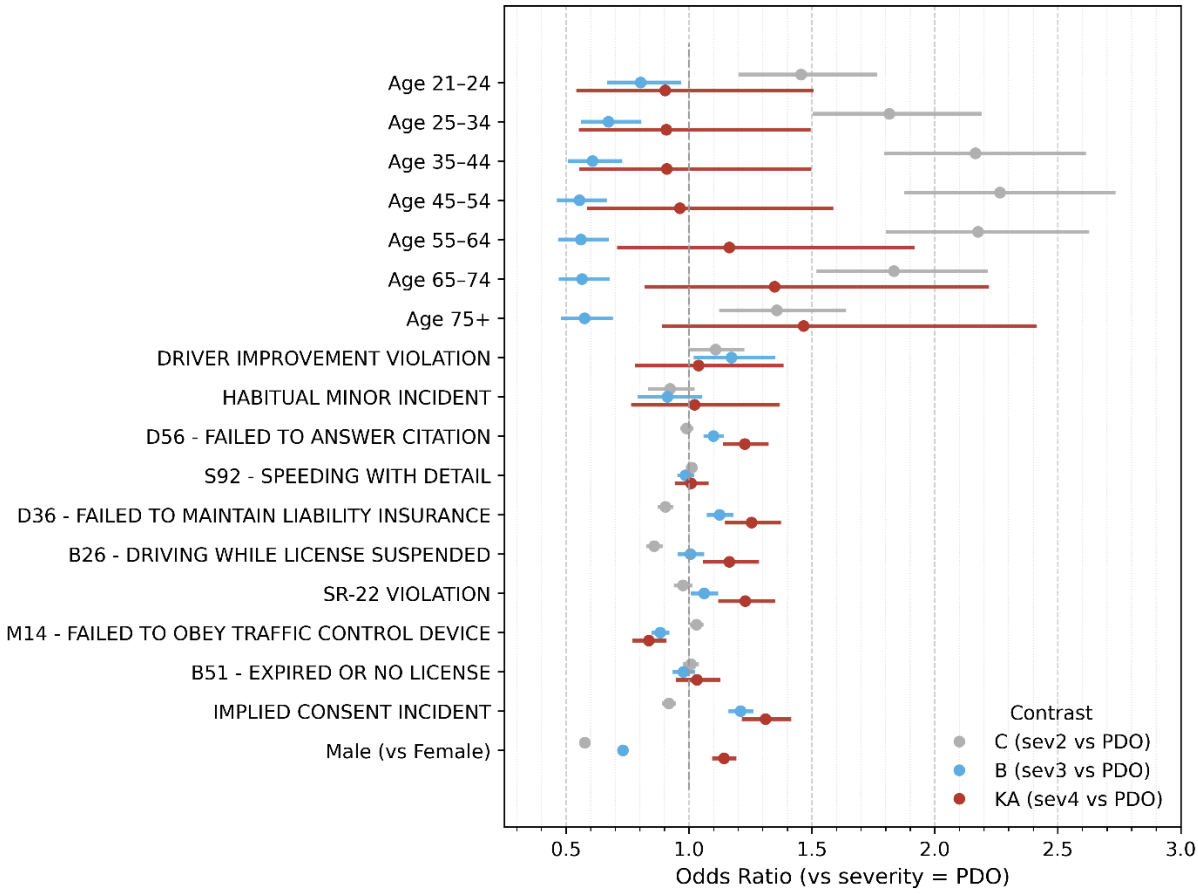


Figure 6.10 Probability of Increased Crash Severity by Violation Code and Demographics

7.0 RECOMMENDATIONS FOR OPTIMIZATION

This dataset is novel in that it is the first dataset in the state of Oregon that links different aspects of traffic safety data together through several related data sources. During the course of the process, several inefficiencies were uncovered that could be resolved to provide a more streamlined linkage process.

7.1 WORKFLOW OPTIMIZATION

The current database linking process requires extensive data cleaning and validation steps to ensure that all data inconsistencies are identified and removed. One source of computational efficiency could come from the use of blocking variables, such as county. To leverage this technique fully, having the addresses of people would be useful. While QA/QC is an important part of any process, there are several steps that could be taken to remove burden on the process.

