Institute of Actuaries of Australia

Statistical Case Estimation for Long Term Claimants

- Uncovering Drivers of Long Term Claims Cost in Accident Compensation

Prepared by Mitchell Prevett and David Gifford

Presented to the Institute of Actuaries of Australia
XIth Accident Compensation Seminar 1-4 April 2007
Grand Hyatt Melbourne, Australia

The Institute of Actuaries of Australia
Level 7 Challis House 4 Martin Place
Sydney NSW Australia 2000
Telephone: +61 2 9233 3466 Facsimile: +61 2 9233 3446
Email: actuaries@actuaries.asn.au Website: www.actuaries.asn.au

Tables of contents

Abstract

This paper investigates the drivers behind long term care costs for long term claimants in a no-fault scheme using predictive claim modelling. We will consider the various predictors which are available, and discuss their usefulness in projecting the costs for these claimants.

We will also discuss an approach to modelling and projecting these costs on an individual claimant basis. The approach involves separating predictive variables into static and dynamic groups and further defining a single dynamic state variable which captures the majority of the information in the dynamic predictors. We will describe a transition model approach which is then constructed to project forward the dynamic predictor variables as an input into the claim cost model. The claim cost models predict the claim payments using all known static predictors and the constructed dynamic claim state variable. We will discuss the structure of these models and the insights into the drivers of cost derived from them.

Finally, we will discuss some of the systemic factors which need to be considered in estimating the lifetime cost of such claimants, as well as the usefulness of such an approach in the management of these claims.

Key words: *attendant care costs, long term care costs, long tern claimants, catastrophic claimants, predictive claims modelling, static and dynamic, transition model, drivers, systemic factors, claims management*

1 Introduction, Background and Acknowledgements

The purpose of this paper is to discuss the approach to identifying the drivers of the cost of long term claimants in accident compensation. A thorough understanding of these drivers is then used in designing a modelling methodology for projecting the costs of these claims.

Long term claimants, while relatively small in number, represent a significant proportion of outstanding claims liabilities for schemes such as the TAC. Furthermore, such benefits will increase as a proportion of total liabilities as such claimants will typically continue to receive long term care benefits until death, while less seriously injured claimants will typically cease receiving benefits at an earlier stage.

There is also significant uncertainty associated with such liabilities, due to the very long term nature of the benefits and the uncertain impact of:

- Changes in claimants' long term care needs;
- The availability of care; and
- The cost to be paid for care.

While these factors mean that any discussion regarding these liabilities should be accompanied by appropriate disclaimers regarding the associated uncertainties, in our view it is still desirable to understand the cost drivers of these claims and use these to predict the ultimate cost of the claims. In other words, such an approach should reduce the uncertainty to the "unknown" items listed above, and reduce the uncertainty associated with "known" items (i.e. those captured in the data).

Long Term Claimants/Long Term Care

The model described in this paper has been derived for long term care (attendant care and accommodation) payments in respect of claims within the TAC's Community Support Division. This group primarily consists of paraplegics, quadriplegics and neurologically impaired claimants.

The general approach outlined in this paper does not depend however on the definition of long term claimants used herein. It may be applied to any group of claimants which require significant amounts of long term care, where a reasonable history exists.

Claim Cost Drivers and the goals of accident compensation schemes

The primary purpose of the model documented in this paper is the prediction of claims costs in respect of long term claimants, based on an examination of data items or "Claim Cost Drivers" which are found to have a predictive relationship with claims costs.

We note however that for schemes such as the TAC, the claims cost associated with long term claimants (or indeed, any claimants) is only one of several outcomes used in measuring the success of the management of such claimants. In addition to cost, claimant satisfaction and successful integration into the community following an accident can equally be regarded as demonstrating successful claims management.

Predictive Claims Modelling / Statistical Case Estimation

A Statistical Case Estimation ("SCE") model is a particular type of predictive claims model which provides individual estimates of future claim costs arising from existing, open claims. These SCEs are predicted via a statistical model using the individual characteristics of each claim.

Predictive claims model are particularly useful to organisations because they implicitly link underlying drivers to outcomes of interest at the individual claim level. The models themselves help us perform three key tasks:

- Provide a stronger link between changes in the claims processes and reserving
- Provide estimates of future claim costs arising from existing, open claims
- Provide an understanding and quantification of the drivers of a claim

There are many benefits from using predictive claims models in a number of applications including:

- Better usage of individual claims data for forecasting and reserving
- Updating reserves/re-reserving for changes to legislation
- Better understanding of the claims drivers, enabling implementation of more appropriate claims management initiatives
- Improved quantification of benefits from/monitoring of claims management initiatives
- Identification of high risk claims

Background to the TAC

The Transport Accident Commission provides benefits and compensation to parties injured in transport accidents in Victoria since 1 January 1987. The scheme differs to most other Australian compulsory third party schemes in that most benefits are provided on a "no-fault" basis. This means that most benefits are provided to injured parties regardless of whether they were at fault or not. The benefits provided by the scheme are typically paid on an ongoing basis and cover components such as medical treatment, hospital costs and paramedical, economic loss or income replacement, rehabilitation costs and long term care costs.

In the year ended 30 June 2006 the TAC paid benefits totalling approximately $675 million, while the discounted central estimate of outstanding claims liabilities, excluding claims handling expenses and prudential margin, was approximately $4.75 billion. These figures demonstrate the very long duration of the TAC's liabilities, a significant proportion of which relate to the ongoing attendant care and accommodation costs provided in respect of long term claimants.

Acknowledgements

We would like to acknowledge the TAC who has graciously allowed us to publish this paper, containing results relating specifically to their long term claimant portfolio. As accident compensation within Australia moves towards no-fault coverage of people severely injured in transport accidents, the expertise and support of the TAC will be invaluable.

We would also like to thank John Walsh, Ian Reed and Nina Nissan for their time, spent reading our paper, and for their valuable comments.

2 Statistical Case Estimation Modelling

Considerations for Model Design

When designing and constructing an SCE model we need to balance a number of objectives to ensure that the model is suitable. The final model needs to:

- Capture the material drivers of claims cost and hence the mix of claim types in a portfolio
- Be transparent such that we can identify which drivers are included in the model and how they effect the results
- Balance the credibility and stability of drivers with their "predictiveness" (the ability of the model to predict the ultimate cost for each claim)
- Be stable in times where experience is also stable
- Be responsive in times when emerging experience is changing
- Identify superimposed inflation and trends in claim drivers

In design phase of any SCE model we need to have consideration of these objectives and test a number of approaches to arrive at the final design.

Static and Dynamic Drivers

The model design is based on individual claim characteristics which are proven to be good drivers of claim costs. However, a complication arises with the use of some of the characteristics due to their instability. We can first classify them into three broad groups to better understand this issue.

1 Static predictors. These are largely known at the beginning of the business process or claim life cycle and will not change subsequently. Examples include gender and date of accident;

2 Dynamic (foreseeable) predictors. These are predictors that will change in the future in a foreseeable way, for example the age of the customer / claimant will change in the future but can be calculated exactly at any specific date. Another example in accident compensation is duration since accident.

3 Dynamic (stochastic) predictors. These are predictors that will change over the lifetime of the claimants in a predictable way but with a stochastic or random element. Examples in accident compensation include litigation / common law status, injury severity, impairment level, care needs, rehabilitation plan, investigation status and residential status.

The third category of drivers is the most problematic from a modelling perspective. An individual claim characteristics model can be constructed using historic dynamic predictors where they were known with certainty. However for prediction purposes the model is applied to open claims in the future, when the value of future predictors are not known. Treating dynamic predictors as if they are static will cause biased results (consistent over or under estimation). A simple example of this bias is presented below:

- A claim is currently not litigated but there is a 25% probability that the claim will proceed to common law before finalisation.
- Let us assume that litigated claims cost $100,000 and non-litigated claims cost $50,000.

- Taking the litigation status as if it is static the SCE for the claim will be $50,000. There will be no allowance for the probability of proceeding to common law.
- The unbiased SCE should represent the probability weighted expected value, E(SCE) = 25% x $100,000 + 75% x $50,000= $62,500.

Our approach involves a detailed investigation into which predictors would be deemed static and which were dynamic. It was not surprising that many of the dynamic variables are in fact the most predictive of the claim cost. To use dynamic stochastic predictors in the SCE model we must build a separate model for the forecasting of these predictors which is discussed further below.

