Using statistical modelling to predict crash risks, injury outcomes and compensation costs in Victoria.

Renee Schuster[a], Michael Nieuwesteeg[a], Amanda Northrop[a], Cameron Lucas[b], Daniel Smith[b]

a. Transport Accident Commission, Victoria, Australia

b. Taylor Fry Consulting Actuaries

Abstract

Background: In 2011, Victoria's Transport Accident Commission (TAC) built a rich linked crash database to explore the research question: "What are the significant variables in predicting crash risk, injury outcomes and compensation costs when controlling for all other variables"?

Aims: The core aims of the TAC Road Safety Risk Models project were to conduct sophisticated analyses of available data to identify key drivers of road trauma, injury severity and compensation costs, as well as identify key target markets.

Method: The project began with an intense data build involving the sourcing, linking and cleansing of road safety and related data. This included crash and compensation data, as well as exposure data on Victorian licence holders and registered vehicles. Detailed injury data was also obtained. A series of statistical models were then developed to examine the relationship between person, vehicle and crash variables, along with injury severity and compensation costs.

Results: A number of pre-crash variables were found to be significant predictors of crash risk and severity including vehicle, person and geo-demographic variables. Injury severity was found to be the most significant variable at predicting compensation costs.

Conclusions: The established database provides a benchmark for future Road Safety policy analysis, particularly with consideration given to the cost of injury to society. With the prospect of new and improved data availability for key input datasets, the TAC has begun to update the linked dataset and refresh the models to identify new relationships.

Introduction

Transport Accident Commission Road Safety Research

The Transport Accident Commission (TAC) is a statutory no-fault compensation scheme that provides coverage for all persons injured in transport accidents in Victoria. The TAC is a "no-fault" insurance scheme. The reference to "no-fault" means that medical benefits will be paid to an injured person regardless of who caused the crash. The TAC is funded by compulsory payments made by Victorian motorists as part of the vehicle registration and annual renewal process.

A key function of the TAC is "to promote the prevention of transport accidents and safety in use of transport" (Transport Accident Act 1986). This means that the TAC is also responsible for delivering public education and Road Safety programs aimed at reducing road trauma. The TAC works in partnership with Victoria Police (Police), VicRoads and the Department

of Justice (DoJ) to deliver these objectives. Research has always played a significant role in developing TAC Road Safety and Marketing initiatives.

The TAC and other Road Safety agencies rely heavily on police reported crash data to inform strategies and measure progress. The TAC has long maintained a link between data held on its claimants and data recorded by police about the crash. This enables the construction of a linked dataset, thus providing a rich source of crash information supplemented with injury outcomes. In recent years the TAC has engaged widely with its stakeholders to increase its evidence base and enhance analytics. More recent acquisitions include regular snapshots of the VicRoads Vehicle Registration and Licence Holder databases, more detailed information on injury classifications and severity and estimates of lifetime cost of TAC claims. The TAC has also utilised geocoding software to improve address accuracy and append geographic based socio-demographic data.

The research strategy has now begun to move beyond the acquisition and improvement of data towards data exploration, and the discovery of insights that provide clear direction to the Road Safety and Marketing program.

Road Safety Risk Models Project

In 2011, the TAC compiled a rich multi-source crash database to explore the research question: "What are the significant variables in predicting crash risk, injury outcomes and compensation costs when controlling for all other variables"?

The core aims of the TAC Road Safety Risk Models project were to conduct sophisticated analyses of Road Safety and related data to identify:

- key target markets, and
- key drivers of road trauma, injury severity and TAC compensation costs.

The project progressed throughout 2011 and involved building a suite of statistical models to identify significant variables when predicting crash probability and crash severity. The TAC contracted Taylor-Fry Consulting Actuaries to work with analysts from the Road Safety & Marketing Team to develop and deploy the Models.

Method

Data Build, Exploration and Preparation

During the data build phase, multiple data sources were used to create a database of all persons and vehicles involved in Victorian road crashes between 2006 and 2010. In addition to the linked crash database, the TAC Project team also prepared exposure datasets to facilitate estimates of crash probability.

A diagram of the data build phase including administrative input datasets, data enhancements and output datasets is provided as Appendix 1. A broad overview for each of the elements of the data build phase is also provided.