7.1.1 Data Field Standardization

Deterministic linkage relies on standard, uniform fields to be able to make consistent and accurate links. Without standardization of the inputs, the likelihood of the code throwing an error or generating unintelligible results increases. Throughout the creation of the linked datasets, multiple data field inconsistences were uncovered that could streamline the deterministic model.

7.1.1.1 Inconsistencies Within the Same Dataset

There were a few inconsistencies that occurred within the same dataset that generated errors. This prevented new datasets from being linked to earlier versions without additional data cleaning and code writing.

Changing Field Names Across Dataset Versions – Occasionally when new data was requested from the agency, the new spreadsheet would use a fieldname with a slight variation. While these were likely added to clarify the field, they created an attribute that could no longer be directly linked to existing dataset without additional coding. An example of this was changing the title of a data field from “Verdict” would become “Verdict Date” between versions.

Changing File Names Between Data Pulls – There were occasions where a new data pull would be sent as a file that had a different naming convention that previous files. The code is configured to merge files with very specific naming, so these slight variations resulted in the code viewing the file as missing.

To resolve these issues, the Oregon DMV could consider standardizing all fields across their datasets and documenting why these fields should remain standardized. If an improvement need to be made, they should be well documented and implemented across all datasets to reduce the number of times that new code needs to be developed.

7.1.1.2 Inconsistencies Within Datasets from the Same Agency

Of the four datasets, three were provided by the Oregon DMV. While most of the information was standardized between the three datasets, there were some formatting discrepancies that required additional lines of code to address.

Name Formatting – There were several different ways in which naming was inconsistent across the DMV datasets. Some datasets had additional spaces after names that other datasets omitted, which requires data cleaning to standardize. Additionally, the inclusion of middle name initials was sometimes excluded from datasets which make deterministic linkage more difficult by removing an identifying field. This is especially important for entries with common names. An example of this was the last name Smith appearing as "Smith " in the Driver Data and "Smith" in the Accident Data.

DOB Formatting – The code was written to handle dates in the format MM/DD/YYYY. Before running the code, all datasets had to be reformatted to ensure that the code could recognize the dates as valid. An example of this was "01/01/2020 00:00:00AM" where the time would need to be removed before the merging could be conducted.

Identifier Formatting – ODLs were treated as different data types between datasets. In order for the code to merge the datasets, the ODLs had to be converted to the same data type. An example of this was that Driver Data formatted ODLs are numeric fields, while Accident Data treated ODLs as strings.

To resolve these inconsistencies, a standard format should be implemented across all DMV datasets. This makes locating information faster and more efficient, both manually and via software.

7.1.1.3 Inconsistencies Across Archives

Because data was collected from both the DMV and the ODOT Crash Data System (CDS), there were issues with different serial number formats. This required the creation of a new serial number to combine the datasets, which required additional time and computational power.

Serial Number Formatting – Each unit has a different way of organizing the serial number in the Accident Data and Crash Data. The Crash Data creates a new serial number based on the serial number found in the Accident Data. It also uses leading zeros and specialized numerical prefixes to indicate duplicates and reassigned counties. Because of these different methodologies, the merge requires custom logic to handle the multitude of combinations that the system could generate. An example of this was that The Accident Data would use the code "01-1234" to represent an accident and the Crash Data would use "01234" as the new serial number. Another common occurrence was that if the Accident Data had a duplicate entry using the code "01-12345," the Crash Data would use the serial number "92345."

This ID system is non-intuitive, and it makes records harder to link because the serial number may change depending on the circumstances. The units should consider

collaborating to create a system that relies on a standard serial number format while finding ways to communicate all necessary information. Examples of this could be creating a new field to identify entries that are duplicates or removing the leading zeros from the ODOT CDS formatting.

7.2 DOCUMENTATION OF CURRENT DATA PRACTICES

The current documentation for the four datasets is not easy to find and is often distributed by agency employees who have experience working with the datasets. This method is sufficient for day-to-day operations, but it creates issues when attempting to document the linkage process of the combined dataset due to the sheer number of variables (which may go by different names depending on the unit) and unit codes. This report has provided detailed documentation on the linkage process, but there are areas for improvement that would make the documentation even more comprehensive.