Forecasting Dynamic Variables

Based on the previous discussion about static and dynamic variables we have established that the use of dynamic variables as if they are static will result in biased predictions. It is also generally the case that the dynamic variables are in fact the most predictive of the future claim payments. We have attempted to summarise the majority of the information in the dynamic predictors into one dynamic claim state variable. This will retain the ability of the model to use current dynamic characteristics but only requires the projection of one variable into the future. The claim state is a combination of the most important dynamic characteristics and is projected into the future. If we wish to build a model with dynamic claim characteristics we need to know how these characteristics change over time. A "Transition Model" approach can be used to project forward dynamic variables in the following way:

1 The dynamic variable needs to be redefined to contain a number of states in which the claim can fall (e.g. 2 states – zero payments or non-zero payments in the past year)

2 The current state for each claim is the starting point for projection

3 We estimate the probability that each claim will "transition" from one state to another over the next time period (i.e. 1 year)

4 The process is repeated for each subsequent period, one at a time, using the probability of being in each state in the previous period – this process is known as "chaining".

In summary, the transition model predicts the probability of a claim "transitioning" or moving between 2 of the claim states, at any future point in time. We chose to only model and project a single state variable with a reasonably small number of states (say 4 or 5) because the transition models are pivotal to forecasting a sensible and realistic claims cost. Small changes in the probabilities of transitioning can result in very large changes in the projected cost for each claim and hence we have adopted the approach of modelling one simple dynamic variable in a very rigorous way as opposed to modelling many complicated variables less rigorously.

How do we decide on the Claim States?

Firstly we have undertaken an exercise to determine which of the modelling variables are dynamic and which are static. It was self evident that indeed all variables capturing information relating to past payments were dynamic and although some of the other modelling variables demonstrated some dynamicity we decided to assume they were static for modelling.

The next step is to order the dynamic variables by their predictiveness of the future payment levels for long term care benefits. The final step involves simplifying the significant dynamic variables systematically into a single variable while attempting to:

- Maintain enough claims in each state to enable stable estimation of transition probabilities
- Achieve a reasonable level of homogeneity within each state
- Retain a large amount of the predictiveness from all of the dynamic variables
- Achieve a definition that is sensible and interpretable.

The state definition we have derived is presented below. Note that we have removed "nuisance claims" by implementing a small payment threshold on the definitions.

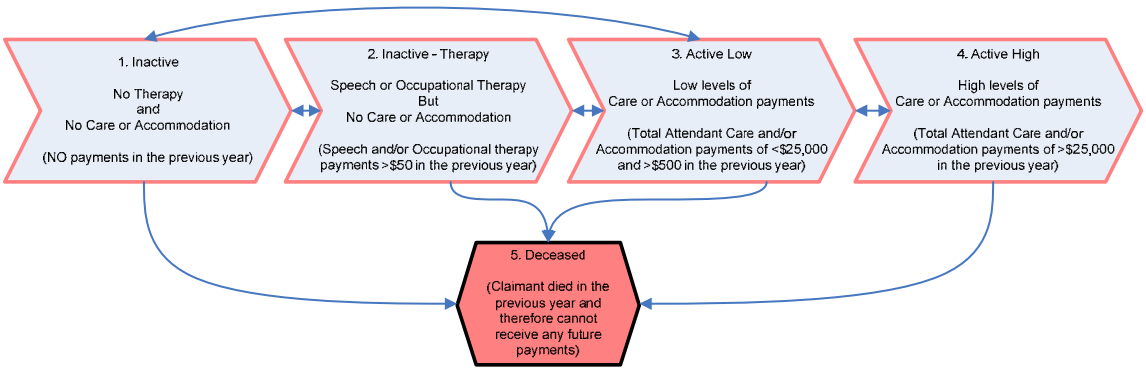

Figure 1 – Dynamic Claim States

1 Inactive - A claim is termed "inactive" when there are payments of $500 or less made for long term care costs (the payment type that we are modelling). There are also no payments for speech or occupational therapy. An inactive claim in state 1 is currently not receiving any payments but will have a probability of becoming active in the future, (which may result from a change in the claimant's needs or family circumstances).

2 Inactive Therapy – These claims are currently not receiving long term care payments but **are** receiving payments for speech or occupational therapy. These claims will also have a probability of becoming active in the future and this is generally more likely than for state 1 claims.

3 Active Low– An active claim is one that is currently receiving payments for long term care. We have sub-divided these claimants into those with low levels of payments/ benefit utilisation who are likely to be receiving low levels of part time care where the primary care is provided by family members.

4 Active High – These claims are more likely to be receiving full time care on an ongoing basis.

5 Deceased – Benefits are provided for the full lifetime of the claimant and as such any care costs will cease upon death.

There was a further motivation for separating the Active claims into the high and low groups which was related to modelling the costs which will be discuss further in the next section.

Forecasting Cashflows

Using the transition model we can attach to each claim a probability of being active (or receiving long term care payments in the next year). The payment amount model predicts the expected payment given the claim is in the active state. The predicted payment level for any claim is based on the probability of the claim being Active (at the future time period) multiplied by the expected payment level, given the claim is Active.

We have chosen to construct 3 payment models for this purpose. Our initial analysis determined that the past payment level is in fact the best predictor of future payment levels however there were a number of further considerations when deciding how to incorporate past payments as a predictor in payment amounts model.

1. Active High - Rate of Change Model (percentage increase in payments)

This model is used to predict the expected payment amount that would be received in the Active High state. For those claims **currently** in the Active High claim state we use a model which predicts the percentage change in yearly payment amounts. This approach has been adopted for the following reasons:

- It incorporates the fact that the most significant predictor of next year's payment is the current year's payment.
- Those claims in the Active High state are expected to be the long term claims receiving regular payments and this approach works well for these claims.

2. Active High - Payment Per Active Claim (PPAC) Model

This model is also used to predict the expected payment amount that would be received in the Active High state. It is used for those claims which are currently **not** in the Active High state. There are a number of reasons why we believe using the Rate of Change model based on the previous year's payment level as a predictor is unsuitable:

- A large number of claims currently in the Inactive states will have a $0 payment amount and a percentage increase on this level is not sensible.
- Those claims in the Active Low payments are more likely to be receiving volatile and infrequent payments and again a percentage increase approach will not produce sensible results.

3. Active Low - Payment Per Active Claim (PPAC) Model

This model is used to predict the expected payment amount that would be received in the Active Low state and is used for all claims which are in the Active Low state (including those currently in the Active Low state). The expected payments to be received in the Active Low state are based on a simple PPAC model which does not rely on the previous year's payments.

Overall Model Structure

The approach we have developed has been largely driven by the factors above, in particular:

- The need to use dynamic predictors in the estimation of annual claims cost
- The need to allow for changes in these dynamic predictors over time
- The fact that past payments are the most predictive of future payments
- Using past payments is not possible for claims or claims with zero or low levels of payments

The diagram below demonstrates how all of the model components come together to produce the total liability estimates.