The project then progressed to the data exploration and preparation phase, which involved data familiarisation, assessment of data quality and suitability for modelling, data cleansing and the preparation of final datasets. During this phase, the project teams worked together to develop an optimal modelling plan that met the objectives of the project while fitting within

the limitations of the available data. For example, it was necessary to separate out the probability modelling into person and vehicle models as we had no information on the usual driver(s) of a given vehicle for those vehicles that are not crash involved; that is, it could not be assumed that a registered vehicle owner was the usual driver of that vehicle. Furthermore, a decision was made to separate out the person and vehicle probability models into single and multiple vehicle crashes as "fault" information was not always available in the crash data; however fault could be assumed in the single vehicle crash models.

Modelling

Table 1 below outlines the series of models that were subsequently developed.

Table 1: Summary of TAC Road Safety Risk Model

	Cost Severity	Injury Severity	Vehicle Probability	Person Probability
Model Details	Generalised Linear Model (GLM) fitting the natural log of cost.	A series of binomial GLM models using a logit link function.	Two binomial GLM models using a logit link function.	Two binomial GLM models using a logit link function.
	Models the no-fault lifetime cost of a TAC claim.	Models the probability of a TAC claim being minor / moderate / serious / severe injury.	Models the probability of a registered vehicle being involved in a road accident which resulted in a TAC claim in a single year.	Models the probability of a licence holder having a road accident and making a claim where they were the driver of a vehicle in a 5 year period.
		A series of models were developed.	Separate models for single vehicle and multiple vehicle crashes.	Separate models for single vehicle and multiple vehicle crashes.
Input Data	Claimants.	Claimants.	(Crashed and claimed) Vehicles. Registered Vehicles (Exposure).	Claimants. Licence Holders (Exposure).
Model Notes	Includes pre and post crash variables.	Used only variables known prior to the crash.	Used only variables known prior to the crash.	Used only variables known prior to the crash.
	Only variables relating to the claimant and the vehicle they were occupying were used. Details of other vehicles involved in the crash were not.	Only variables relating to the claimant and the vehicle they were occupying were used. Details of other vehicles involved in the crash were not.	Only variables available in both the "crashed and claimed vehicles" file and VicRoads registration file were used.	Only variables available in both the claimants file and VicRoads licence file were used.
	Very high and low cost claims were excluded (nb: removed 9%).	Only uses claims where an injury score was available (approx 55%).		Only used claimant records where the claimant was the driver or rider.

The first step of the modelling phase involved variable testing and selection. A large number of variables from the input data sets were initially included in the models. The modelling then undertook an iterative approach whereby the least significant variables were omitted one at a time. Once insignificant variables were omitted, the process progressed to the simplification of continuous and categorical variables. The continuous variables (such as age and income) were split into two or more ranges based on a visual analysis of the plotted observations. Splines were then fitted and tested to ensure the slope of each range was statistically different from the next. Categorical variables (such as vehicle make and model) were "grouped" where they were not significantly different from each other. Interaction effects between variables were also examined. Interaction effects are used when the effect of two variables combined is significantly different to the sum of the effects of each individual variable.

Results

Unlike the Cost Severity Model, the Injury Severity Models used variables known prior to the crash only; thereby identifying some useful risk variables to consider in developing future road trauma prevention strategies. Furthermore, unlike the probability models where the variables we could include were greatly constrained by the exposure datasets, we were able to test many more variables with the Injury Severity Models. For these reasons, this paper

presents more detailed results on the Injury Severity Model. High level results only from the other models are summarised thereafter.

Injury Severity Modelling Results

The Injury Severity Models predict the probability of a TAC Claim being minor, moderate, serious and severe injury severities; using Maximum Abbreviated Injury Score (AIS). AIS is an anatomical-based coding system to classify and describe the severity of specific individual injuries. A score of between 1 and 6 (labelled as minor, moderate, serious, severe, critical and maximum) is assigned to each individual injury. Maximum AIS is the score of the person's most severe injury. Due to the very small number of high severity claims in the input datasets, claimants with a Maximum AIS score of 4 and above were grouped into the "severe" injury group for the purpose of these models.