7.2.1 Creation of a Comprehensive Data Dictionary

The consolidation of all attributes into one data dictionary would be useful for all future endeavors that use the combined dataset. By providing clarity on the definition of all variables, administrative burden could be reduced by shortening the amount of time spent searching for explanations. If the dataset is non-numerical in nature and utilizes abbreviations, the dictionary should also provide explanations for all possible entries.

7.2.1.1 Cataloging Codes for Exceptions and Special Cases

During the creation of the data dictionary, a separate section should be dedicated to special cases where the data entry deviates from the established rules. For example, ODOT CDS assigns the serial number “99999” to fatal collisions with no driver record, like in a hit and run. These special exceptions are not self-explanatory, so explicitly spelling them out could save time and create a form of knowledge retention for future research and exploration.

7.2.1.2 Cataloging Internal Data Cleaning Practices

As seen in previous examples, agencies have implemented data cleaning practices, such as altering the serial number of duplicates. These internal practices could be catalogued in the data dictionary to understand what has already been done to address data cleaning and how future users may filter out previously completed work to avoid redundant steps.

7.2.2 Documentation of Data relating to Driver Improvement Programs

One of the many reasons for creating a combined state-wide dataset was the assessment of the efficacy of the HTO program and other related behavioral programs. However, data on these programs is sparse within the data from both the DMV and CDS databases. To more accurately assess the success of these programs, there are several data fields that would be useful:

- Program start date

- Program end date (or duration of program)
- Program completed (yes or no)
- Contractor/agency who delivered the program (if appropriate)

The list above is not exhaustive, and any other variables that could assist researchers in evaluating intervention programs would increase the utility of the database as a tool for DMV program evaluation.

7.3 SUMMARY OF RECOMMENDATIONS

Recommendations for optimizing the workflow of future dataset merging fall into three broad categories: data standardization, geospatial data collection, and documentation of existing procedures. The improvements itemized in each of these categories would reduce confusion for future users and allow for streamlined data linkage that requires less complicated logic to process. While it is understood that large administrative tasks are difficult, especially those that require collaboration across units, these improvements will save time in the long run as new data is integrated.

8.0 CONCLUSION

This report summarizes current best practices on crash data linkage in the state of Oregon, and how a linked crash database can be applied to the analysis of the efficacy of DMV programs such as the Habitual Traffic Offender program. Using driver, verdict, crash, and accident data from the Oregon DMV and ODOT CDS, a combined database was created using deterministic linkage methods that could link demographic information to crash data. The database was used to generate visualizations of citation and collision history, showing that trends in sex, age, and driver experience can be observed across several years of data. Three GLM were used to evaluate speed-related crashes, DUII-related crashes, and all-type related crashes. Additionally, a multinomial logit model evaluated the effect of the top ten citation types in the verdict data on the level of crash severity. Results revealed that drivers who received speeding or DUII citations before the occurrence of a crash were very likely to be involved in speeding or DUII related crashes. Considering the odds ratios for speeding-related and DUII-related crashes by age, it was found that younger drivers tend to engage in speeding but are less likely to be involved in alcohol-related crashes due to the age limit. The report also generated several recommendations for Oregon transportation agencies on how to best align their data management practices to streamline future database integration and answer detailed questions regarding the use of driver improvement practices.

8.1 LIMITATIONS

The primary limitation of the current database is a lack of clear data on the driver improvement programs in the state of Oregon. Without data on when participants entered and graduated from these programs, the closest approximation available in the current data is when a driver exceeds the stated thresholds for citations or collisions within a specified timeframe. There is also no data on which program a participant entered, meaning that this information would have to be extrapolated from the types of citations garnered. The indirect nature of these observations introduces significant uncertainty into any conclusion drawn from analysis and prevents recommendations specific to any program.

8.2 RECOMMENDATIONS

The database in its current form produced several insights that could be used to enhance current driver behavioral interventions, such as the Habitual Traffic Offender program. Statistical analysis revealed that drivers are most likely to be involved in a crash in the two years immediately following the issuance of their first citation, This window could prove useful when evaluating license suspension terms, which generally fall in the six-month to one year range for programs such as the Driver Improvement Program. Another interesting observation is that the odds of being involved in a collision increase 47% after drivers receive their first citation and 303% when they receive between two to five citations. Most intervention thresholds are closer to three citations, so this information could be used to justify earlier interventions. Lastly, citations relating to violations such as administrative non-compliance and DUII has a much stronger correlation with increased crash rates, so focusing on flagging those types of behaviors as more serious may improve the efficacy of the system. While these findings would still require policy

review and further analysis, they represent potential routes for the Oregon DMV to explore in the future.