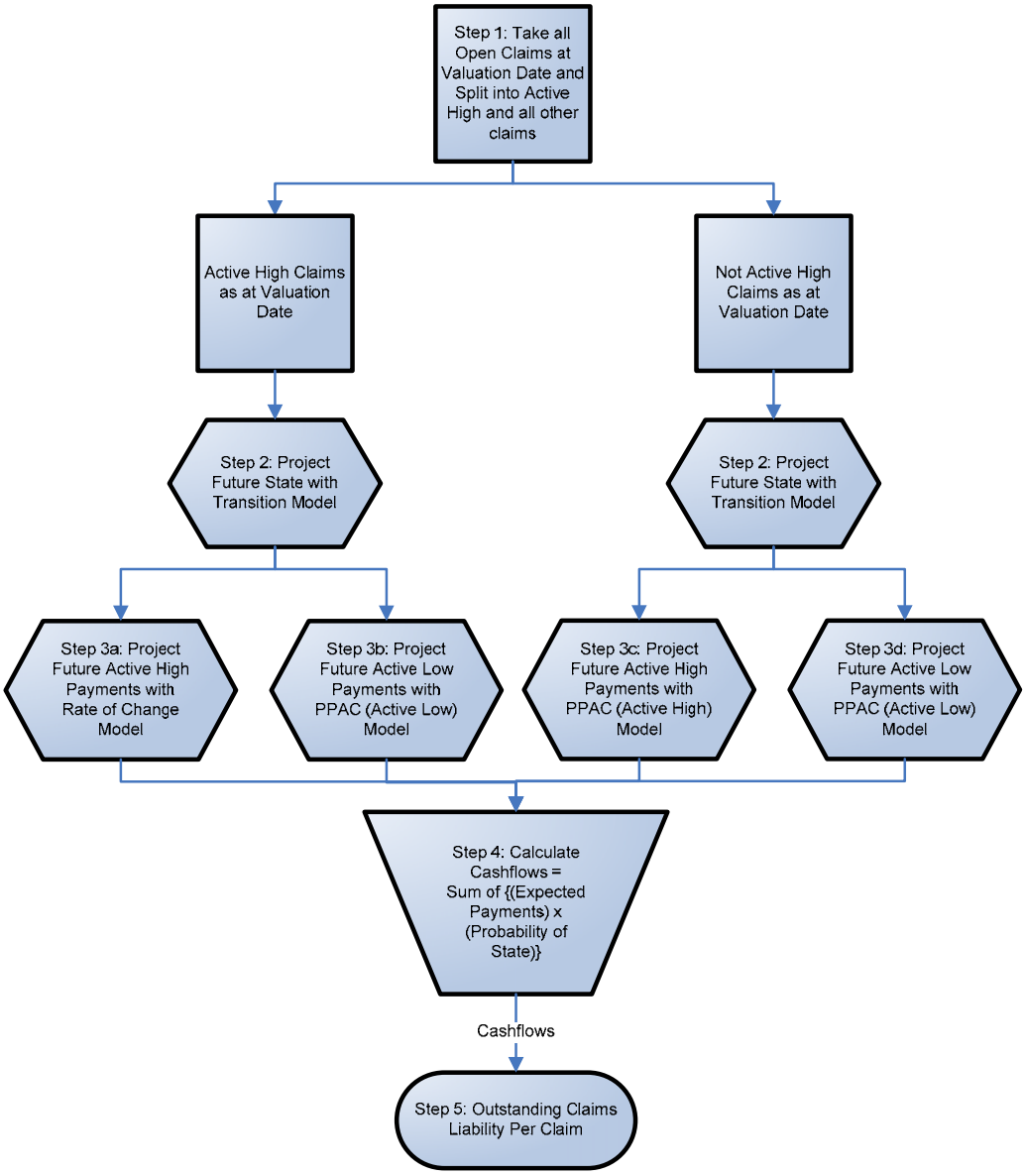

Figure 2 – Overall SCE Model structure

3 Model Construction

Data

The universe of data used for constructing the model needs to match the claims to which the model will be applied. The dataset used for modelling purposes included around 3,500 claims of which around 5% related to quadriplegia, 35% to severe acquired brain injuries and the remaining 61% were termed non-catastrophic.

As mentioned previously the analysis in this project relates to the long term care payment type as it is the most significant from a valuation perspective. The model was constructed to predict the amount of long term care payments for the claims in the modelling dataset, based on payments made between 1997 and 2005.

To predict the amount of long term care payments for each claim in the next year, we had a collection of claimant characteristics available including:

- Injury class (e.g. Quad C5, Severe ABI – 1, Fractures, Fatal)
- Functional code (e.g. minimum function, dependent in most tasks)
- Mobility code (e.g. No use of arms/legs, some use of arms/legs)
- Functional Independence Measure (FIM)
- Functional Assessment Measure (FAM)
- Age at injury and current age
- Gender

To predict the payment level on each claim in the next year we can also use the past payment levels at each point in time in the dynamic state variable. The past payments available for use in the dynamic state variable included:

- Attendant care payments
- Accommodation payments
- Speech therapy payments
- Occupational therapy payments

During the initial stages of the project we undertook a detailed investigation into the integrity of each of these data items. This investigation included looking at the distribution for each of the variables and comparing this with the relevant codes or values that could be taken, as well as identifying any missing or incorrectly coded values. We found that in general most variables were well recorded however the FIM and FAM variables were incomplete for a large proportion of claims.

TAC have undertaken to collect the FIM and FAM variables on all catastrophically injured claims (i.e. Quadriplegics and Severe ABIs) however as at the time of constructing the SCE model only around half of these claims had been assessed and coded. As a result of the incomplete data, the main model for statistical case estimation purposes does not incorporate FIM and FAM. However, as an exploratory exercise to understand the potential importance of these variables we have constructed additional models with them included which will be discussed later in this chapter.

The second phase of our data investigations included comparing each of the potential predictor variables with the other predictors to understand any correlations between them. This stage is essential to understanding which variables may not add anything to the model over and above other variables, and also understanding the relationships between variables. As injury classes for spinal injuries are constructed based on mobility codes, the injury class and mobility code for these claimants is 100% correlated. Similarly the injury class and functional code for severe brain injuries are also 100% correlated. Much of the explanatory power of the functional and mobility codes would therefore already be contained in the injury class. For model construction it was therefore cleaner to incorporate any additional information from the mobility and functionality codes into the injury class variable. As such we created an "expanded" injury class with additional levels for some injuries based on either the mobility of functionality codes, where there was enough data to do so.

The final phase of the data investigations was to understand how dynamic each of the modelling variables was. The process for testing this property is to take the value of each variable across the historic snapshots for that claim and calculate the proportion of successive records that are changing. For the variables with a significant level of dynamic records we then further calculate the pattern of records changing by development year to determine if the dynamic property stabilises after a period of time. On this basis we determined that the variable which measures injury severity demonstrates some dynamic behaviour at early durations, however is more stable at later durations, and for the purposes of our model was regarded as static.

Transition Models

One of the first steps in constructing the individual transition models is to investigate which of the transitions are worth modelling and which are too immaterial for consideration. The table of the overall transition probabilities is useful in making these decisions. We note that we have only modelled transitions between the states 1 to 4 i.e. not including the death state. The probabilities of transitioning to the death state were based on mortality assumptions which, for the purposes of this paper, will not be discussed in detail.

Table 1 – Overall Transition Probabilities

Starting State	Transition to	Number of transitions	Probability of transition
Active High	Active High	2473	90.5%
Active High	Active Low	232	8.5%
Active High	Inactive Therapy	13	0.5%
Active High	Inactive	15	0.5%
Active Low	Active High	317	10.9%
Active Low	Active Low	2033	70.1%
Active Low	Inactive Therapy	212	7.3%
Active Low	Inactive	339	11.7%
Inactive Therapy	Active High	41	1.9%
Inactive Therapy	Active Low	217	10.3%
Inactive Therapy	Inactive Therapy	902	42.8%
Inactive Therapy	Inactive	946	44.9%
Inactive	Active High	137	1.0%
Inactive	Active Low	325	2.3%
Inactive	Inactive Therapy	730	5.1%
Inactive	Inactive	13019	91.6%

These overall probabilities showed that:

- 91% of Active High claims continue to be Active High in the next year.
- Around 9% of the Active High claims transition to Active Low and essentially 0% transition to either of the Inactive states
- The Active Low state is a "transient" state with many claims transitioning to Active High or the Inactive states each year.
- Inactive Therapy is a feeder state for Active Low
- Most Inactive claims remain Inactive.

Using this analysis we were able to determine that some transitions are so infrequent that we ignore them for modelling purposes. In total there are 12 possible transition probabilities required to be modelled (given that the probability of entering the death state is based on the mortality assumptions). Generalised Linear Models (GLMs) were used to estimate the transition probabilities as described with the following process:

- All claims in the starting state (or from state) were used in the modelling universe for this particular transition
- The outcome variable modelled was a binary variable taking the value of 1 if the claim transitioned to the ending state (or to state) and zero otherwise.
- Any static or dynamic foreseeable claim characteristics could be used in the models (e.g. current age and age at injury, number of years since accident, year of accident, gender, category of injury, severity of injury).
- The model goodness of fit was assessed using a variety of tests including actual vs. expected comparisons for each variable in the model.
- When combined for each claim for projection, the transition probabilities must be rescaled to ensure that they add to 100%.

To provide some background into the process and diagnostics used for the transition models we will outline an example model. The example is the transition from the Active High state to the Active High state. This is a crucial model for the overall SCE as it represents the "continuance" probability for remaining on high levels of long term care payments, for the most severe claims (alternatively, 100% less this probability is the "discontinuance" probability). The final GLM fitted is presented in the table below.

Table 2 – Parameter Estimates for Transition Model from Active High to Active High

Variable	Function	Parameter Estimate	Log Odds Ratio	Signifcance (Pr > ChiSq)
Intercept		2.1888		<.0001
Development year	min((max(devyear-1,0)),7-1)	0.1704	118.6%	<.0001
Current age	min((max(currage-10,0)),25-10)	0.065	106.7%	<.0001
Injury class	Non Catastro	-2.7655	6.3%	<.0001
Injury class	Other Sev ABI	-2.4875	8.3%	<.0001
Injury class	Paraplegia - 100% disrptn of funct	-1.9375	14.4%	<.0001
Injury class	Sev ABI - 2, Mobility code gt 4	-0.8117	44.4%	0.0169
Injury class	Sev ABI - 3, Mobility code gt 5	-1.5747	20.7%	<.0001
Injury class	Sev ABI - 3, Mobility code le 5	-1.0048	36.6%	0.0008
Injury class	Sev ABI - 4, Other mobility code	-2.0766	12.5%	<.0001
Injury class	Base	0	100.0%	.