Table 2: Injury Severity Models - Significant Variables

Person	Age
	Gender
	Licence Type
	Seatbelt / Child Restraint / Helmet
	Seating Position
Geo-demographic	Proportion with at least a bachelors degree (in local area)
Crash	Crash Date
	Speed Zone
Vehicle	Vehicle Intent
	Vehicle Offending
	Vehicle Type
	Year of Manufacture

The following charts present raw probabilities and pure effect relativities for selected significant variables in the Injury Severity Models. The probability charts on the left show the actual relationship between the predictor variable (e.g. age) and the modelled variable (in this case, injury severity). This is equivalent to the raw, un-modelled data without controlling for other predictor variables. The relativity charts on the right hand side show the pure effect on the model (after controlling for other significant predictors) for values of the predictor variable. For continuous variables, a value of 100% translates to no effect, a value of less than 100% translates to a lower severity and values greater than 100% translate to a higher severity. For categorical variables, one factor was chosen as a base which would obtain 100% relativity, and the other factors would be given a relativity score in relation to it.

Figure 1: Injury Severity Models – The Effect of Age at Crash

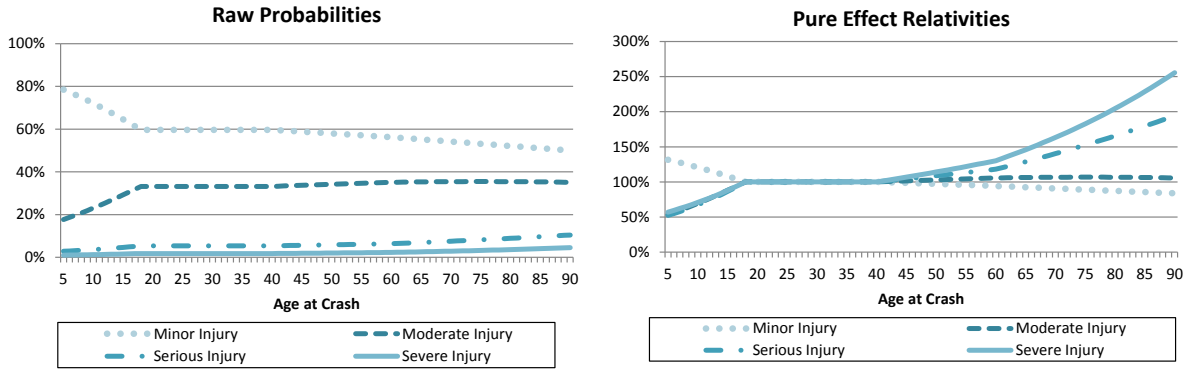


Figure 2: Injury Severity Models – The Effect of Gender

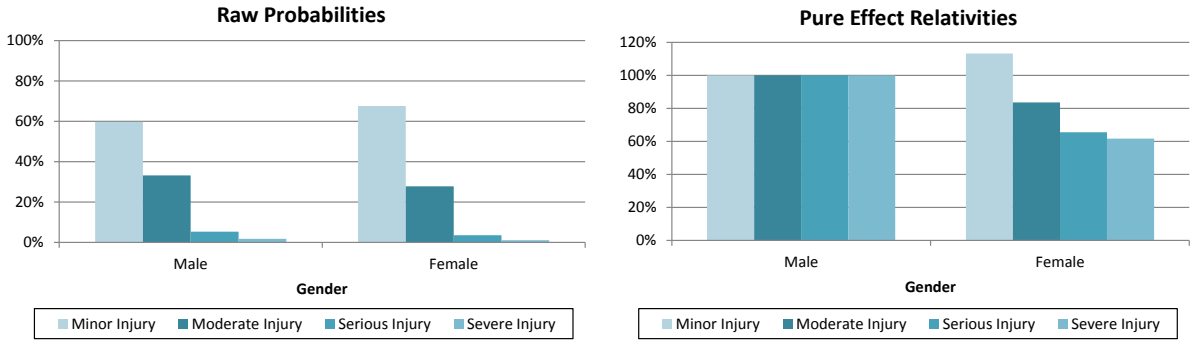


Figure 3: Injury Severity Models – The Effect of Licence Type

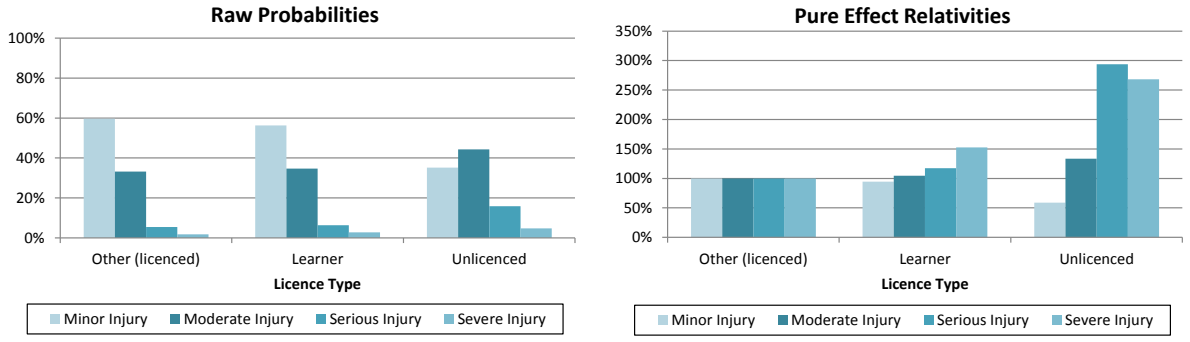


Figure 4: Injury Severity Models – The Effect of Seating Position and Seatbelt/Helmet