8.3 FUTURE RESEARCH OPPORTUNITIES

The utility of the integrated dataset could be improved in the future if additional data sources, such as driver improvement program rosters, could be linked to the dataset using deterministic linkage. This would allow for the analysis of programs on a more granular level, resulting in actionable recommendations for agencies on how best to improve program outcomes. In addition to improving the analysis of the driver improvement programs, there are also opportunities that may arise if georeferenced data is used to validate and expand the crash data. While state-wide georeferenced crash databases do exist, these databases do not have the ability to study the citation history surrounding the crash. This could be useful in analyzing questions such as whether people who violate license suspensions are repeatedly doing so in the same geographic areas (which may indicate routine trip patterns), which may help inform policy aimed at increasing compliance. By linking these additional datasets to the database in future projects, the database may be able to generate better conclusions regarding risky driving behaviors and how agency policy can be used to reduce risky driving outcomes.

Moreover, future analyses could also start with a simple figure showing crash counts by severity, with separate panels for PDO, injury, serious injury, and fatal crashes, so the relative scale of each outcome is clear. Additional models could then focus on specific outcomes: (1) DUII-related serious and fatal crashes, (2) any serious or fatal crash, and (3) crashes on higher-speed roads (for example, posted 35 or 40 mph and above), while carefully handling missing posted speed information.

Model performance could be extended by adding basic predictive metrics, such as pseudo-R2, if computationally practical for the full linked dataset. Age effects could also be refined by testing narrower groups among young drivers, such as 16-18 and 19-21, to see whether patterns differ within the under-21 population. Finally, repeating key analyses by severity (PDO, injury, serious injury, fatal) would help identify whether the associations with demographics, experience, and citation history change as severity increases.

9.0 REFERENCES

Auguste, M., & Pawelzik, J. (2024). Linking crash and breathalyzer data in Connecticut. *Traffic Injury Prevention*, https://doi.org/10.1080/15389588.2024.2314589

Bamney, A., Pantangi, S. S., Jashami, H., & Savolainen, P. (2022). How do the type and duration of distraction affect speed selection and crash risk? An evaluation using naturalistic driving data. *Accident Analysis and Prevention, 178*https://doi.org/10.1016/j.aap.2022.106854

Brief penalties for revoked driver's license: Habitual traffic offenders (HTO). National Conference of State Legislatures (NCSL). (2022, March). https://www.ncsl.org/transportation/penalties-for-revoked-drivers-license-habitual-traffic-offenders-hto

Cook, L. J., Thomas, A., Olson, C., Funai, T., & Simmons, T. (2015, July). Crash Outcome Data Evaluation System (CODES): An examination of methodologies and multi-state traffic safety applications. (Report No. DOT HS 812 179). Washington, DC: U.S. Department of Transportation, National Highway Traffic Safety Administration.

Crash outcome data evaluation system (CODES). NHTSA. (n.d.). https://www.nhtsa.gov/crash-data-systems/crash-outcome-data-evaluation-system-codes

Dezman, Z., Andrade, L. d., Vissoci, J. R., El-Gabri, D., Johnson, A., Hirshon, J. M., & Staton, C. A. (2016). Hotspots and causes of motor vehicle crashes in Baltimore, Maryland: A geospatial analysis of five years of police crash and census data. *Injury*, http://dx.doi.org/10.1016/j.injury.2016.09.002

Doidge, J. C., & Harron, K. (2018). Demystifying probabilistic linkage: Common myths and misconceptions. *International Journal of Population Data Science, 3* https://doi.org/10.23889%2Fijpds.v3i1.410

FindLaw. (2023, January). California Code, Vehicle Code - VEH § 14601.3 | findlaw. https://codes.findlaw.com/ca/vehicle-code/veh-sect-14601-3/

Glitsch, E., & Knuth, D. (2016). Key aspects of successful rehabilitation after repeated or serious driving offenses. *Traffic Injury Prevention, 17*336-345. http://dx.doi.org/10.1080/15389588.2015.1082176

Jones, S. (1986). The effectiveness of habitual traffic offender license revocation in Oregon. Oregon Motor Vehicles Division (DMV). Salem, OR.