The significant predictors included in this model were the development year, current age of the claimant and the injury class. The model demonstrates the following effects:

- Development Year - As the number of years since accident increases, the probability of remaining in the Active High state increases up to year 8 and then remains flat.
- Current Age of Claimant - A claimant's probability of remaining in the Active High state increase steadily between ages 10 and 25.
- Injury Class - The more severe the injury the more likely that a claimant will remain in the Active High state. The base group includes all quadriplegic claims and the most severe acquired brain injuries (ABIs). All other groups have parameter estimates that are negative and hence the predicted probabilities of remaining Active High will be lower.

Rate of Change in Payments Model

The rate of change model predicts the percentage increase in this year's payments over last year. The model is fitted using a GLM. The Observations used for modelling included:

- Claims that have been in the Active High state and then stayed in the Active High state for the next year. This was to ensure that claims receiving partial year of payments were not distorting the model.
- For fatal claims, we have included all past service years up to but not including the year of death.
- We have excluded observations for claims in development year 0, because it is a partial year of payments.

All available predictors were tested in the model however the significant predictors included in the model were:

- Previous year payments - Higher percentage increases for lower payment levels are observed.
- Injury class – For high severity Injuries, payments increase at a greater rate than other injuries
- Current age – Claimants over the age of 60 increase at a lower rate than those under the age of 60.

For completeness we note that the model fitted was an Inverse Gaussian, with a log link, where the target was the payments in the next year and the offset was the log of the payments in the previous year (to ensure we are modelling the rate of change in payments).

The table below shows the resulting predicted percentage increase in payments over the previous year based on the fitted GLM. The table demonstrates that for Severe Injuries aged under 60 with current long term care payments of $25,000 to $35,000, the model predicts that payments in the next year will be 23% higher than the current year. We note that some of the rates are negative which will decrease the payment level over successive years to the lower bands.

Table 3 – Predicted Rate of Change for Currently Active High Claims

Current Age	Less than 60		Greater than 60	
Injury Class ***	Severe	Other	Severe	Other
Previous Year Payments for ATC and ACC				
25,000 to 35,000	23.2%	13.4%	17.4%	8.1%
35,000 to 50,000	13.5%	4.5%	8.1%	-0.5%
50,000 to 100,000	8.4%	-0.2%	3.3%	-4.9%
Greater than 100,000	2.2%	-5.9%	-2.6%	-10.4%

*** The Severe group contains
'Sev ABI - 1', 'Sev ABI - 2', 'Quad - C1 - C4', 'Quad - C5'

The payments used in the rate of change model (as well as the PPAC models) have been inflated to current values and the rates of change shown are therefore over and above base inflation.

The table below is the complete model with parameter estimates, multiple effect for each parameter estimate and significance.

Table 4 – Parameter Estimates for Rate of Change Model

Variable	Function	Parameter Estimate	Multiple Effect	Signifcance (Pr > ChiSq)
Intercept		-0.0611	94%	0.0013
Injury Class	(injclass in ('Sev ABI - 1','Sev ABI - 2','Quad - C5','Quad - C1 - C4'))	0.0828	109%	<.0001
Previous year payments for ATC and ACC	(Annual Payments <= 35000)	0.1872	121%	<.0001
Previous year payments for ATC and ACC	(35000 < Annual Payments <= 50000)	0.1047	111%	<.0001
Previous year payments for ATC and ACC	(50000 < Annual Payments <= 100000)	0.0587	106%	0.0033
Current Age	(current age > 60)	-0.0483	95%	0.0006
Service Year	(srvyear in (1999,2000))	0.0563	106%	<.0001

We note that this model includes an additional binary parameter for the service years 1999 and 2000. This parameter is a historic correction for higher than average increases in payment levels for these 2 years. This correction can be considered as a stepped superimposed inflation change.

Payment Amount Models

For claims currently not in the Active High state, we have used a model which predicts the total annual payments that a claim receives given that the claim is in an Active state and therefore receives a payment. The model constructed is the average Payment Per Active Claim (PPAC), which doesn't depend on the previous year's payment level. Two models are fitted for payments which will be received in the Active High state and payments which will be received in the Active Low state.

The observations used for the models included:

- Claims that transition into the Active High or Active Low state in the next year.
- For fatal claims, we included all past service years but excluded the year of death.

- We excluded claims in development year 0, since the first projection development year will always be 1.

The model fitted for Active High payments was a gamma GLM with a log link on the total payments in the next year where they are greater than $25,000. The parameter estimates from the model are presented in the table below.

Table 5 – Parameter Estimates for Active High Payment Amount Model

Variable	Function	Parameter Estimate	Multiple Effect	Significance (Pr > ChiSq)
Intercept		10.4165	33,406	<.0001
Injury class	Paraplegia - 100% disrptn of funct, Sev ABI - 4, Full use of arms & legs, Other Head - Other functional code	-0.2037	82%	<.0001
Injury class	Quad - C1 - C4	0.7103	203%	<.0001
Injury class	Quad - C7 - C8	-0.2904	75%	0.0004
Injury class	Sev ABI - 1, Sev ABI - 2, Mobility code gt 4	0.2832	133%	<.0001
Injury class	Sev ABI - 2, Mobility code le 4, Quad - C5	0.4984	165%	<.0001
Injury class	Sev ABI - 3, Mobility code gt 5	-0.1391	87%	0.0003
Injury class	Sev ABI - 4, Other mobility code	-0.4327	65%	<.0001
Injury class	Sev ABI - 5, Full use of arms & legs	-0.7126	49%	<.0001
Injury class	Other	0	100%	.
Current Age	Linear from 0 to 15	0.0498	105%	<.0001
Current Age	Linear from 40 to 60	-0.0098	99%	<.0001
Current Age	Linear from 60 to 70	-0.0215	98%	<.0001
Impairment %	< 50%	-0.1517	86%	<.0001
Impairment %	> 50%	0	100%	.

Some of the insights from this model include:

- Injury Class - Quads C1-C4 are twice as costly as the ‘Other’ injury group.
- Current age – Increasing costs up to age 15, decreasing costs from age 40 to 60, more dramatic decreases from age 60 to 70.
- Impairment Range - Injuries with less than 50% impairment are 24% lower than those with 50%+.

The Active Low model is of less significance however it still provides us with many insights. The parameter estimates from this model are presented below.

Table 6 – Parameter Estimates for Active Low Payment Amount Model

Variable	Function	Parameter Estimate	Multiple Effect	Signifcance (Pr > ChiSq)
Intercept		9.9212	20,357	<.0001
Injury class	'Sev ABI - 1','Sev ABI - 2','Quad - C5','Quad - C1 - C4','Quad - Unknown'	0.1911	121%	0.0093
Injury class	Other	0	100%	.
Current Age	Linear from 5 to 18	-0.049	95%	<.0001
Current Age	Linear from 18 to 45	0.0095	101%	0.0007
Current Age	Linear from 45 to 65	-0.0096	99%	0.0065
Impairment %	< 50%	-0.1188	89%	0.0036
Impairment %	> 50%	0	100%	.

Some of the insights from this model include:

- Expanded Injury Class - Severe injuries are 21% higher than the 'Other' injury group.
- Current age - Decrease at 5% per year of age up to 18, increase at 1% per year of age from 18 to 60, decrease at 1% per year of age from 45 to 65.
- Impairment Range - injuries with less than 50% impairment are 11% lower than those with 50%+.

Payment Amount Models with FIM and FAM

In the data section, we previously discussed that the FIM and FAM variables were not populated for around half of our Active High claims and as such we could not use this variable for SCE modelling purposes. We did however wish to understand how useful the variables would be in predicting future claims costs and in what way the data might be incorporated in the future. To answer these questions we reconstructed the Rate of Change and Active High payment amount models, adding in these variables and using only the claims where it was populated. In each case we took the current model (without FIM and FAM) as the starting point for constructing the new models. Before constructing these models we have investigated the relationship between the target and the FIM and FAM variables.