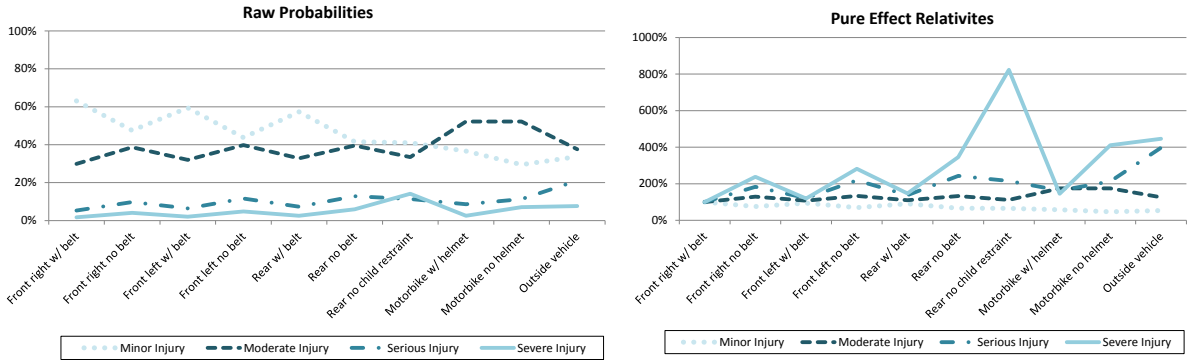


Figure 5: Injury Severity Models – The Effect of Speed Zone

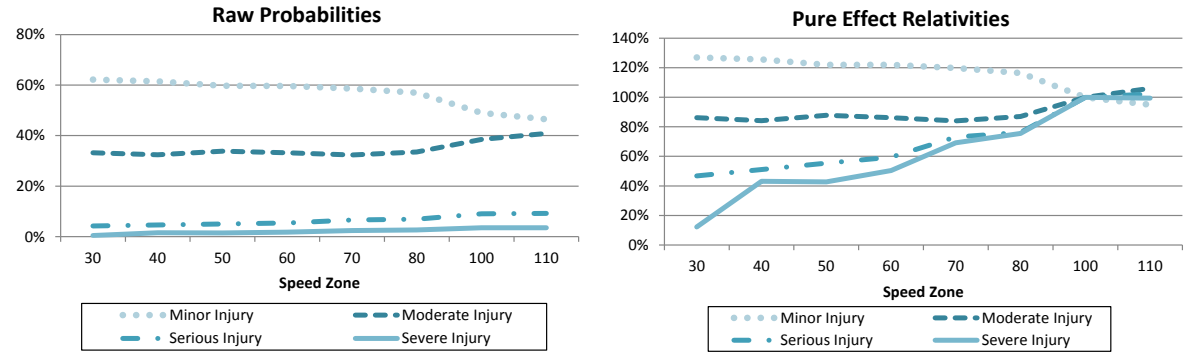
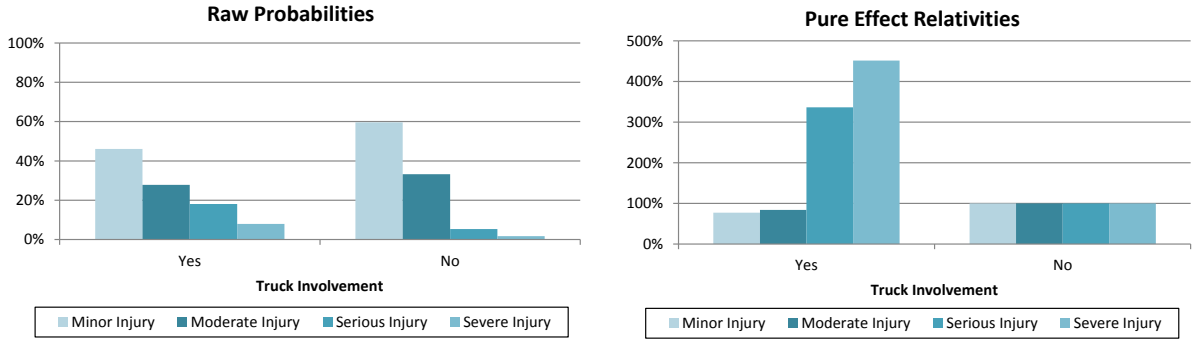


Figure 6: Injury Severity Models – The Effect of Truck Involvement



Selected observations from the results of the Injury Severity Modelling include:

- Severity of injury increases with age.
- Males were at a significantly higher risk of serious injury than females.
- Learner drivers have relatively more severe crashes, but unlicensed drivers are far more likely to have serious or severe crashes.
- Passengers in general are worse off than drivers. Rear seat occupants specifically were at 50% higher risk of severe injury than drivers.
- Seat belt use and helmet use is extremely protective, particularly for children.
- Motorcyclists have more severe injuries, particularly if a helmet is not worn.
- Pedestrians are slightly worse off than a motorcyclist not wearing a helmet.
- The faster the speed, the more severe the injuries.

- Risk of serious injury was 4.5 times higher in truck involved crashes compared to standard vehicles.

Overall Project Findings

Injury severity was found to be the most significant variable at predicting TAC compensation costs. As expected, the cost of a claim grows with increasing injury severity.

18 year olds are by far the most at risk. Males are worse than females in terms of probability of single vehicle crashes and severity of all crashes. Motorcyclists contribute significantly to the probability of a claim for males, especially in the 30-50 age group. When motorcyclists are excluded, males and females have a similar claim probability distribution when single and multiple vehicle crashes are combined. Motorcyclists in general have a higher claim probability, are more likely to be involved in single vehicle crashes, have more severe injuries (which are exacerbated if a helmet is not worn) and have slightly higher compensation costs for similar severity of injury.

Geo-demographic variables, and particularly socio-economic variables, have a significant influence on injury severity, claim probability and compensation costs. Language barriers tend to increase claim costs but potentially lead to a lower probability. Increased income and education in the area where a claimant lives lead to lower probability, lower severity and lower costs for injuries of the same severity.

Newer cars are less likely to be involved in a serious crash, and when they are, the compensation costs are relatively lower. Some vehicle makes, models and types are more prone to single or multiple vehicle crashes. For example, Commodores and Falcons are more likely to be involved in single vehicle crashes than small or expensive cars. The impact of different vehicles is typically watered down in multiple vehicle crashes given that fault was not considered. It is important to note the potential bias in interpreting the results of the vehicle probability model and the impact of different vehicles given that driver characteristics were not included. For example, the results indicating Commodores and Falcons are high risk vehicles may be caused by the types of drivers of these vehicles, rather than the vehicles themselves.