Karimi, S., Hosseinzadeh, A., Kluger, R., Wang, T., Souleyrette, R., & Harding, E. (2024). A systematic review and meta-analysis of data linkage between motor vehicle crash and hospital-based datasets. *Accident Analysis and Prevention, 197* https://doi.org/10.1016/j.aap.2024.107461

Kweon, Y.-J. (2011). Chapter 8: Crash Data Sets and Analysis. In *Handbook of Traffic Psychology*. essay, Elsevier Inc.

Lee, J., Park, B.-J., & Lee, C. (2018). Deterrent effects of demerit points and license sanctions on drivers' traffic law violations using a proportional hazard model. *Accident Analysis & Prevention*, *113*, 279–286. https://doi.org/10.1016/j.aap.2018.01.028

Mayhew, D. R., Simpson, H. M., & Pak, A. (2003). Changes in collision rates among novice drivers during the first months of driving. Accident analysis & prevention, 35(5), 683-691.
MMUCC guideline model minimum uniform crash criteria ... (2024, January). https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813525
Montana Code Annotated 2023 (MCA 2023). 61-11-212. Penalties, MCA. (n.d.). https://leg.mt.gov/bills/mca/title_0610/chapter_0110/part_0020/section_0120/0610-0110-0020-0120.html
Oregon Department of Transportation (ODOT). (2022). Oregon suspension/revocation/cancellation guide. https://www.oregon.gov/ODOT/Forms/DMV/7484.pdf
Oregon Department of Transportation (ODOT). (2021). Oregon Transportation Safety Action Plan. https://www.oregon.gov/odot/Safety/Documents/2021_Oregon_TSAP.pdf
Oregon DMV. (n.d.). *Suspensions, revocations and cancellations*. Oregon Department of Transportation: Suspensions, Revocations and Cancellations: Oregon Driver & Motor Vehicle Services: State of Oregon. https://www.oregon.gov/ODOT/DMV/pages/driverid/suspreasons.aspx
Owens, J. M. (2023). Strategies to Improve State Traffic Citation and Adjudication Outcomes (No. BTSCRP Project BTS-04).
Parrish, K. E., & Masten, S. V. (2014). The problem of suspended and revoked drivers who avoid detection at checkpoints. *Traffic Injury Prevention*, *16*(2), 97–103. https://doi.org/10.1080/15389588.2014.909592
Sagberg, F., & Ingebrigtsen, R. (2018). Effects of a penalty point system on traffic violations. *Accident Analysis and Prevention, 110*71-77. http://dx.doi.org/10.1016/j.aap.2017.11.002
Savolainen, P. T., Gates, T. J., Qu, T., Bamney, A., & Jashami, H. (2022). Developing A Consistent Data Driven Methodology to Multimodal, Performance Based and Context Sensitive Design.
Savolainen, P. T., Gates, T. J., Kassens-Noor, E., Gupta, N., Mahmud, M. S., Kay, J., Johari, M. M., . . . Geedipally, S. (2022). Evaluating the Impacts of the 2017 Legislative Mandated Speed Limit Increases.
Sayers, A., Ben-Shlomo, Y., Blom, A. W., & Steele, F. (2016). Probabilistic record linkage. *International journal of epidemiology*, *45*(3), 954–964. https://doi.org/10.1093/ije/dyv322
Tainter, F., Fitzpatrick, C., Gazillo, J., Riessman, R., & Jr, M. K. (2020). Using a novel data linkage approach to investigate potential reductions in motor vehicle crash severity – An evaluation of strategic highway safety plan emphasis areas. *Safety Research, 74*9-15. https://doi.org/10.1016/j.jsr.2020.04.012
2022 Oregon Motor Vehicle Traffic Crashes Quick Facts. (2024, October). Oregon Department of Transportation (ODOT). www.oregon.gov. https://www.oregon.gov/odot/data/pages/crash.aspx