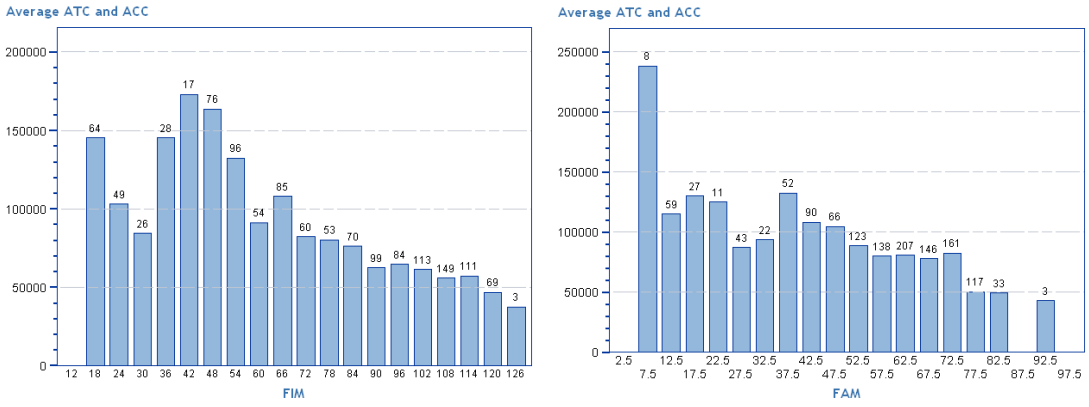

Figure 3 – Average Long Term Care Costs vs. FIM and FAM

The graphs above show the average long term care costs (represented by the bar height and read off the left axis and the number on the top of each bar represents the number of claim records used for each bar. For the FIM graph, there are increasing payment levels as the FIM decreases, to around 40.Below 40 the payment levels drop significantly and subsequently increase again from 30 to 20.

The FAM variable appears to be linearly related to long term care costs, such that decreasing FAM scores result in higher costs but with apparent less significance that the FIM variable.

When modelling with FIM and FAM we have adopted the approach of treating the 2 variables as separate and independent. In practice the FAM score is not used in isolation, the FIM and FAM are added together. However, from a modelling perspective we wanted to determine what the FAM added over and above the FIM and using the FIM plus the FAM would cause multicollinearity issues in the model which would cloud this analysis. Multicollinearity is the situation in which two or more predictors (or subsets of predictors) are strongly (but not perfectly) correlated to one another, making it difficult to interpret the strength of the effect of each predictor (or predictor subset)[1].

The parameter estimates from the Rate of Change model are shown below.

Table 7 – Parameter Estimates for Rate of Change Model

Variable	Function	Type of Variable	Parameter Estimate	Multiple Effect	Signifcance (Pr > ChiSq)
Intercept			-0.0848		0.0002
_bbinjclassHigh	(injclass in ('Sev ABI - 1', 'Sev ABI - 2', 'Quad - C5', 'Quad - C1 - C4'))	Binary	0.0682	107%	0.0002
_bbinf_subamt1	(Annual Payments <= 35000)	Binary	0.2086	123%	<.0001
_bbinf_subamt2	(35000 < Annual Payments <= 50000)	Binary	0.1371	115%	<.0001
_bbinf_subamt3	(50000 < Annual Payments <= 100000)	Binary	0.0782	108%	0.0009
_bbcurrage60	(current age > 60)	Binary	-0.0971	91%	0.0008
_bbFIM75	(FIM <= 75)	Binary	0.0578	106%	0.0021

We can see from this model that a binary variable for FIM less than or equal to 75 is added to the model and the positive estimate for this variable indicates that the, after allowing for payment level, claims with a low FIM score have higher increases in payments. The FAM variable did not add anything to the model once FIM was already included. Also note that the service year effect was not significant once FIM was added which may either mean that the service years 1999 and 2000 increases are explained by the FIM variable or alternatively the model has gained too many parameters and the service year effect is no longer significant. We believe that the latter explanation is more likely to be true.

This is because the FIM and FAM variables used in this model are not historical and likely to be low level dynamic. Due to the fact that these variables have only recently been coded, we cannot attain a

[1] Source: nature.com - http://www.nature.com/nrg/journal/v4/n2/glossary/nrg996_glossary.html

history for each claim. Through discussion with TAC we believe that for most claimants, where the injury has stabilised, the FIM and FAM would not change over time. However, some of the claims in our data period from 1997 to 2005 would have deteriorated significantly which would have resulted in significant increases in costs and as such the FIM parameter effect fitted may be somewhat artificial (though we believe this to be the smaller part).

The parameter estimates from the Active High payment amount model are presented below.

Table 8 – Parameter Estimates for Active High Payment Amount Model with FIM and FAM

Variable	Function	Parameter Estimate	Multiple Effect	Significance (Pr > ChiSq)
Intercept		10.2949	29,581	<.0001
FIM	(FIM <= 20)	0.4533	157%	<.0001
FIM	max(FIM-25,0)	0.025	103%	<.0001
FIM	max(FIM-47,0)	-0.0345	97%	<.0001
FAM	FAM	-0.0038	100%	0.0041
Injury class	Quad - C1 - C4	0.9786	266%	<.0001
Injury class	Sev ABI - 1,Sev ABI - 2, Quad - C5	0.4343	154%	<.0001
Injury class	Sev ABI - 4, Other mobility code, Quad - C7 -	-0.6997	50%	0.0019
Injury class	~BaseInjuryGrp	0	100%	.
Current Age	Linear from 0 to 15	0.0587	106%	<.0001
Current Age	Linear from 60 to 70	-0.0379	96%	<.0001
FIM x Injury class	max(FIM-25,0) x Quad - C1 - C4	-0.0129	99%	<.0001
FIM x Injury class	max(FIM-25,0) x Sev ABI - 1,Sev ABI - 2, Quad - C5	-0.0031	100%	0.0173
FIM x Injury class	max(FIM-25,0) x Sev ABI - 4, Other mobility code, Quad - C7 - C8	0.0064	101%	0.0403
FIM x Injury class	max(FIM-25,0) x ~BaseInjuryGrp	0	100%	.
Dev Year	(devyear=1)	-0.2447	78%	0.0116

The FIM variable is complicated to fit in the Active High payment amount model as the function fitted for FIM needs to accommodate the shape observed in Figure 3 above. We also need to add various interaction effects between the FIM and some injury class groups. This indicates that the effect of FIM is different across the various injury types. The FAM variable is also a significant predictor of future claim cost and is added to the model as a simple linear effect. In general, the FIM is much more significant than the FAM.

The addition of these 2 variables also reduces the reliance on the injury class variable and as such we are able to condense the level of detail for this variable in the model. Also the impairment percentage is no longer significant after allowing for FIM and FAM in the model. In conclusion, the FIM and FAM are highly predictive of future costs and reduce reliance on other variables in the model however, we need to be cautious as the data is not captured historically and hence the effects identified may be partly artificial (though we believe this to be the smaller part).

4 Model Evaluation

Performance as a Predictive Model

When building an SCE model we are attempting to balance a number of requirements for the final model which we discussed in the considerations for model design section earlier. The predictiveness of the model is one such aspect that we should consider however, the most predictive model is not always the most sensible or stable for forecasting and reserving purposes. Nevertheless we perform a detailed set of diagnostics around this aspect to determine if the fitted model is reasonable. Below is a short summary of these diagnostics for the interested reader but this does not form an integral part of the themes from this paper.

For each of the transition models we compare the actual and expected transition probabilities from the model by each of the predictors in the model. We also compute the gains chart for each transition modelled as with Brookes and Prevett 2003. The diagnostics were completed on the same sample as that used to construct the models because in all cases there were insufficient observations to allow us to hold back a dataset specifically for testing purposes. We will not show these diagnostics in this paper but in general the fitted models provided a reasonable level of predictiveness.

For the Rate of Change and Active High payment amount models we were able to hold back a random 30% of the data for testing purposes (around 700 observations). Based on this testing dataset we see that the model is highly predictive of the future claims cost. Using a crude statistical goodness of fit measure, the R-square, we see that it explains almost 90% of the total variation in claims costs.

For the Active High payment amount model we also hold back the testing dataset. Here we see a significant reduction in the predictive ability of the model such that it only explains around 32% of the total variation. This is not unexpected as we have removed the most important predictive variable (past payments). For further details on the predictiveness evaluations for these 2 models we refer the interested reader to Appendix A.

Performance as a Valuation Model

In setting balance sheet reserves for long term claimants there are several considerations which lead to an approach based on individual claimant reserves being preferable to aggregate valuation models:

- There is typically significant variability in the number of high cost claimants per accident year;
- There is also significant variability in the severity and hence cost of individual claimants;
- The age of individual claimants will impact the duration and hence liability associated with each claim;
- Reported claims comprise a large proportion of the liability, as such claims will typically be identified shortly after accident.

For more recent accidents more weight will typically be given to “average” payment assumptions rather than the payments made to individual claimants, as these can varying significantly in the first few years after accident. Other individual claimant characteristics can still be useful in setting individual reserves such as:

- Age
- Sex
- Injury severity or Functional/mobility assessment.

There are several reasons why caution needs to be taken in setting balance sheet reserves based on individual claimant models.