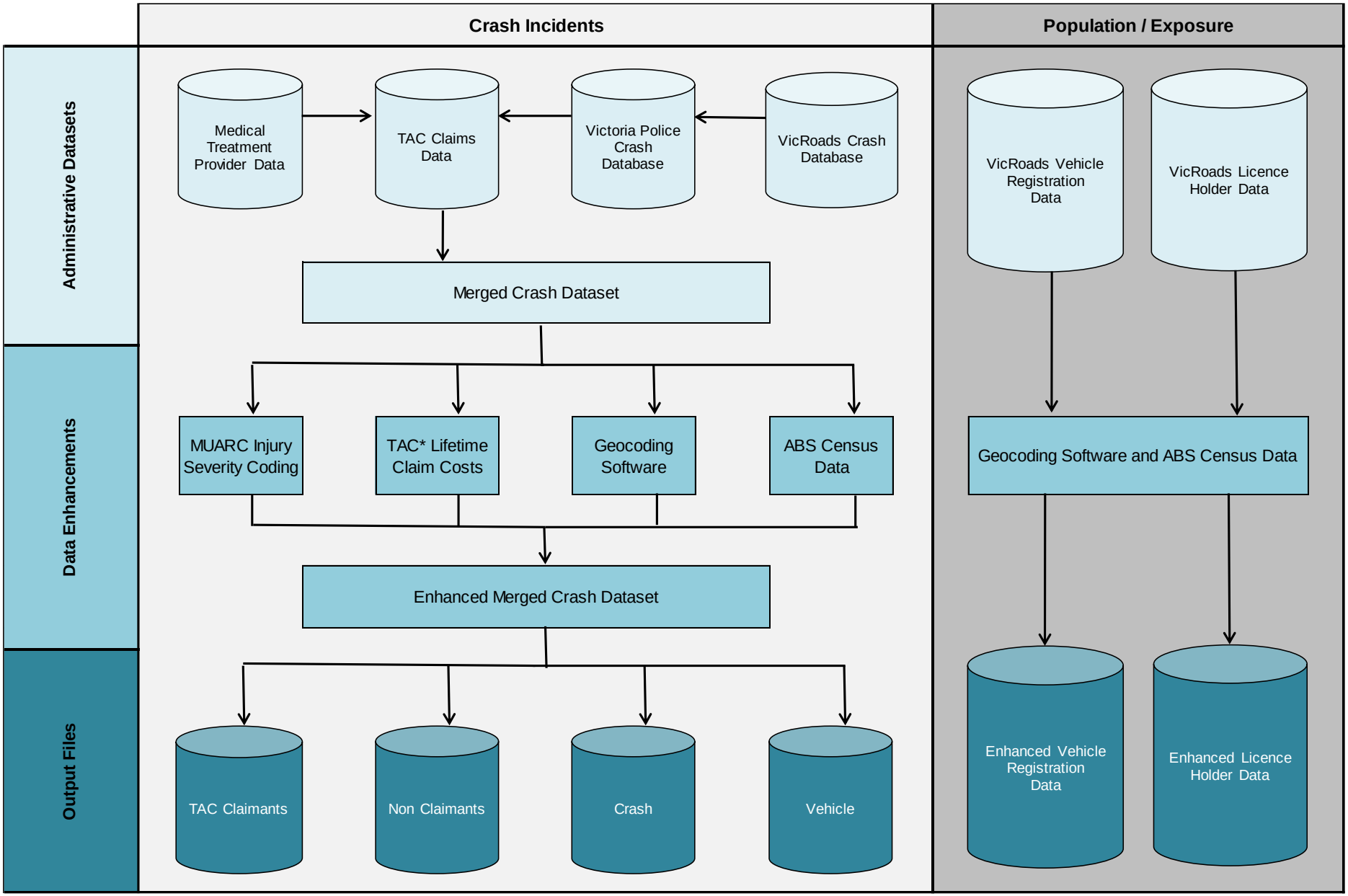
Conclusions

The established database provides a benchmark for future Road Safety policy analysis, particularly with consideration given to the cost of injury to society. The TAC has begun to update the linked dataset with updated data, and some new and improved data acquisitions. This includes updated TAC lifetime compensation cost estimates, updated ABS Census data, more detailed vehicle specification data and vehicle crash worthiness ratings. The TAC now also has more regular snapshots and a wider range of data on Victorian licence holders from VicRoads, including demerit point history and licence conditions. Future work will also explore utilising self-reported injury data collected by the TAC claims department, in addition to injury codes obtained from the hospitals, to create a richer dataset of injury severity. Future work may also entail sourcing private car insurance data to better understand the population of road users that are not involved in crashes. Although a large suite of variables were tested in the initial models, many more will be tested in future. These include

whether the claim had previous psychology, chiropractic or physiotherapy treatment and whether they had pre-existing drug or alcohol issues.