In the case of the TAC and other schemes managing such claims, the long term claimant pool is far from maturity. Systemic effects which have not been observed in the data may therefore emerge over time. These may include (but are not limited to):

- An increased expectation for long term care;
- A change in the provision of care. Typically some care is provided by families, but over time for a fixed group of claimants the ability of families/friends to provide care would be expected to reduce. In setting individual estimates based on past trends, it is important to consider the extent to which such long term impacts are represented by the data;
- The overall impact of the ageing population on the availability and cost of long term care;

Also, the amounts previously paid for long term care and other benefits may have been impacted by previous legislative and claims management environments. Any model derived based on the past data may therefore incorporate such previous environments in future projections, which may not be appropriate.

This situation is a potential issue with **any** model which relies on the past for predicting the future, and always needs to be recognised in setting balance sheet reserves. The very long term nature of these benefits for these claimants, however, means that the impact of any variation between past and future parameters (particularly growth factors, which have a multiplicative effect) can be significant.

Incurred Cost Development

One of the key diagnostics we perform to ensure that a valuation model is reasonable, is to track the incurred cost development from this model over a long period of time. The idea is that the actuarial incurred cost is estimated using the Statistical Case Estimation model at a valuation date in the past and subsequently re-estimated each year following this to observe if there is any upward or downward development. Any upward or downward trending in this incurred cost indicates that the model is potentially under or over stated.

The process for development of these charts is undertaken as follows:

1 All open claims as at 30 June 1997 are scored and total SCEs are calculated.

2 Payments made on these claims over next year are collected.

3 SCEs on these claims are recalculated as at 30 June 1998.

4 Steps 2 and 3 are repeated for each year up to 2006.

5 The total of SCEs plus payments made at each valuation date is graphed.

When performing this analysis we discovered that the incurred cost charts demonstrated an unusual shape which through further investigation we established was linked to the actual mortality experience over the period being higher than expected. As such we re-performed the analysis without including the mortality decrement and excluding any deaths from the actual experience. This enabled us to determine if all the other SCE model components were performing well.

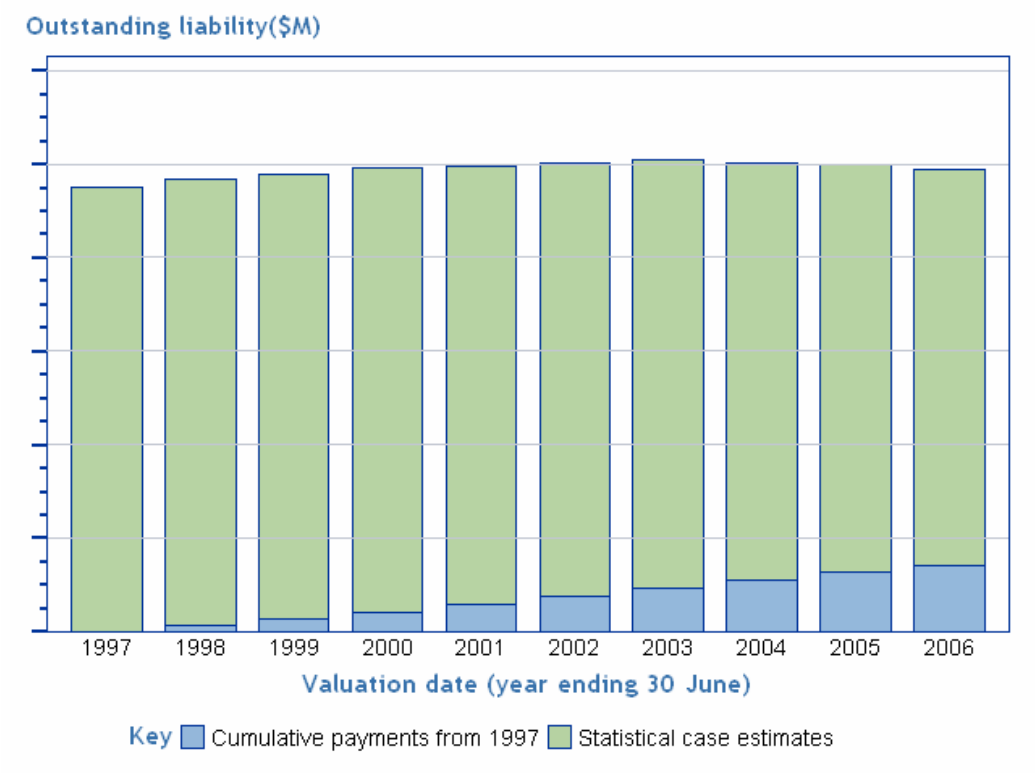

Figure 4 – Incurred Cost Development Chart (excluding mortality)

The total liabilities and payments presented in this chart are in current values as at 30 June 2006. We can see that the incurred cost increases over 6 years by approximately 1% pa and then decreases over the next 3 years. The total increase over the 9 years is less than 0.5% pa. Around 15% of the total liability at 1997 is paid out over the 9 years which demonstrates the extreme long tailed nature of the liability. In conclusion, we believe this analysis shows that the SCE model will likely develop in a stable way and it is not obvious that it will be either over or under stated.

One of the key benefits of an SCE model is the ability to divide the liability estimates into any subset of claims desired. The graph below shows the same incurred cost development charts as above, for the major claim types – non-catastrophic, quadriplegics, and severe acquired brain injuries.

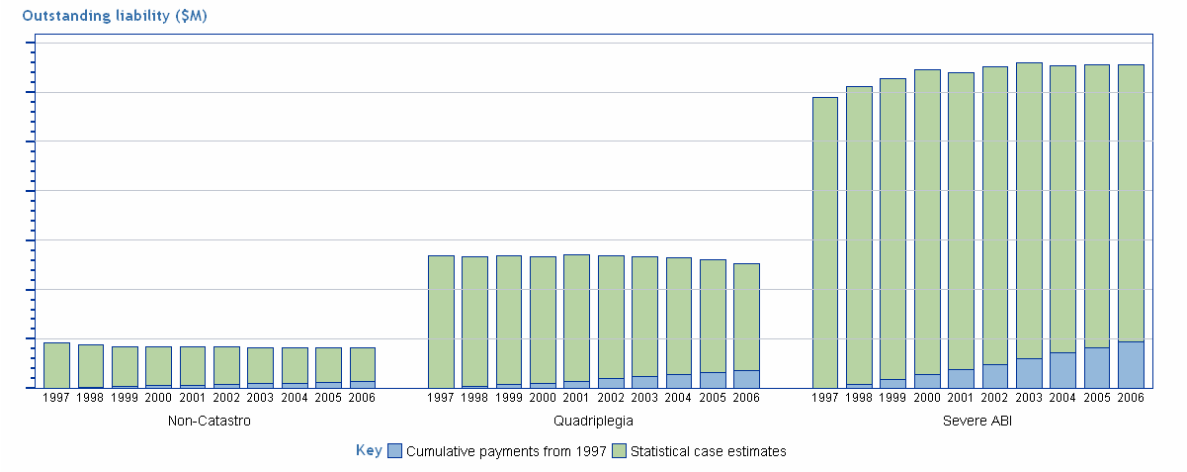

Figure 5 – Incurred Cost Development Chart (excluding mortality)

Here we see that much of the increase in the incurred costs for the early years, is largely driven by the severe ABI claims and the decreases in the later years are as a result of the quadriplegic claims.

5 Insights into Claims Cost Drivers

Key Drivers

Constructing statistical models to predict claims costs involves testing and deciding which claim characteristics are significant drivers for inclusion in the predictive model. For the TAC long term claimants we have tested each of the variables available in 3 different models. The first one allows for past payments as a predictor in the model. It is clear that past payment levels are generally a good predictor of future payments. The key drivers for this model are presented in the pie chart below with their final predictive contribution to the model after all other variables are allowed for, represented by the size of each slice.

To determine the contribution of the variables in the pie chart we have undertaken the following process.

- The unexplained component is based on 100% minus the R-square statistic calculated using the independent test dataset. Whist we recognised that the R-square statistic is not an ideal goodness-of-fit measure in all cases it is widely used and has the intuitively appealing interpretation of "representing the proportion of the variation explained by the model".
- The contribution of the variables is based on the "Type 3 Chi Square" statistic from the fitted GLM model. The Type 3 statistic represents the reduction in the goodness-of-fit should the key variable be removed and all other variables remain in the model. This is a common approach to assessing the importance or impact of each variable in many statistical and data mining situations.
- For interaction effects in the model we have taken the simple approach of assigning half of the Type 3 statistic to each of the variables in the interaction (all of our interactions only contain 2 variables).

In summary, this approach to determining the key drivers and their contribution is a balance between relying on appropriate statistical measures while still maintaining a pragmatic approach. We believe that it is unlikely that the high level messages will be much different under any reasonable alternative.