Limitations of previous models will be explored and many aspects of the models will be reviewed and refreshed. This will ensure the models reflect the current state of play and new relationships are identified.

Appendix 1


Crash Incidents
Population / Exposure
Administrative Datasets
Data Enhancements
Output Files
Medical Treatment Provider Data
TAC Claims Data
Victoria Police Crash Database
VicRoads Crash Database
VicRoads Vehicle Registration Data
VicRoads Licence Holder Data
Merged Crash Dataset
MUARC Injury Severity Coding
TAC* Lifetime Claim Costs
Geocoding Software
ABS Census Data
Geocoding Software and ABS Census Data
Enhanced Merged Crash Dataset
TAC Claimants
Non Claimants
Crash
Vehicle
Enhanced Vehicle Registration Data
Enhanced Licence Holder Data

Administrative Datasets

TAC Claims Data: data is largely collected for the purpose of claims management and processing. The dataset includes demographic data, residential address at claim lodgement, occupation, injuries sustained, medical treatments received and selected crash related information.
Medical Care Provider Data: For each TAC claimant, the TAC receives detailed data on injuries being treated by medical care providers. These providers include the Department of Health, surgeons, doctors, physiotherists and counsellors.
Victoria Police Traffic Incident System (TIS): Contains information on all traffic accidents reported to Victoria Police. This includes information on crash involved persons and vehicles, crash circumstances, crash location, road (and roadside) features and conditions, and weather conditions.
VicRoads Road Crash Information System: Contains a subset of TIS; all persons involved in accidents where at least one person was injured. VicRoads has a team of coders that validates and revises selected data collected by police members, particularly in relation to crash location and road characteristics.
The VicRoads Vehicle Registration Information System: Holds information on all vehicles registered in Victoria. It includes information such as registration number, Vehicle Identification Number (VIN), vehicle specifications (e.g. make, model, type, class etc.), garage address and details of the vehicle owner.
The VicRoads Driver Licensing System: Holds information on all persons holding a Victorian Driver/Rider licence. This includes (for each licence holder) basic demographic information, residential address, and licence type and proficiency.

Data Enhancements

Monash University Accident Research Centre (MUARC) Injury Severity Coding: MUARC assisted with mapping hospital injury codes to a range of injury severity measures; including the Abbreviated Injury Scale (AIS). AIS is an anatomical-based coding system to classify and describe the severity of specific individual injuries. This new data provided a more simple numerical method for grading and comparing claimant injuries by severity.
TAC Modelled Lifetime Claim Costs: TAC actuaries calculated estimates of outstanding claim liabilities, which were added to "to date" claim payments to estimate the lifetime cost of individual claims. These costs were all indexed to values as at June 2011.
Intech IQ Standardiser: a software package designed to validate and correct address data, and subsequently undertake geocoding; which involves assigning geographic coordinates and other geographic codes (including ABS Census Collection District (CCD)) to each address. Claimant first known address, the residential address of all licence holders and garage address for all vehicles were validated and geocoded where possible.
The Australian Bureau of Statistics 2006 Census data, aggregated to Collection District and postal area was obtained. Socio-demographic data (such as ancestry, income and education) was appended to person and vehicle level address data where possible; including claimant address, licence holder address and vehicle garage address.

Output Files

The **Claimant**s file contains a record for each claim in the period 2006 to 2010 from all persons who had a claim accepted by the TAC. This dataset contained approximately 85,000 records.
The **Non-Claimants** file contains all persons who were involved in a road traffic accident in the period 2006 to 2010 but did not have an accepted claim with the TAC. This dataset contained approximately 470,000 records.
The **Vehicle** file contains a list of all vehicles involved in a transport accident (with or without a TAC claim), which was reported to Victoria Police and/or the TAC over the study period. This dataset contained around 400,000 records.
The **Crash** file contains a list of all transport accidents (with or without a TAC claim), which was reported to Victoria Police and/or the TAC over the 5 years of interest. The final dataset contained approximately 240,000 records.