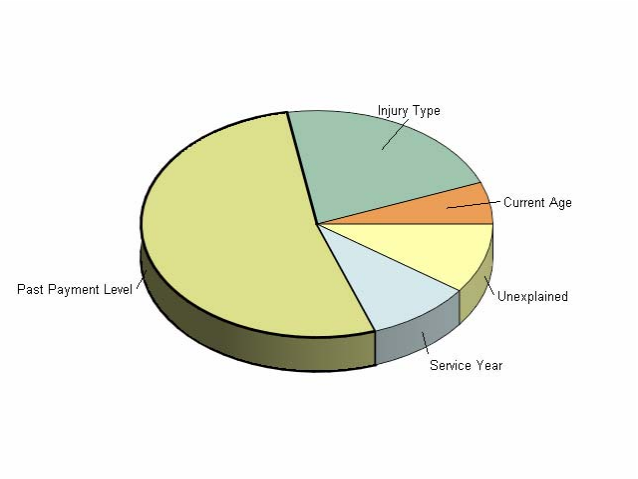

Figure 6 – Key Claims Cost Drivers with Past Payments

We see that the past payment level is the most important driver contributing a little over half of the total predictive ability. The injury type of the claimant contributes a further quarter and the service year of payment and current age of the claimant also contribute to a lesser extent. The unexplained component shows how much of the variation in claims cost is still unexplained by these key drivers. For this model the unexplained component is only around 10%.

The second model used does not include past payments as a key driver. This analysis is to determine which drivers are the strongest for a claim when we don't currently know what the payment level is (e.g. for new claims or claims that have no payment history). The contribution of each of the drivers from this model are presented in the chart below.

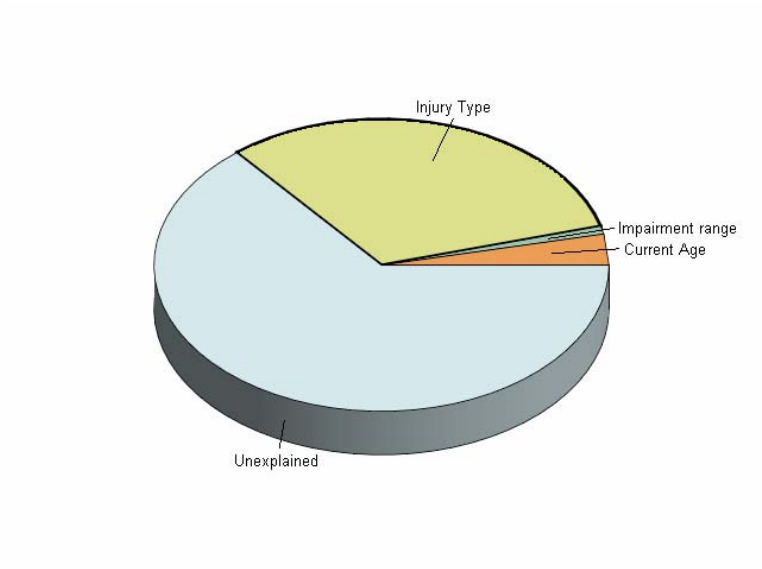

Figure 7 – Key Claims Cost Drivers without Past Payments

This pie chart shows that a large majority of the variation in the claims cost is actually unexplained by any of the claims drivers we use. While this is unfortunate, a model with a large unexplained component can still be very useful for prediction of small groups of claims but may not perform well on individual claims. The most important driver by far in the model is the injury type while the impairment range and current age make much smaller contributions.

The final model we have used to uncover the key drivers of claims cost excludes past payments as in the second approach but includes the Functional Independence Measure ("FIM") and Functional Assessment Measure ("FAM"). These measures are a scoring system for the estimate the level of ongoing care needs required for each claim.

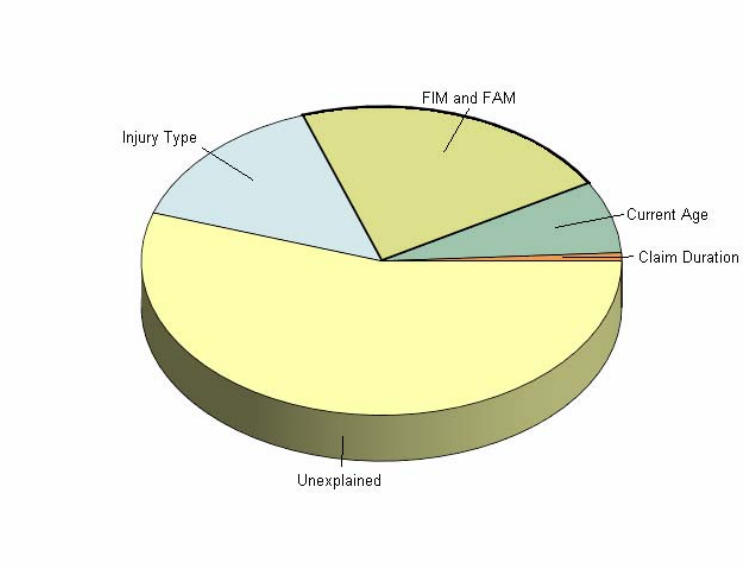

Figure 8 - Key Claims Cost Drivers without Past Payments and with FIM and FAM

The unexplained component from this analysis demonstrates around half of the variation in claims costs. The FIM and FAM variables added to this model are the most important driver of claim cost. Injury type is the second most significant predictor but represents a significantly smaller contribution as a result of FIM and FAM being the model. Current age has increased in its contribution to this analysis while claim duration has remained a small contributor.

The magnitude of the unexplained component demonstrates that there are a number of factors contributing to variation in claims costs which are not included in the model. Anecdotal evidence suggests that other data items such as family circumstances and claimant activity in community participation can be important drivers of claim costs, and while this information is often known, it is not necessarily recorded in a manner which makes it amenable to statistical analysis). It would therefore appear that where there is some evidence that such data **can** assist in predicting claims costs that some effort be made to collect and record it in a systematic manner.

Hypotheses Tested

The key drivers highlighted in the previous section provide a significant insight into understanding the claims process and what impacts upon it. Also of interest is what potential or hypothesised variables were tested and turned out not to be significant claims cost drivers. During the exploratory data analysis and model construction phase we tested the following drivers.

Table 9 –Predictors used in the Final SCE Model

Variable	Rate of Change model	Active High payment model	Active Low payment model	Transition models	Comment
Age	☐	☐	☐	☐	Includes age at accident and age at injury.
Attendant care and Accommodation payments	☐	☐	☐	☐	Included in the Rate of Change Model and incorporated as a transition state.
Duration since accident	☐	☐	☐	☐	Only significant for transition models.
FAM	☐	☐	☐	☐	* Significant if applicable, does not add much more to FIM.
FIM	☐	☐	☐	☐	* Significant if applicable
Impairment range	☐	☐	☐	☐	Mildly significant in the PPAC Models
Injury class	☐	☐	☐	☐	Includes the category of injury (Quad, Para, Severe ABI, Non-Cat) and severity of injury (high, mid, low).
Therapy payments	☐	☐	☐	☐	Includes speech and occupation therapy. Incorporated as a transition state.

Table 10 –Predictors Tested and NOT used in the Final SCE Model

Variable	Rate of Change model	Active High payment model	Active Low payment model	Transition models	Comment
Functional code	☒	☒	☒	☒	Correlated with injury class, so not used in the models.
Mobiliity code	☒	☒	☒	☒	Correlated with injury class, so not used in the models.
Service profile	☒	☒	☒	☒	Correlated with injury class, so not used in the models.
Days since discharge	☒	☒	☒	☒	Data issues with the discharge date based on the most recent discharge rather than the first.
Residential status	☒	☒	☒	☒	Significant but is dynamic, not used in models.
Days in accommodation	☒	☒	☒	☒	Significant but is dynamic, not used in models.
Days in attendant care	☒	☒	☒	☒	Significant but is dynamic, not used in models.
Gender	☒	☒	☒	☒	Tested, not significiant
Year of accident	☒	☒	☒	☒	Tested, not significiant

It may be the case that the driver was not significant because either:

- The data captured to reflect the driver is not appropriate or reliable
- The historical experience for the driver and outcome does not reflect what we expect to see in the future

Quantification of Drivers

We can examine the drivers tested and incorporated in the model as described above to determine the effect they have on the overall cost of the claims. A final summary of the key insights derived from the models are presented below.

- Current payment level is the best predictor of next year payments (except when the current level is low, < $25,000)
- Injury class is the next strongest predictor of payment level (in both the Rate of Change and Payment Amount models)
- Payment levels decrease after the claimant reaches age 60
- Accommodation status (defined as accommodation payments > $0) was not a significant predictor

- FIM and FAM were not initially used because of missing values and a lack of history

Effects Across Time

We have undertaken analysis into the effects of the dynamic variables in the model across time. Some of the insights derived from the model are presented below.

- Young claimants (under 21) on high levels of long term care are up to 3 times more likely NOT to continue at these high levels (than older claimants).
- Claimants between 16 and 25 are almost twice as likely to cease low levels of long term care and receive only Therapy payments (than all other ages).
- Children less than 16 are 3 times more likely to take on low levels of long term care while they are receiving only Therapy payments (than older claimants).
- Quadriplegics and Sev ABI -1 claims are around 10 times more likely to continue on high levels of long term care than Non-Catastrophic claims.
- Changing from low levels to high levels of long term care is more likely as the claimant gets older.
- Sev ABI – 3 and Quads C6-C8 are 4 times more likely than non-catastrophic claims, to take on low levels of long term care while they are receiving no payments at all.

6 Model Applications

Claims Management

SCEs/Individual estimates can be used to assist in claims management in a number of ways.

The actual payments being made to each claimant can be compared with the payments estimated by the SCE. Outliers can then be detected and reviewed on a case-by-case basis. Often these may result from individual circumstances which are not captured by the available data. This may however lead to additional data items being captured, which in turn will improve the SCE model and lead to less outliers in future.

There may be some cost drivers which are not able to be accurately captured and/or modelled. This is a feature of any statistical model however and should not preclude the use of SCEs/Individual models for long term claimants

SCEs can also be used to track the performance of various cohorts of claims over time. This can be used in assessing the performance of the claims management function.

It will often be the case that claims staff will estimate the care needs of claimants over coming years. These estimates can be compared with the SCE and ideally over time both the SCEs and staff estimates will converge, based on observed past patterns.

7 Summary and Conclusions

General model structure

In order to model long term care costs for the Transport Accident Commission, we have constructed a statistical case estimate model which consists of a number of components:

- A transition state model consisting of four states (other than mortality), namely Inactive, Inactive Therapy, Active Low and Active High
- A rate of change model for claims which remain in the Active High State and
- Two payment models, one for claims which transfer to the Active High State (from a different state) and one for any claim in the Active Low State.

Key predictive variables

In modelling the transition between various states, the most important variables are age, duration since injury and injury class.

In modelling the payments for claims currently in the Active High state, the most important variable is the current level of payments. Also important is injury class and age.

The Functional Independence Measure (FIM) is also a useful predictor of claims costs when available.

Applications

An individual model is preferable to an aggregate model in the valuation of long term claimants where sufficient data exists.

Caution needs to be taken with systemic effects which have not been observed in the past data but which may therefore emerge over time.

The amounts previously paid for long term care and other benefits may have been impacted by previous legislative and claims management environments. Any model derived based on the past data may therefore incorporate such previous environments in future projections, which may not be appropriate. The very long term nature of these benefits for these claimants, however, means that the impact of any variation between past and future parameters (particularly growth factors, which have a multiplicative effect) can be significant.

Predictive modelling can also be an important tool in claims management.

8 Bibliography

Brookes R & Prevett M, 2004, *Statistical Case Estimation Modelling - An Overview of the NSW WorkCover Model*, Presented to the Institute of Actuaries of Australia Accident Compensation Seminar 28 November to 1 December 2004.

Appendix A

Model Evaluation – Gains Charts, Actual vs. Expected, R-Square and Co-efficients of Variation

Gains charts plot the cumulative total cost captured by a model. Claims are ranked by predicted payment size. The total payments within each percentile of the cases is summed and divided by the total cost of all cases. A good model will capture a large proportion of the total cost in the upper percentiles. This plot is analogous to the 80/20 rule, where we capture 80% of the cost with 20% of the effort. The 80% would be the gains for the top 20% of cases, as ranked by the model.
Gains plots for the modelled predictions are compared to two other benchmarks.

1 No model – in this situation we would be randomly selecting 20% of the cases from the population and by general reasoning only 20% of the total cost would be captured. A completely random selection of cases will always produce this result and as such will lead to a gains plot which is a straight line from 0% to 100%.

2 Perfect model – here the model ranks cases from highest to lowest exactly, producing the theoretical best model possible. To determine the gains we would achieve from such a model we can simply rank the cases by their actual values.

In the plots below, the cumulative percentages for the gains plots are on the right hand size axis and the band percentiles on the horizontal axis. The green line is the gains plot of the predictions and from the graph, the top 5% of predicted claims represents approximately 40% of the total cost. The purple line shows the cumulative costs of the actual payments and from the graph, the top 5% of claims represents approximately 50% of the total cost. A good model is as close as possible to the purple line and far above the diagonal line. By itself, the gains chart does not indicate a great deal about the goodness-of-fit of the model. However, it becomes very useful when comparing two competing models (where higher gains means a better model).

We also use the root average squared error (RASE) as a measure of the goodness-of-fit of the model. This statistic is similar to the standard deviation or spread of the residuals and hence a lower RASE indicates a better model. A related measure is the root mean square error (RMSE) which differs because it adjusts for the number of parameters in the model. For data mining problems with large samples sizes the differences between these statistics is minimal. Often we take the RMSE or RASE one step further and divide it by the mean target value to calculate the coefficient of variation ("CoV" or "CV") for the model residuals. Again, a lower CV for model residuals indicates a better model fit. Caution should always be exercised when interpreting these measures because they can be seriously affected by outliers.

The red and blue lines in the plot are read from the left axis and represent actual and expected average values within each percentile of observations. Tabulations of these average values within each decile (group of 10 percentiles) are calculated to allow the comparison of the smaller average values (in the lower deciles) which are not distinguishable in the plot.

To generate the tabulations and plot lines, the predictions were again ranked in ascending order and assigned into bands. Then the average target and average prediction value were calculated for each band. The bands adopted for the plot lines are percentiles and those adopted for tabulation are decile bands and also the top 10 percentile bands.

A good model had similar average actual and expected values across the range of bands with no observable or consistent bias. An even better model has a greater differentiation in predicted values, again without observable bias.

Model Evaluation for the Active High Rate of Change and PPAC Models

The actual vs. expected and gains chart for this model on the testing data is shown below.

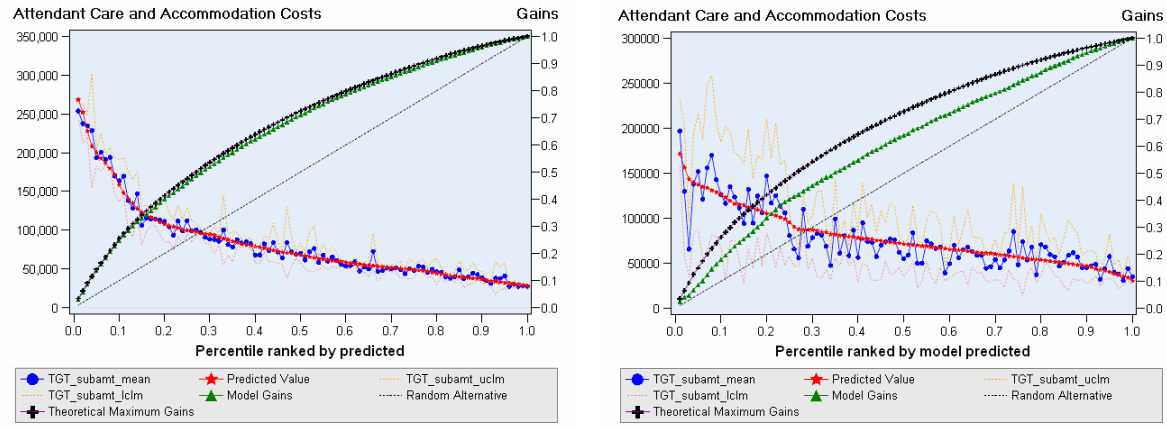

Figure 9 – Actual vs. Expected and Gains Chart for Rate of Change (left) and Active High Payment Amount (right) Models

The Rate of Change model on the left demonstrates a very close fit between actual and expected payments however, this is not surprising given that the past year's payments are included in the model and these payments are such a good predictor of the future year. This model has an R-square value of 87% and a CoV of 23% on the test dataset.

For the Active High payment amount model on the right shows we also compute the actual vs. expected and gains charts based on a testing dataset. This model shows considerably more variability in the actual payments around the model expectation. We also see that the differentiation of the model is much less with predictions only stretching up to $150,000 while the Rate of Change model reaches $250,000. The gains line is also considerably worse than the Rate of Change model as with the model R-square (32%) and CoV (56%). Although the performance of this model is poor compared with the Rate of Change model it is still very useful and in fact the best alternative for the situation where the claim has no previous payments to use for prediction.

We also complete the same diagnostics on the Active Low payment amount model but have not presented the results in this paper. In general, the model has only a low level of predictiveness but is robust and does not form a material part of the overall SCE model.